RU EN ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ ISSN 2079-3316
научная статья ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И МАШИННОЕ ОБУЧЕНИЕ

УДК 004.934:004.032.26
10.25209/2079-3316-2026-17-3-315-362

Влияние начального темпа обучения на свойства нейронных сетей в задаче распознавания речи

Ильдус Рустемович **Садртдинов**✉
Национальный исследовательский университет «Высшая школа экономики», Москва, Россия
✉*isadrtdinov@hse.ru*

Аннотация. Начальный темп обучения стохастического градиентного спуска играет определяющую роль в качестве получаемого решения и структуре выучиваемых нейронной сетью признаков. В недавних работах для задачи классификации изображений установлена таксономия трех режимов начального темпа обучения и выделен узкий диапазон высоких значений, обеспечивающий наилучшее качество после дальнейшего дообучения с низким темпом обучения или усреднения весов.

В статье эта методология распространяется на задачу распознавания речи, где таксономия режимов воспроизводится в основных чертах, а выделенный диапазон высоких значений сохраняет превосходство по итоговому качеству дообучения и усреднения весов и линейную связность полученных решений в пространстве весов. Картина выучиваемых признаков воспроизводится в более мягкой форме: с ростом темпа обучения опора моделей смещается к нижним кепстральным коэффициентам и низкочастотной части мел-спектрограммы, однако внутри области высоких значений это смещение не выделяет диапазон наилучшего качества. Дополнительно предложено усечение мел-частотного кепстра как переносимый инструмент анализа выученных признаков, применимый к любой модели, работающей со спектрограммой.

Таким образом, выделенный режим не является артефактом задачи классификации изображений, а успешно переносится на задачу распознавания речи.

Ключевые слова и фразы: темп обучения, стохастический градиентный спуск, распознавание речи, линейная связность, выучивание признаков

Для цитирования: Садртдинов И. Р. *Влияние начального темпа обучения на свойства нейронных сетей в задаче распознавания речи* // Программные системы: теория и приложения. 2026. **Т. 17**. № 3(72). С. 315–362. https://psta.psiras.ru/read/psta2026_3_315-362.pdf

© Садртдинов И. Р. 2026 CC BY

Введение

В современном мире многие задачи анализа данных, от классификации изображений [2, 3] и распознавания речи [4, 5] до генерации текстов [6, 7] и предсказания пространственной структуры белков [8], успешно решаются с помощью применения искусственных нейронных сетей [9, 10]. Несмотря на столь значимый успех нейросетевых моделей в самых разных доменах данных, важным ограничением их внедрения остается низкая интерпретируемость [11, 12], т.е. возможность объяснить, почему модель выдает то или иное предсказание для конкретного примера. Попытки интерпретации нейронных сетей упираются в высокую размерность пространства параметров и в сложный, нелинейный процесс оптимизации [13, 14]. Таким образом, исследования, направленные на улучшение понимания свойств нейросетевых моделей и процесса их обучения не теряют актуальности до сих пор.

В силу невыпуклости функции потерь нейронных сетей и использования стохастических алгоритмов оптимизации, правильный выбор гиперпараметров, например темпа обучения (англ. learning rate), оказывается ключевым для получения решения с высоким качеством. Согласно множественным исследованиям [15–17], именно темп обучения играет определяющую роль при определении характеристик решения, к которому сходится процесс оптимизации. В частности, существующие практические рекомендации [18–22] предлагают начинать обучение с относительно высокого значения темпа обучения, поскольку он позволяет избегать минимумов с низкой обобщающей способностью [15, 23–25] и высокой кривизной функции потерь [16, 17], а также добиваться более удачной структуры признаков, воспринимаемых итоговой моделью [26, 27]. Тем не менее, большинство предыдущих работ изучает влияние высокого темпа обучения исключительно на синтетических данных или на относительно простых постановках, например, классификации изображений.

Данное исследование во многом вдохновлено работой [27], в которой изучается влияние начального темпа обучения на обобщающую способность моделей, геометрию пространства весов и выучивания признаков моделями. Впрочем, авторы ограничиваются рассмотрением задачи классификации изображений, в то время как данная работа расширяет их результаты на звуковые данные, а именно на задачу распознавания речи.

Основной вклад представляемой работы состоит в следующем:

- Таксоноомия трех режимов темпа обучения [24] вместе с подрежимом 2A [27] воспроизводится для задачи распознавания речи в практической (не масштабно-инвариантной) постановке.

- Структура подрежима 2A при этом уточняется: различные его диагностические свойства выполняются на разных диапазонах PLR (см. разделы 3 и 4).
- Анализ выученных признаков из [27] переносится на звуковой домен, и обнаруживается, что преимущество моделей 2A по качеству локализовано в низкочастотных компонентах мел-спектрограммы; при этом опора на первые кепстральные коэффициенты отделяет режим 1 от режима 2 в целом, но внутри последнего подрежимы 2A и 2B по этому признаку неразличимы (см. раздел 5).

Полученные результаты показывают, что подрежим 2A не ограничен задачей классификации изображений и может проявляться и в других доменах данных (аудио) и целевых задачах (распознавание речи).

1. Обзор смежных работ

1.1. Масштабно-инвариантные нейронные сети

Слои нормализации – необходимый компонент подавляющего большинства современных архитектур нейронных сетей: BatchNorm [28] используется в сверточных сетях типа ResNet [29], а LayerNorm [30] – в трансформерах [31]. Помимо ускорения и стабилизации обучения [32, 33] и сглаживания ландшафта функции потерь [34], слои нормализации обладают важным побочным эффектом: они делают параметры предшествующего им слоя *масштабно-инвариантными* (МИ), иными словами, выход слоя нормализации не зависит от нормы параметров предыдущего слоя.

Следовательно, слои нормализации существенно меняют динамику процесса оптимизации. Показано, что BatchNorm неявно настраивает темп обучения [35], а направление вектора весов на самом деле эволюционирует не согласно номинальному темпу обучения (англ. learning rate, LR) η, а согласно *эффективному темпу обучения* (англ. effective learning rate, ELR) $\eta_{\text{eff}} = \eta/\|\boldsymbol{w}\|^2$, зависящему от текущей нормы параметров $\|\boldsymbol{w}\|$ [36]. Аналогично, в масштабно-инвариантных сетях коэффициент регуляризации весов λ (англ. weight decay) перестает играть роль классического регуляризатора и вместо этого управляет темпом обучения через норму параметров [37]; данный эффект подтвержден и уточнен в работах [38, 39].

Из-за зависимости ELR от нормы параметров прямое исследование влияния номинального LR на современные архитектуры осложняется: даже при фиксированном η эффективный темп обучения непрерывно меняется по мере изменения нормы весов в процессе обучения. Поэтому аккуратный анализ роли темпа обучения требует либо явного контроля за нормой параметров, либо перехода к полностью масштабно-инвариантным (МИ) сетям, то есть таким, у которых выход всей сети целиком не зависит от нормы параметров.

Такое ограничение достигается заморозкой параметров последнего линейного слоя (т.н. классификационной головы, англ. classifier head) и удалением обучаемых аффинных параметров в слоях нормализации [35, 40]. В дальнейшем сеть обучается проективными алгоритмами оптимизации на сфере фиксированного радиуса [24]. В такой постановке фиксация LR приводит и к фиксации ELR, что делает результаты пригодными для строгого анализа. Именно такой методологии придерживаются работы [24, 27].

В рамках этого подхода изучен ряд нетривиальных эффектов равновесной динамики МИ-сетей. С использованием аппарата стохастических дифференциальных уравнений в работе [40] показано, что стационарное распределение параметров на сфере зависит от произведения $\eta\lambda$ (т.н. внутреннего темпа обучения, англ. intrinsic learning rate).

В статье [41] обнаружено, что при обучении с регуляризацией весов МИ-сети могут испытывать периодические неустойчивости, в которых значения нормы весов и функции потерь демонстрируют циклическое поведение. Ключевой для дальнейшего изложения результат получен в статье [24], где показано, что в зависимости от величины ELR динамика МИ-сети при обучении на фиксированной сфере попадает в один из трех качественно различных режимов (см. ниже). Работа [25] расширяет эти наблюдения на квадратичную ошибку.

1.2. Влияние темпа обучения на оптимизацию

Темп обучения – ключевой гиперпараметр градиентной оптимизации: из-за невыпуклости функции потерь нейронной сети и стохастичности используемых алгоритмов оптимизации именно он во многом определяет, к какому именно решению сойдется процесс обучения [15–17, 42–44]. Существующие практические рекомендации [19–22] предлагают начинать обучение с относительно высокого значения LR, поскольку он позволяет избегать минимумов с низкой обобщающей способностью [15, 23–25, 44, 45] и высокой кривизной функции потерь [16, 17], а также добиваться более удачной структуры признаков, воспринимаемых итоговой моделью [26, 27, 46–50].

С точки зрения кривизны функции потерь, шум стохастического градиентного спуска (англ. stochastic gradient descent, SGD) [51] увеличивается вместе с величиной LR, что не позволяет оптимизации задерживаться в минимумах с высокой остротой (англ. sharpness) [15]. В частности, механизм «катапультирования» [16] описывает, как траектория обучения с высоким LR выбрасывается из окрестности острых минимумов в более

пологие области пространства весов. Работа [17] показывает похожий эффект с другой стороны: градиентный спуск с фиксированным LR самостоятельно стабилизируется на границе устойчивости (англ. edge of stability), что напрямую связывает величину LR с достижимой остротой минимума.

С точки зрения выучивания признаков, высокий темп обучения также способствует выучиванию более разреженного набора признаков. SGD с высоким LR тяготеет к решениям с разреженной активацией скрытых нейронов [46] и к «стохастическому коллапсу» в сторону более простых подсетей [47]. Похожий эффект наблюдается и для отдельных нейронов: высокий LR приводит к обучению «пороговых нейронов» в ReLU-сетях [48], а также к более разреженным решениям в линейных диагональных сетях [49]. Кроме того, высокий темп обучения позволяет отфильтровывать «противоборствующие» ложные признаки (англ. spurious features), общие для объектов разных классов [50].

Центральной для дальнейшего изложения является таксономия трех режимов обучения, предложенная в [24]: в зависимости от величины эффективного темпа обучения динамика попадает в

1) режим *сходимости* (низкие LR, параметры монотонно сходятся к минимуму),
2) режим *хаотического равновесия* (средние LR, функция потерь стабилизируется на ненулевом уровне с шумовыми колебаниями) или
3) режим *расходимости* (высокие LR, качество модели соответствует случайному угадыванию).

Отметим, что режим сходимости в данной таксономии достижим только для *перепараметризованных* (англ. overparameterized) сетей, т.е. таких, у которых число обучаемых параметров существенно превышает объем обучающей выборки, что позволяет SGD достигать пренебрежимо малой ошибки на обучающей выборке, то есть интерполировать данные [52–54].

Для сетей с ограниченной емкостью (недопараметризованных) (англ. underparameterized) при малых LR шумовая компонента стохастического градиента не позволяет параметрам сойтись к точке минимума: вместо этого траектория стабилизируется в стационарном распределении вблизи точки минимума, а не в ней самой [55], на манер режима хаотического равновесия. Авторы также отмечают, что режим 2 — это тот диапазон, в котором стартует подавляющее большинство практических расписаний темпа обучения, поскольку он не приводит ни к преждевременной сходимости, ни к расходимости. При этом сама работа [24] не дает ответа на вопрос, какие именно значения LR внутри этого широкого режима предпочтительны для начала обучения.

Именно этот вопрос детально исследуется в [27], продолжающей и уточняющей таксономию [24]. Там исследуется описанный спектр темпов обучения с точки зрения качества дальнейшего *дообучения* с низким LR или *усреднения весов* (англ. stochastic weight averaging, SWA). В рамках усреднения весов вместо одного финального решения берется средний вектор весов нескольких контрольных точек (англ. checkpoints), сохраненных вдоль траектории обучения, что позволяет добиваться более высокого качества по сравнению с одной моделью [56]. Авторы показывают, что режим хаотического равновесия неоднороден и распадается на два подрежима с существенно различными свойствами итоговых решений: *2A* — нижняя часть режима непосредственно над порогом сходимости (т.е. границей между 1 и 2 режимами), обеспечивающая наилучшее качество после дообучения и усреднения весов, и *2B* — его верхняя часть, при которой это преимущество теряется. Кроме того, авторы отмечают два важных различия между 2A и 2B:

1) *Геометрия пространства весов.* LR из подрежима 2A локализуют выпуклую область пространства весов, содержащую только качественные, линейно связные минимумы. Под линейной связностью (англ. linear mode connectivity) [57–59] здесь понимается следующее: если взять два обученных решения и линейно проинтерполировать между ними в пространстве весов, то вдоль всего отрезка ошибка будет оставаться низкой. Поскольку функция потерь нейронных сетей невыпукла, между двумя независимо обученными моделями обычно возникает барьер ошибки, и в этом смысле линейная связность — это скорее исключение, чем правило.

   Отсутствие барьера означает, что оба решения лежат в одной и той же области пространства весов, в которой функция потерь ведет себя выпукло, а значит, и все промежуточные точки отрезка между ними тоже являются качественными решениями. Именно поэтому усреднение весов таких решений не портит модель, а улучшает ее [56]. LR из подрежима 2B, напротив, приводят к обширным областям с разнообразными, линейно не связными минимумами, и усреднение решений, полученных из разных запусков, в этом случае не улучшает качество.

2) *Выучивание признаков.* LR из подрежима 2A приводят к разреженному набору наиболее релевантных для задачи признаков: например, к смещению в сторону средних частот при классификации изображений. LR из подрежима 2B, напротив, постепенно теряют разреженность, а затем и способность извлекать какие-либо полезные признаки из данных.

1.3. Перенос выводов на звуковой домен

Большая часть рассмотренных выше результатов получена либо на синтетических данных, либо на задачах компьютерного зрения. Тем не менее, эти направления исследований находят параллели и в звуковом домене. Связь между остротой минимума и обобщающей способностью показана и для звуковых моделей: в задаче детекции дипфейков [60] авторы измеряют остроту минимума и показывают, что она растет именно при сдвиге домена, то есть на данных, не встречавшихся при обучении, а явное уменьшение остроты статистически значимо коррелирует с улучшением обобщающей способности. Этот результат подтверждает, что связь остроты и генерализации, установленная для изображений [15, 23], оказывается актуальной и для звуковых моделей.

Отдельного внимания заслуживает исследование [61], изучающее эффект различных расписаний разогрева темпа обучения (англ. learning rate warmup) при обучении моделей распознавания речи. Авторы показывают, что звуковым моделям требуется существенно более медленный разогрев LR, чем это принято для текстовых трансформеров: более высокий LR на этапе разогрева ускоряет сходимость, но не улучшает итоговое качество модели. Схожий по духу результат получен и для другого ключевого гиперпараметра SGD – размера пакета (англ. batch size): в работе [62] показано, что качество итоговой акустической модели определяется в первую очередь суммарным объемом использованных данных, то есть произведением размера пакета на число итераций, а не размером пакета самим по себе.

Работа [63] анализирует, какие частоты выучивает первый сверточный слой модели для распознавания речи, обучаемой на сырой звуковой волне. Авторы сравнивают частоты выученных фильтров со стандартной шкалой мел-фильтров, которая обычно используется для ручного построения звуковых признаков, и показывают, что сеть выучивает собственный набор фильтров, чье распределение по шкале частот заметно отличается от мел-банка. Иными словами, модель формирует свою частотную структуру, специфичную для задачи, а не воспроизводит заранее заданную акустическую шкалу.

Более ранние работы [64, 65] демонстрируют похожий эффект на другом уровне: в классических акустических моделях, обучаемых не на сырой звуковой волне, а на спектрограммах, отдельные нейроны скрытых слоев постепенно становятся избирательны к конкретным фонетическим признакам, например, к способу артикуляции звука. Эта избирательность заметна уже в первых слоях и усиливается по мере углубления сети. Оба

наблюдения мотивируют перенос анализа выучивания признаков из подрежима 2A [27] на звуковой домен: в задаче распознавания речи роль релевантных признаков естественным образом берут на себя именно частотные полосы и фонетические характеристики, которые выучивают акустические модели.

2. Методология исследования

Эксперименты следуют методологии из статьи [27]. А именно, рассмотрим двухстадийное расписание темпа обучения, которое состоит из *предобучения* с постоянной величиной **P**LR (англ. pre-training learning rate), которое продолжается либо *дообучением* с низким **F**LR (англ. fine-tuning learning rate), либо *усреднением весов* SWA. Для SWA обучение продолжается с величиной темпа обучения PLR, в течение которого собирается несколько контрольных точек для усреднения. Основная цель исследования – определить, как величина PLR влияет на обобщающую способность и свойства итоговой модели.

2.1. Данные

Эксперименты проводятся на наборе данных для распознавания речи LJSpeech[1]. Всего набор содержит 13 100 записей от одного диктора продолжительностью от 1 до 10 секунд на английском языке. Записи представляют собой выдержки из аудиокниг.

Случайная подвыборка из 50 записей используется для отладки экспериментов, еще 50 записей – в качестве тестовой выборки. Из оставшихся 13 000 записей для обучения используются только 5%, т.е. 650 записей. Этот выбор продиктован тремя соображениями.

Во-первых, при таком размере обучающей выборки модель становится перепараметризованной, что позволяет экспериментировать с режимом сходимости, который не наблюдается при рассмотрении полной выборки в 13 000 записей.

Во-вторых, скудность обучающей выборки гарантирует разную обобщающую способность при обучении с различными LR.

В-третьих, в данном случае исследование становится особенно актуальным, поскольку изучает поведение моделей в условиях ограниченной обучающей выборки, что часто встречается на практике.

[1] https://keithito.com/LJ-Speech-Dataset/

Рассматриваются три независимо разыгранные подвыборки такого размера; каждая из них генерируется один раз и переиспользуется на обеих стадиях обучения, так что предобучение и соответствующее ему дообучение всегда видят одни и те же данные (см. раздел 2.5). При предобработке записи приводятся к частоте дискретизации 16 кГц. В качестве метрики качества используется ошибка на уровне слов (англ. word error rate, WER) [66], стандартная для задачи распознавания речи.

2.2. Признаковое описание

Для каждой звуковой записи вычисляется оконное преобразование Фурье (англ. short-time Fourier transform, STFT) с окном Ханна длины 400 отсчетов (25 мс) и шагом 160 отсчетов (10 мс) при размере преобразования $N_{\mathrm{FFT}} = 512$. Сигнал симметрично дополняется отражением по краям (англ. reflect padding), поэтому число временных шагов равно $T = 1 + \lfloor L/160 \rfloor$, где L – длина записи в отсчетах. К квадрату модуля спектра применяется банк из 64 треугольных мел-фильтров, покрывающих полосу от 0 до 8000 Гц; выходы этих фильтров далее называются компонентами мел-спектрограммы. К ним применяется логарифм $\log(x + 10^{-5})$.

Полученная мел-спектрограмма нормируется отдельно для каждой записи: из каждой из 64 компонент вычитается ее среднее по времени, а результат делится на стандартное отклонение по времени. Внутри пакета записи дополняются нулями (англ. padding) до длины самой длинной из них, при этом статистики нормировки считаются только по фактическим временным шагам, а дополненные шаги обнуляются. В противном случае короткие записи в пакете с одной длинной оказались бы смещены относительно нуля, что исказило бы статистики последующих слоев BatchNorm.

2.3. Архитектура нейронной сети

В экспериментах используется сверточная архитектура Jasper-5x3 [67], которая принимает на вход логарифм мел-спектрограммы из 64 компонент. Модель семейства Jasper-BxR состоит из B блоков по R подблоков. Один подблок – это последовательность из одномерной свертки вдоль оси времени, слоя BatchNorm, функции активации ReLU и слоя Dropout [68]. Внутри блока все свертки имеют одинаковый размер ядра и одинаковое число каналов и не меняют временное разрешение: шаг свертки (англ. stride) равен 1, а входная последовательность симметрично дополняется нулями на $(k-1)/2$ отсчетов, где k – размер ядра. Каждый блок снабжен остаточной связью (англ. residual connection): вход блока пропускается через свертку с ядром размера 1 и BatchNorm, после чего прибавляется к выходу BatchNorm последнего подблока, то есть до его функции активации.

Входной блок сети — это свертка из 64 в 256 каналов с ядром размера 11 и шагом 2, за которой следуют BatchNorm и ReLU; шаг 2 вдвое понижает временное разрешение. Далее идут пять блоков по три подблока, параметры которых приведены в таблице 1.

Таблица 1. Параметры блоков архитектуры Jasper-5x3

Блок	Размер ядра	Число каналов	Dropout
1	11	256	0.2
2	13	384	0.2
3	17	512	0.2
4	21	640	0.3
5	25	768	0.3

Выходной блок состоит из свертки из 768 в 896 каналов с ядром размера 29 и коэффициентом расширения (англ. dilation) 2, свертки из 896 в 1024 канала с ядром размера 1 (обе сопровождаются BatchNorm, ReLU и Dropout с вероятностью 0.4) и финальной свертки из 1024 каналов в $V = 29$ с ядром размера 1, которая для каждого временного шага выдает логиты по словарю. Свертки, за которыми следует BatchNorm, не имеют свободного члена (англ. bias): его роль берет на себя сдвигающий параметр самой нормализации. Свободный член есть только у финальной свертки, отображающей признаки в логиты словаря. Суммарно модель содержит около 107.7 млн параметров.

Используется без каких-либо специальных схем принятая в PyTorch [**69**] инициализация весов. Веса всех сверток разыгрываются из равномерного распределения $\mathcal{U}(-1/\sqrt{n}, 1/\sqrt{n})$, где n — произведение числа входных каналов на размер ядра; из того же распределения разыгрывается свободный член финальной свертки. Масштабирующие и сдвигающие параметры BatchNorm инициализируются единицами и нулями соответственно.

Аугментации данных намеренно не используются, чтобы в дальнейшем оценить влияние темпа обучения на устойчивость моделей (см. раздел 5). Стоит отметить, что рассматриваемая стандартная архитектура Jasper не является полностью масштабно-инвариантной, поэтому описанные эксперименты скорее воспроизводят раздел 6 из работы [**27**]. Такой выбор сделан в пользу приближения экспериментальной постановки к практическим сценариям обучения звуковых моделей.

2.4. Функция потерь и декодирование

Модель работает с символьным словарем из $V = 29$ токенов: пустой символ (англ. blank), пробел, апостроф и 26 строчных букв латинского

алфавита. Транскрипции приводятся к нижнему регистру, типографский апостроф заменяется на ASCII-вариант, все символы вне словаря (включая знаки препинания) удаляются, а последовательности пробелов схлопываются в один. Числа и сокращения в LJSpeech уже развернуты в словесную форму, поэтому дополнительная нормализация текста не требуется.

Сеть выдает последовательность логитов длины $T' = \lfloor (T-1)/2 \rfloor + 1$, где T – число временных шагов мел-спектрограммы; двукратное сокращение возникает из-за шага 2 во входном блоке. К логитам применяется логарифм softmax по оси словаря, при обучении минимизируется функция потерь CTC (англ. connectionist temporal classification) [70] с пустым символом, стандартная для задачи распознавания речи.

Декодирование выполняется жадно (англ. greedy decoding): на каждом временном шаге берется наиболее вероятный токен, затем подряд идущие одинаковые токены схлопываются в один и удаляются пустые символы. Ни лучевой поиск (англ. beam search), ни внешняя языковая модель не используются, поскольку интересны относительные различия между режимами темпа обучения, а не абсолютное качество распознавания.

Величины WER и CER вычисляются как расстояние Левенштейна между декодированной гипотезой и нормализованной эталонной транскрипцией на уровне слов и символов соответственно. Ошибка по всей выборке при этом считается не как среднее ошибок отдельных записей: расстояния Левенштейна суммируются по всем записям выборки, а полученная сумма делится на суммарное число слов (для CER – символов) во всех эталонных транскрипциях. Тем самым каждое ошибочно распознанное слово вносит в итоговую величину одинаковый вклад независимо от того, в какой записи оно встретилось.

Альтернативный вариант, при котором ошибка сначала считается для каждой записи отдельно, а затем усредняется по выборке, придавал бы коротким записям непропорционально большой вес. При любом замере качества сеть переводится в режим теста (англ. inference mode): Dropout отключается, а BatchNorm использует накопленные статистики.

2.5. Гиперпараметры оптимизации

В качестве алгоритма оптимизации используется SGD с инерцией $m = 0.9$ [71], коэффициентом регуляризации $\lambda = 10^{-4}$ и фиксированным размером пакета $B = 32$. Таким образом, одна эпоха обучения (то есть один полный проход по обучающей выборке) состоит из $\lfloor 650/32 \rfloor = 20$ итераций. И предобучение, и дообучение выполняются в течение 8000 эпох. Внутри каждой стадии темп обучения постоянен: ни разогрев (англ. warmup), ни затухание темпа обучения, ни обрезка градиентов (англ. gradient clipping) не применяются, поскольку любая из этих техник смешивала бы эффект исследуемой величины LR с эффектом расписания.

Для SWA новые контрольные точки сохраняются каждые 5 эпох после конца предобучения, т.е. 8005, 8010, и т.д. После усреднения весов статистики BatchNorm пересчитываются одним проходом по обучающей выборке, поскольку усредненному вектору весов не соответствуют статистики ни одной из усредняемых моделей [**56**]. Рассматриваются 16 значений PLR из списка $\{10^{k/4} | k = -12, \ldots, 3\}$ и 4 значения FLR $\{10^{k/2} | k = -6, \ldots, -3\}$. Обе сетки логарифмические: шаг по PLR равен $10^{1/4}$, по FLR – $10^{1/2}$.

Границы сетки PLR выбраны так, чтобы покрыть всю таксономию целиком: нижнее значение лежит заведомо внутри режима 1, а верхнее является наибольшим, при котором обучение еще не расходится в арифметике с плавающей точкой. Все значения FLR не превосходят порога сходимости, благодаря чему дообучение сходится при любом PLR, а верхнее из них совпадает с самим порогом, что позволяет наблюдать эффект «катапультирования».

Каждая конфигурация обучается трижды, и все приводимые далее статистики усреднены по этим трем запускам. Запуски различаются одновременно по трем источникам случайности: используется своя подвыборка обучающих данных из 650 записей (см. раздел 2.1), своя случайная инициализация весов сети и свой порядок формирования пакетов. Тем самым разброс между запусками отражает совокупное влияние как выбора обучающих данных, так и случайности самой процедуры оптимизации.

Выбросы при усреднении по трем запускам не отбрасывались и не подвергались усечению: во всех случаях приводятся выборочное среднее и стандартное отклонение по всем трем запускам, а конфигурации с большим разбросом обсуждаются отдельно (см. раздел 4).

2.6. Программно-аппаратная конфигурация

Все эксперименты реализованы на языке Python с использованием библиотек PyTorch и torchaudio. Один запуск отвечает одной паре (PLR, seed) на стадии предобучения и одной тройке (PLR, FLR, seed) на стадии дообучения. Каждому запуску выделяется один графический ускоритель NVIDIA V100, четыре ядра CPU и 24 ГБ оперативной памяти. Обе стадии предобучения занимают не более 3 часов. Суммарно проведено $16 \times 3 = 48$ запусков предобучения и $16 \times 4 \times 3 = 192$ запуска дообучения, не считая замеров качества, геометрии пространства весов и устойчивости моделей, которые требуют значительно меньшего вычислительного времени.

3. Обобщающая способность моделей

Прежде чем оценивать обобщающую способность моделей, предобученных с различными PLR, необходимо удостовериться, что воспроизводятся режимы обучения в таксономии [24].

На рисунке 1 изображена функция потерь CTC во время предобучения для разных значений PLR. Вертикальные пунктиры проведены по крайним

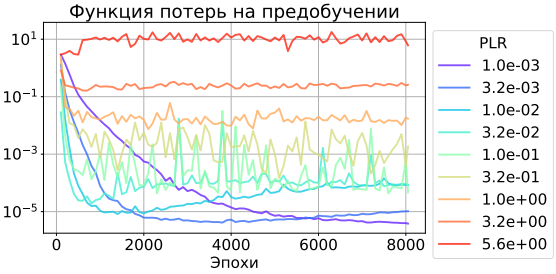

Рисунок 1. Функция потерь CTC

значениям PLR, отнесенным к подрежиму 2A: слева от левого пунктира лежит режим 1, справа от правого – подрежим 2B. Кривые функции потерь приведены для одного запуска

На рисунке 2 представлено среднее значение WER по итерациям в стабильной фазе обучения (то есть когда CTC перестает уменьшаться) для разных значений PLR. Показаны среднее и стандартное отклонение логарифма ошибки по трем запускам и последним 500 эпохам. Для корректного усреднения логарифма все значения WER $< 10^{-8}$ заменялись на 10^{-8}, что привело к ненулевым значениям WER в режиме 1.

Рисунок 2. Среднее значение WER по итерациям в стабильной фазе обучения

Соотношения между размером набора данных и количеством параметров модели оказалось достаточно, чтобы гарантировать перепараметризованность: низкие значения PLR достигают значения функции потерь

$< 10^{-3}$, а также (около-)нулевых значений WER. Тем не менее, многие значения PLR в промежутке $[5.6 \cdot 10^{-3}, 10^{-1}]$ спорадически достигают нулевого уровня WER в стабильной фазе, поэтому однозначно провести границу между режимами 1 и 2 только по значению метрики на обучающей выборке оказывается невозможным.

Вместо этого обе границы проводятся по геометрии пространства весов, а именно по барьеру ошибки на *обучающей* выборке между моделью, дообученной с высоким FLR, и точкой усреднения весов (см. раздел 4). Такой критерий не использует отложенные данные и меняется на порядки, а не на проценты, поэтому устойчив к разбросу между запусками. Обе границы отвечают качественной смене поведения этой одной величины: в режиме 1 барьер близок к 1, в подрежиме 2A обращается в ноль, а в подрежиме 2B возрастает вновь. Тем самым режим 1 отвечает PLR $\leqslant 10^{-2}$, подрежим 2A – промежутку $[1.8 \cdot 10^{-2}, 1]$, а подрежим 2B – значениям PLR $\geqslant 1.8$.

Качество на отложенной выборке служит при этом не основанием, а согласующимся наблюдением: минимум WER после предобучения приходится на первые значения подрежима 2A, непосредственно примыкающие к нижней границе. Это воспроизводит наблюдение работы [**27**], где наилучшее качество в точке предобучения также достигается сразу над порогом сходимости, а в практической постановке граница между режимами и вовсе проводится по этому максимуму качества с оговоркой о приближенном характере такого выбора.

Отметим, что обе границы локализованы лишь с точностью до шага сетки PLR, а конкретные значения представляют собой соглашение. Такая размытость закономерна: при обучении во всем пространстве весов норма параметров меняется по ходу оптимизации, поэтому фиксированному номинальному темпу обучения не отвечает фиксированный эффективный темп (см. раздел 6.3), и переходы между режимами оказываются более плавными, чем при обучении на сфере фиксированного радиуса. По этой же причине выбор границ не сопровождается статистическими критериями: раз переход не является скачком, «истинного» значения границы, которое можно было бы оценивать, не существует, а разрешение по оси PLR в любом случае ограничено шагом сетки.

Также стоит отметить, что при обучении не на сфере, а во всем пространстве весов, вообще говоря, режим 3 обычно не наблюдается (он оказывается нестабильным и приводит к расхождению процесса оптимизации в арифметике с плавающей точкой). Однако наибольшее рассмотренное значение PLR $= 5.6$ проявляет все признаки 3 режима, в частности, качество на уровне случайно инициализированной модели (WER равный 1 как на обучающей, так и на тестовой выборке). Большие же значения PLR действительно приводят к расходимости. Поэтому в дальнейшем анализе это значение PLR отнесено к режиму 2.

На рисунке 3 изображены WER на тестовой выборке при дообучении с различными значениями FLR (слева) и усреднении весов с различным числом моделей алгоритмом SWA (справа). Показаны среднее и стандартное отклонение по трем запускам. Проанализируем далее качество моделей

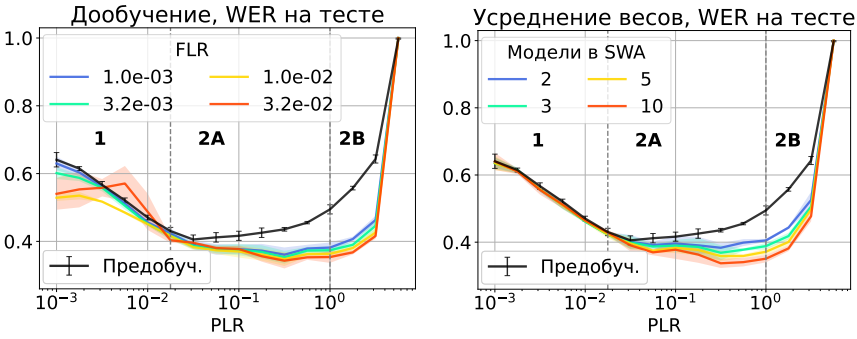

Рисунок 3. WER после дообучения и усреднения весов алгоритмом SWA: черная линия показывает качество после предобучения, вертикальные пунктиры проведены по крайним значениям PLR, отнесенным к подрежиму 2A.

для каждого режима в отдельности.

3.1. Режим 1

В режиме 1 качество на предобучении растет вместе с величиной PLR, что согласуется с результатами [24, 27]. При этом модель сходится в окрестность точки минимума и дальнейшее дообучение с низким FLR, равно как и усреднение весов, не способно снизить ошибку. Тем не менее, при значительном увеличении темпа обучения FLR $\gg$ PLR, наблюдается эффект «катапультирования» [16], при котором WER становится заметно ниже, чем в точке предобучения (см., например, PLR $= 10^{-3}$, FLR $= 3.2 \cdot 10^{-2}$). Однако ошибка все равно не дотягивает до уровня, наблюдаемого при обучении с таким же PLR из случайной инициализации. В итоге двухступенчатое расписание LR при начале обучения в режиме 1 лишено смысла: наилучшее качество достигается при постоянном высоком значении PLR.

3.2. Подрежим 2A

Как обсуждалось выше, подрежим 2A отвечает промежутку PLR $\in [1.8 \cdot 10^{-2}, 1]$, причем обе его границы проведены по поведению барьера ошибки на обучающей выборке (см. раздел 4). Качество после предобучения

достигает минимума в начале этого промежутка и далее снижается по мере роста PLR, однако именно в этом подрежиме достигается наилучшее качество после дообучения и усреднения весов. Верхняя граница согласуется и с поведением SWA: при больших PLR качество усреднения весов заметно ухудшается, хотя на трех запусках это ухудшение становится уверенно различимым лишь начиная с $\mathrm{PLR} = 3.2$.

При этом один из результатов работы [27] подтверждается только частично: различные значения FLR совпадают по качеству только для $\mathrm{PLR} \leqslant 10^{-1}$, далее же следует расслоение, свойственное подрежиму 2B из оригинальной статьи. Это наблюдение позволяет предположить, что четкое разграничение между подрежимами 2A и 2B избыточно для практических моделей, и некоторые особенности подрежима 2A (например, достижение оптимального качества после дообучения или эквивалентность FLR с точки зрения финального качества) могут наблюдаться для разных промежутков PLR. В дополнение, поведение алгоритма SWA аналогично исходным результатам: ошибка уменьшается при увеличении числа моделей, пока не насыщается.

3.3. Подрежим 2B

Поведение моделей в подрежиме 2B полностью аналогично оригинальной работе [27]: качество предобучения, дообучения и усреднения весов деградирует при увеличении PLR, пока не достигает уровня случайно инициализированной модели. Иными словами, при двухступенчатом расписании LR, начало обучения в подрежиме 2B оказывается менее удачным выбором по сравнению с подрежимом 2A. Примечательно также, что при предобучении с наибольшим $\mathrm{PLR} = 5.6$ качество дообучения остается на уровне случайного, т.е. это значение PLR приводит к менее удачной инициализации весов даже по сравнению со случайной инициализацией.

4. Геометрия пространства весов

4.1. Описание метрик

В предыдущем разделе установлено, как различные PLR влияют на обобщающую способность моделей. Следующий шаг – сравнить величины начального темпа обучения с точки зрения геометрии пространства весов, т.е. какие конкретно области этого пространства исследует SGD при различных PLR.

В частности, рассматриваются три конечные модели:

(1) дообученная с низким $\mathrm{FLR} = 10^{-3}$,

(2) дообученная с высоким $\mathrm{FLR} = 3.2 \cdot 10^{-2}$ и
(3) SWA из 5 моделей.

Между любыми двумя из них вычисляется угловое расстояние, т.е. угол между векторами весов; эта метрика показывает, насколько две модели удалены друг от друга *геометрически*. В дополнение вычисляется барьер ошибки, который замеряет величину ошибки на отрезке между двумя точками при интерполяции в пространстве весов, см. пример на рисунке 4.

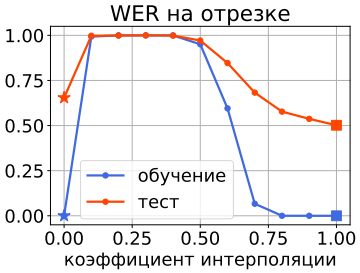

Рисунок 4. Пример интерполяции в пространстве весов между двумя моделями, обученными с низким $\mathrm{FLR} = 10^{-3}$ (левый конец, звезда) и высоким $\mathrm{FLR} = 3.2 \cdot 10^{-2}$ (правый конец, квадрат). Обе модели получены предобучением с $\mathrm{PLR} = 10^{-3}$.

Таким образом, эта метрика показывает, находятся ли две точки в окрестности одного и того же оптимума функции потерь. В литературе принято считать, что модели из окрестностей разных оптимумов являются *функционально* различными, о чем и сигнализирует высокий барьер ошибки [57–59]. Низкий барьер ошибки, напротив, характерен более похожим друг на друга линейно связным моделям [57–59]. В отличие от работ [27, 59], в которых барьер вычисляется сравнением ошибки внутри отрезка со значением на его концах, здесь используется более простая метрика $B(\theta_1, \theta_2)$, которая равна максимуму ошибки на отрезке:

$$B(\theta_1, \theta_2) = \sup_{\alpha \in [0,1]} E\big((1-\alpha)\theta_1 + \alpha\theta_2\big), \tag{1}$$

где θ_1, θ_2 обозначают два вектора весов, а $E(\cdot)$ – значение ошибки в некоторой точке пространства весов. Такое упрощение обусловлено поведением барьера для более высоких значений ошибок: при интерполяции между моделями с $\mathrm{WER} = 1$ то же самое значение наблюдается и на отрезке между ними, что формально говорит о нулевом барьере. Поэтому для согласованного сравнения всех PLR поправка на значения в конечных точках не делается, т.е. в примере выше значение барьера будет равно 1. Барьер ошибки вычисляется как на обучающих, так и на тестовых данных.

4.2. Результаты

Результат замеров метрик, описанных выше, приведен на рисунке 5.

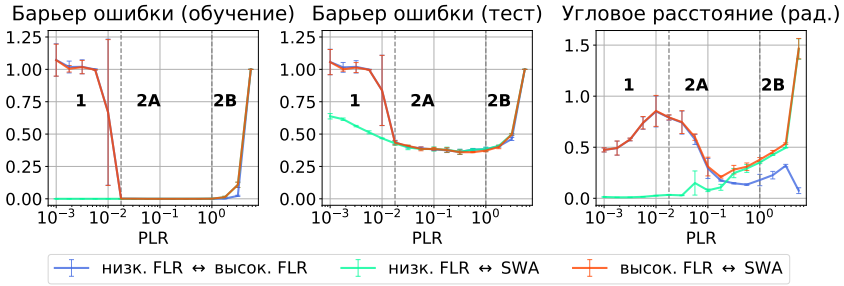

Рисунок 5. Геометрия пространства весов между решениями, полученными с низким FLR $= 10^{-3}$, высоким FLR $= 3.2 \cdot 10^{-2}$ и SWA из 5 моделей. Вертикальные пунктиры проведены по крайним значениям PLR, отнесенным к подрежиму 2A: слева от левого пунктира лежит режим 1, справа от правого — подрежим 2B. Угловое расстояние вычисляется в радианах. Показаны среднее и стандартное отклонение по трем запускам.

Именно барьер на обучающей выборке между моделью, дообученной с высоким FLR, и точкой SWA служит тем критерием, по которому в разделе 3 проведены границы между режимами: нижняя отвечает его исчезновению, верхняя — повторному появлению. Отдельно отметим поведение в точке PLR $= 10^{-2}$, последней перед нижней границей: здесь этот барьер имеет крайне высокую дисперсию по запускам. В одном из трех запусков он оказывается на два порядка ниже, чем в двух других ($2 \cdot 10^{-2}$ против величин, близких к 1), хотя и не опускается до уровня 10^{-3}, характерного для подрежима 2A. Поскольку запуски различаются в том числе обучающей подвыборкой (см. раздел 2.5), точное положение нижней границы зависит от конкретных обучающих данных.

При предобучении в режиме 1 барьер ошибки на обучающей выборке нулевой, а угловое расстояние между дообучением с низким FLR и точкой SWA практически нулевое. Эти показатели свидетельствуют, что две модели фактически представляют одно и то же решение: они имеют сходное качество (см. рисунок 3) и расположены близко друг к другу в пространстве весов.

Барьер ошибки на тестовой выборке также остается относительно невысоким. При этом дообучение с высоким FLR перемещает модель в другую область пространства весов, что подтверждается как более высоким

угловым расстоянием, так и значениями барьеров ошибки, приблизительно равными 1[2]. Такое поведение является прямым подтверждением эффекта «катапультирования» [16], при котором более высокий FLR не может удержаться в остром минимуме, найденном при помощи низкого PLR.

При рассмотрении PLR из подрежима 2A наблюдаются нулевые барьеры на обучающей выборке между всеми тремя моделями. Барьеры на тестовой выборке также оказываются минимальными среди всех промежутков PLR. Таким образом, все три модели оказываются линейно связными и расположенными в одной и той же области пространства весов с низкими значениями ошибки, и основной вывод работы [27] подтверждается.

Однако все три представляют собой разные точки внутри этой области, поскольку значения углового расстояния оказываются отделены от нуля. Более того, расстояние между точками дообучения и SWA возрастает по мере увеличения PLR ближе к подрежиму 2B, т.е. различия между этими решениями становятся более значимыми.

Наконец, при переходе в подрежим 2B все три метрики возрастают, и модели перестают быть линейно связными. Примечательно, что расстояние между моделями, дообученными с различными FLR, растет медленнее, чем от обеих моделей до точки SWA. Более того, для наибольшего PLR $= 5.6$ и угловое, и евклидово расстояния между дообученными моделями резко падают: угловое расстояние составляет около 0.06 радиан (против 0.15–0.33 при соседних PLR $\in [1, 3.16]$), а ℓ_2-расстояние равно 0.7 против значений от 15 до 70 у соседних PLR, то есть уменьшается почти на два порядка.

Вероятно, это вызвано более высокой нормой весов, которая получается при обучении с высоким PLR: в масштабно-инвариантных сетях более высокая норма весов приводит к более низким нормам градиентов, что не позволяет дообучению отойти далеко от точки инициализации. Несмотря на близость в пространстве весов, аналогичную подрежиму 2A, дообученные модели показывают высокий WER, поэтому предобучение со столь высоким PLR не может быть признано удачным с точки зрения качества.

[2] В отличие от точности, WER может принимать значения > 1, так какэта метрика основана на расстоянии Левенштейна.

5. Устойчивость моделей

5.1. Анализ компонент мел-частотного кепстра

В предыдущем разделе мы удостоверились, что геометрия пространства весов моделей для распознавания речи в подрежиме 2A соответствует поведению, которое наблюдается в моделях классификации изображений. Следующий шаг – установить, какие особенности выучивания признаков из данных отличают подрежим 2A.

В оригинальной работе [27] рассматривается двумерное преобразование Фурье над изображениями для частотного анализа, однако здесь мел-спектрограммы, которые подаются на вход в нейросеть, уже являются спектральным представлением. По этой причине обратимся к мел-частотному кепстру [1] (англ. mel-frequency cepstrum), который получается применением дискретного косинусного преобразования к мел-спектрограмме, т.е. в сущности является «спектром спектра».

Кепстральные коэффициенты (англ. mel-frequency cepstral coefficients, MFCC) являются более плотным представлением для аудиозаписей по сравнению с мел-спектрограммой, поэтому далее рассматриваются только старшие коэффициенты в отличие от работы [27], где используется весь двумерный Фурье-спектр изображений. Более предметно, рассматриваются выборки тестовых аудиозаписей, в которых сохранены только несколько кепстральных коэффициентов, а именно $\{(0, 2k-1, 2k) | k = 1, \ldots, 4\}$.

Иными словами, первая выборка содержит коэффициенты $(0, 1, 2)$, вторая – $(0, 3, 4)$, и т.д. Остальные коэффициенты приравниваются к нулю, а затем кепстр конвертируется обратно в мел-спектрограмму с помощью обратного дискретного косинусного преобразования. Нулевой коэффициент сохраняется во всех выборках, поскольку он кодирует суммарную энергию одного временного шага мел-спектрограммы, и его удаление приведет к сдвигу значений спектрограммы и нерепрезентативному качеству моделей.

Далее, для всех четырех выборок с различными кепстральными коэффициентами, WER замеряется на четырех моделях для каждого значения PLR:

(1) предобученная модель;
(2) модель, дообученная с низким FLR;
(3) модель, дообученная с высоким FLR;
(4) точка усреднения весов SWA.

Результаты представлены на рисунке 6.

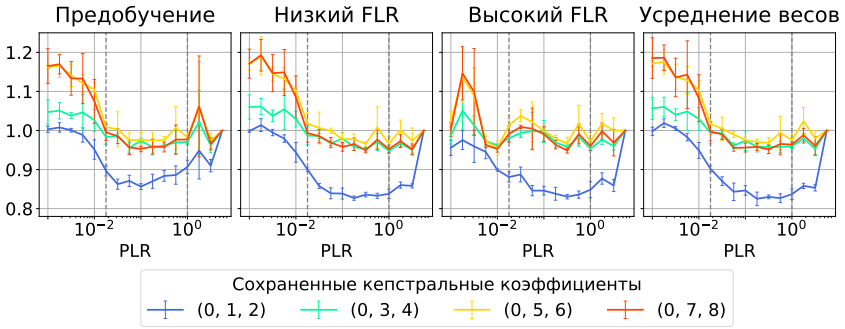

Рисунок 6. WER на тесте для различных компонент мел-частотного кепстра

Вертикальные пунктиры проведены по крайним значениям PLR, отнесенным к подрежиму 2A: слева от левого пунктира лежит режим 1, справа от правого – подрежим 2B. Показаны среднее и стандартное отклонение по трем запускам.

Ошибка моделей оказывается значительно выше, чем при замерах на чистых данных (см. рисунок 3), т.е. модели так или иначе опираются на более широкое подмножество кепстра, нежели рассмотренные тройки коэффициентов. Тем не менее, первые два коэффициента оказываются наиболее значимыми: только с ними часть моделей показывает ошибку, заметно меньшую 1. При этом качество по следующим тройкам коэффициентов лежит в пределах WER $\approx 0.95 - 1.05$, то есть на уровне случайного предсказания или хуже него, поэтому содержательная интерпретация различий между режимами по ним неправомерна. Дальнейший анализ проводится только по тройке $(0, 1, 2)$.

Основное различие проходит по порогу сходимости, а не по границе между подрежимами 2A и 2B. При предобучении в режиме 1 с малыми PLR ошибка на тройке $(0, 1, 2)$ находится на уровне случайного предсказания (WER ≈ 1.0), затем монотонно снижается по мере роста PLR, достигает WER $\approx 0.83 - 0.86$ уже на нижней границе подрежима 2A и далее выходит на плато.

Существенно, что это плато сохраняется и в подрежиме 2B: при PLR $= 1.78$ и PLR $= 3.16$ ошибка на усеченном кепстре (0.86 в обоих случаях) практически совпадает со значениями внутри 2A, хотя качество этих же моделей на чистых данных заметно хуже (см. рисунок 3). Исключение составляет лишь вырожденное значение PLR $= 5.6$, для которого ошибка равна 1 при любом входе. Отметим также, что переход через порог сходимости не является резким: при PLR $= 1.78 \cdot 10^{-2}$, которое отнесено к режиму 1, ошибка уже составляет 0.90.

Таким образом, этот эксперимент воспроизводит выводы работы [27] лишь частично. Он подтверждает более общее наблюдение о том, что высокий темп обучения приводит к более «экономному» представлению данных [46, 47, 49]: модели, предобученные выше порога сходимости, сохраняют часть распознающей способности при сокращении входа до первых двух кепстральных коэффициентов, тогда как модели режима 1 полностью ее теряют.

Отдельно проверим, не объясняется ли наблюдаемое различие просто разным качеством моделей. Для этого сопоставим предобученные модели с совпадающим WER на чистых данных, но по разные стороны от порога сходимости. Модели с PLR $= 10^{-3}$ (режим 1) и PLR $= 3.16$ (подрежим 2B) дают WER 0.641 и 0.642 соответственно, то есть неразличимы на чистом входе, тогда как на тройке $(0, 1, 2)$ их ошибки составляют 1.001 и 0.909; диапазоны по трем запускам при этом не пересекаются ($[0.991, 1.010]$ против $[0.896, 0.927]$). Аналогично ведет себя и пара PLR $= 1.8 \cdot 10^{-3}$ и PLR $= 1.78$ (1.004 против 0.942 при чистых 0.615 и 0.570).

Тем самым устойчивость к усечению кепстра не сводится к общему уровню качества модели: при равном качестве на полном входе модели, предобученные выше порога сходимости, сохраняют заметно больше распознающей способности на первых кепстральных коэффициентах. Однако устойчивость к усечению кепстра не является отличительным свойством именно подрежима 2A: по этому показателю 2A и 2B неразличимы, и его нельзя использовать как диагностический признак для проведения границы между ними.

Стоит также подчеркнуть, что речь идет о сравнении величин WER ≈ 0.83 и WER ≈ 1: обе соответствуют практически неразборчивому распознаванию, и наблюдаемое различие следует трактовать как качественный индикатор структуры выученных признаков, а не как свидетельство работоспособности моделей на усеченном входе.

5.2. Применение аугментаций к мел-спектрограммам

В дополнение к кепстральному анализу изучается устойчивость моделей при применении аугментаций непосредственно к аудиозаписям и спектрограммам. Рассматривается несколько типов аугментаций:

1) **Изменение тональности аудиозаписи.** Поскольку данные содержат записи только одного диктора, модели никак не мотивированы быть устойчивыми к высоте голоса. Этот эксперимент призван проверить, насколько различные PLR чувствительны к изменению тона.
2) **Добавление нормального шума.** Устойчивость различных PLR дополнительно проверяется добавлением к аудиозаписям белого гауссовского шума.

3) **Маскирование компонентов спектрограммы.** Третий эксперимент проверяет, насколько модели полагаются на различные компоненты мел-спектрограммы при распознавании речи. В частности, зануляются несколько самых высоких и самых низких компонент.

Результаты применения аугментаций к аудиозаписям представлены на рисунке 7.

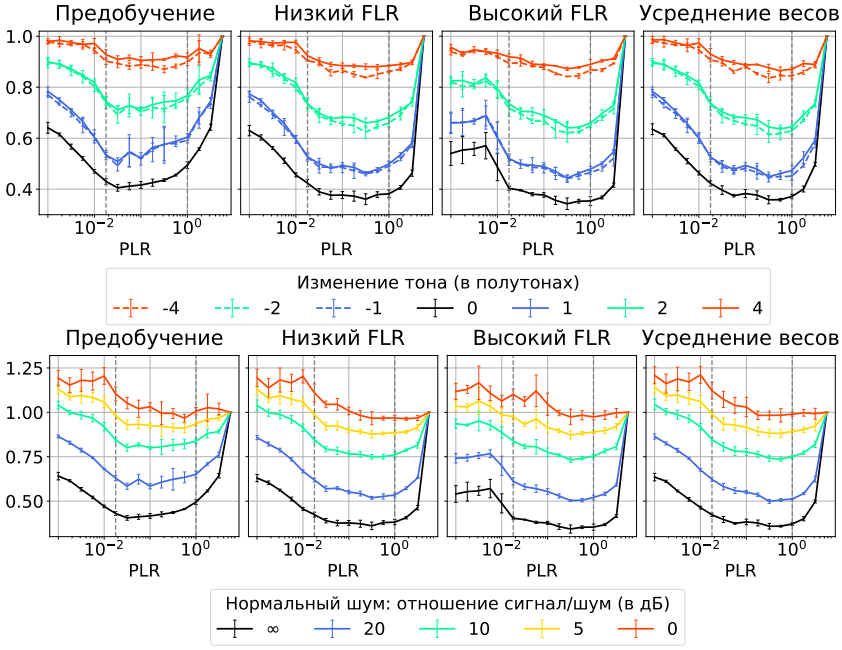

Рисунок 7. WER на тесте при применении различных аугментаций

Показаны среднее и стандартное отклонение по трем запускам при изменении тона и добавлении нормального шума. Черные линии соответствуют качеству на чистых данных. Вертикальные пунктиры проведены по крайним значениям PLR, отнесенным к подрежиму 2A: слева от левого пунктира лежит режим 1, справа от правого – подрежим 2B.

Наблюдается монотонный рост WER при увеличении интенсивности аугментаций, при этом преимущество предобучения в подрежиме 2A сохраняется. Показательно, что понижение и повышение тона на одну и ту же величину приводит к сходным уровням ошибки.

В целом, качество моделей на аугментированных данных очень хорошо коррелирует с качеством на чистых аудиозаписях. Из этого следует важная оговорка: изменение тона и добавление шума не выявляют какой-либо специфической устойчивости моделей 2A, а лишь воспроизводят уже установленное в разделе 3 упорядочение PLR по качеству. Содержательные выводы о структуре выученных признаков поэтому опираются на кепстральный анализ и маскирование мел-компонент, где рассматриваемые преобразования нацелены на конкретные участки признакового описания.

На рисунке 8 показано качество моделей при маскировании спектрограмм.

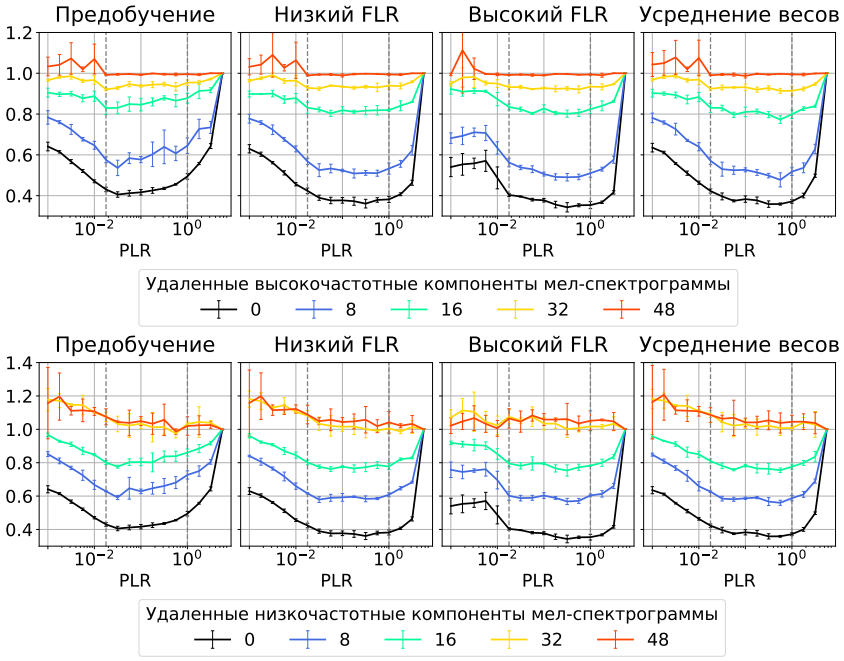

Рисунок 8. WER на тесте при маскировании компонент мел-спектрограммы

Черные линии соответствуют качеству на чистых данных. Вертикальные пунктиры проведены по крайним значениям PLR, отнесенным к подрежиму 2A: слева от левого пунктира лежит режим 1, справа от правого – подрежим 2B. Показаны среднее и стандартное отклонение по трем запускам.

И высокие, и низкие частоты оказываются важными для моделей, поскольку удаление 8 компонент с обоих концов приводит к значительному росту ошибки.

Чтобы сравнить режимы между собой, сопоставим наилучшие значения WER, достигаемые внутри режима 1 и внутри подрежима 2A, для моделей, дообученных с низким FLR $= 10^{-3}$; на чистых данных этот разрыв составляет 0.09. При удалении 8 компонент он сохраняется независимо от того, с какого конца банка фильтров они удалены: 0.08 для верхних компонент и 0.12 для нижних.

Иными словами, слабое маскирование преимущества 2A не затрагивает вовсе, а при удалении нижних компонент даже слегка увеличивает его. Различие проявляется лишь при удалении 16 и более компонент: при удалении нижних компонент разрыв сокращается до 0.07, а при 32 – до 0.04, тогда как при удалении верхних он остается на уровне 0.08–0.11.

Для предобученных моделей и моделей, дообученных с высоким FLR, наблюдается та же картина; для точки SWA сужение разрыва проявляется только при удалении 32 нижних компонент. Это позволяет предположить, что более высокое качество в подрежиме 2A связано в первую очередь с низкочастотной частью спектрограммы.

Отметим, однако, ограниченность этого вывода: при удалении 32 компонент обе конфигурации выходят на уровень практически случайного предсказания (WER $\geqslant 0.93$ при удалении нижних компонент и WER > 1 при удалении верхних). Поэтому сокращение разрыва в этой точке объясняется в том числе насыщением метрики, а не только потерей значимых для моделей 2A признаков.

Тем самым устойчивое различие опирается на одну промежуточную точку (16 компонент), и его подтверждение требует более тонкой сетки по числу удаляемых компонент. С учетом этой оговорки наблюдение дополняет кепстральный анализ и указывает на то, что преимущество подрежима 2A локализовано по частотной оси.

6. Обсуждение результатов

6.1. Устойчивость подрежима 2A к смене задачи и постановки обучения

Один из главных выводов проведенного исследования состоит в том, что подрежим 2A наблюдается в существенно далекой от оригинальной работы [27] экспериментальной постановке. Сменены и сама задача (распознавание речи вместо классификации изображений), и парадигма обучения (стандартный SGD без явной масштабной инвариантности вместо проективного SGD на сфере).

Тем не менее, два ключевых свойства подрежима 2A воспроизводятся в полном объеме:

1) *превосходство по качеству после дообучения и усреднения весов* (см. раздел 3) и
2) *линейная связность решений* (см. раздел 4).

Третье свойство, *специализация выученных признаков* (см. раздел 5), воспроизводится в более мягкой форме: смещение опоры моделей к нижним кепстральным коэффициентам отделяет режим 2 от режима 1, но внутри режима 2 не разделяет подрежимы 2A и 2B, а частотная локализация преимущества 2A проявляется лишь при удалении 16 и более нижних мел-компонент.

Следовательно, подрежим 2A, по-видимому, не сводится ни к особенностям масштабно-инвариантной постановки, ни к специфике задачи классификации изображений. Полученные наблюдения согласуются с тем, что за ним стоит более общее свойство стохастической оптимизации, хотя набор его проявлений заметно меняется от постановки к постановке. В этом смысле результаты работы [27] получают дополнительную поддержку: обучение на сфере удобно для строгого анализа, но основные эффекты 2A наблюдаются и без этого ограничения.

6.2. Граница между подрежимами 2A и 2B

В рассмотренной постановке различные диагностические признаки подрежима 2A перестают выполняться при разных значениях PLR. Эквивалентность различных значений FLR по итоговому качеству дообучения наблюдается только для $\mathrm{PLR} \leqslant 10^{-1}$ (см. раздел 3), тогда как линейная связность решений в пространстве весов (см. раздел 4) и высокое качество усреднения весов (см. раздел 3) сохраняются вплоть до $\mathrm{PLR} = 1$. Формальная граница между 2A и 2B, проведенная по геометрии пространства весов, оказывается заметно выше, чем та, что напрашивается из эквивалентности качества дообучения при различных FLR.

Таким образом, один из результатов [27] воспроизводится лишь частично: в рассмотренной постановке эквивалентность различных FLR не является характеристикой всего подрежима 2A и сохраняется только в его нижней половине.

Среди рассмотренных диагностических признаков наиболее надежен барьер ошибки на обучающей выборке, по которому и проведены обе границы в разделе 3: он не опирается на отложенные данные и меняется на порядки, а не на проценты, поэтому устойчив к разбросу между запусками. Остальные признаки – качество после дообучения, эквивалентность различных FLR и качество усреднения весов – согласуются с проведенными по нему границами, но по отдельности локализуют переход менее уверенно.

По-видимому, в практической постановке границу между 2A и 2B естественнее рассматривать не как единую точку, а как совокупность нескольких независимых переходов, каждый из которых отвечает за отдельное свойство динамики оптимизации. Отсюда следует прикладной вывод: вместо фиксации одного «оптимального» значения PLR внутри 2A имеет смысл использовать расписания темпа обучения, последовательно проходящие весь спектр значений режима. Такое расписание позволяет одновременно задействовать все желаемые свойства 2A, поскольку каждое из этих свойств сохраняется в своем промежутке PLR.

Двухступенчатая схема «предобучение с высоким PLR $\to$ дообучение с низким FLR», использованная выше, служит простейшим примером такого расписания, а стандартные плавно убывающие расписания LR реализуют ту же идею более непрерывным образом и не требуют столь кропотливого подбора начального темпа обучения. Тем самым жесткая привязка к одному постоянному значению PLR внутри подрежима 2A не требуется, и размытость границы между 2A и 2B не является препятствием для практического применения: общепринятые расписания темпа обучения полученные наблюдения не отменяют, а скорее дополнительно обосновывают.

6.3. Скрытая масштабная инвариантность в практической постановке

Наиболее нетривиальное наблюдение геометрического анализа в разделе 4 состоит в аномальном сжатии расстояний между дообученными моделями при наибольшем PLR $= 5.6$: угловое расстояние уменьшается в несколько раз, а ℓ_2-расстояние падает практически на два порядка относительно соседних PLR. Естественное объяснение этому таково: за счет слоев нормализации существенная часть параметров модели эффективно масштабно-инвариантна, а высокий PLR раздувает норму весов этих слоев. Раздутая норма, в свою очередь, подавляет эффективный темп обучения при дообучении, и никакое из рассмотренных значений FLR фактически не сдвигает модель с точки инициализации.

Отсюда следует, что «практическая» постановка, противопоставляемая здесь (а также в разделе 6 оригинальной работы [27]) МИ-обучению на сфере, не является в полной мере не-масштабно-инвариантной. Скорее, она представляет собой гибрид, в котором масштабно-инвариантная динамика присутствует неявно и, как показывает пример PLR = 5.6, может проявляться в неожиданных местах.

Из этого вытекают два практических вывода.

Во-первых, любые сравнения между МИ-обучением на сфере и обучением во всем пространстве весов осмысленны только при явном контроле норм весов слоев.

Во-вторых, при мониторинге обучения имеет смысл логировать нормы весов масштабно-инвариантных слоев, поскольку это почти бесплатная операция, а также она позволяет напрямую отслеживать эффективный, а не только номинальный темп обучения.

Расписания, которые контролируют эффективный темп обучения, а не только номинальный, являются перспективным направлением для будущих исследований.

6.4. Частотная локализация специализации признаков и переносимый инструмент анализа

Кепстральный анализ и маскирование компонент мел-спектрограммы в разделе 5 дают два дополняющих друг друга наблюдения о выученных моделями 2A признаках. Кепстральный анализ показывает, что модели, предобученные выше порога сходимости, лучше моделей режима 1 сохраняют качество распознавания при сокращении входа до нескольких первых кепстральных коэффициентов, а маскирование компонент мел-спектрограммы дополнительно локализует преимущество 2A в нижней части банка мел-фильтров: удаление 16 нижних компонент заметно сокращает разрыв с режимом 1, тогда как удаление такого же числа верхних компонент оставляет его на прежнем уровне. Именно последняя локализация допускает конкретное практическое предсказание относительно узкополосного аудио.

Стандартный телефонный речевой канал (POTS, G.711, GSM-narrowband) обрезает частоты вне полосы $\approx$ 300–3400 Гц. В используемом банке из 64 мел-фильтров, покрывающем полосу от 0 до 8000 Гц (см. раздел 2.2), это соответствует потере примерно 8 нижних и 20 верхних компонент. Существенно, что при этом теряется в том числе тот низкочастотный участок, с которым связано преимущество моделей 2A.

Если наблюдаемая частотная локализация действительно отражает структуру выученных признаков, а не особенности конкретной сетки маскирования, то преимущество 2A по WER на телефонной аудиозаписи должно заметно сузиться, а PLR из верхней части режима 1 или с границы между режимом 1 и подрежимом 2A может оказаться практически предпочтительнее «оптимального на чистом входе». Эта гипотеза может быть проверена без переобучения: достаточно применить полосовой фильтр 300–3400 Гц к тестовым записям и заново замерить WER на уже сохраненных моделях, что оставлено для дальнейшей работы.

Отдельно стоит подчеркнуть методологический вклад раздела 5, выходящий за рамки исследования темпа обучения. В оригинальной работе [**27**] анализ выученных признаков строится на двумерном Фурье-преобразовании изображений. Для аудио-моделей, работающих с мел-спектрограммой, естественным аналогом становится усечение мел-частотного кепстра.

Предложенный инструмент не привязан ни к архитектуре Jasper, ни к CTC-функции потерь, поэтому применим к произвольной модели, принимающей на вход спектрограмму, в том числе к более современным архитектурам Conformer [**72**], wav2vec 2.0 [**73**] или HuBERT [**74**]. Это делает кепстральное усечение самостоятельным методологическим вкладом описанного исследования, полезным и вне контекста исследования темпа обучения.

6.5. Ограничения текущего анализа и естественные продолжения

Описанное исследование сознательно упрощено сразу в нескольких измерениях, и каждое из этих упрощений задает естественное направление для дальнейшей работы.

Прежде всего отметим, что все выводы получены на одном корпусе, одной архитектуре и при фиксированных размере пакета и длительности расписания, а статистики усреднены по трем запускам. Этого достаточно, чтобы констатировать перенос эффектов 2A на распознавание речи, но недостаточно, чтобы говорить об их универсальности: последнее потребовало бы систематического перебора задач, архитектур и режимов параметризации.

Далее, обучение проводилось на 5%-подвыборке LJSpeech. Воспроизведение на полной выборке LJSpeech или на более масштабном наборе данных LibriSpeech [**75**] позволит понять, сохраняются ли наблюдаемые свойства 2A в реалистично недопараметризованных сценариях, характерных для практического обучения моделей распознавания речи.

Отдельного упоминания заслуживает ограниченность инструментов анализа признаков. Усечение кепстра и маскирование мел-компонент выполнены на грубой сетке, а получаемые при этом значения WER близки к уровню случайного предсказания, поэтому выводы раздела 5 следует считать качественными индикаторами, а не количественными оценками степени специализации моделей. Более тонкая сетка по числу удаляемых компонент, а также прямые меры разреженности представлений позволили бы проверить, действительно ли подрежим 2A отличается от остальной части режима 2 по структуре выучиваемых признаков.

Кроме того, аугментации при обучении намеренно не использовались, поэтому открытым остается вопрос о том, взаимодействуют ли эффекты 2A и SpecAugment [**76**] суммарно (в этом случае их разумно комбинировать) или же конкурируют, покрывая один и тот же класс регуляризационных эффектов (в этом случае 2A можно интерпретировать как «бесплатный SpecAugment» с точки зрения итогового качества).

Отметим, какого эффекта от обоих изменений следует ожидать: и увеличение выборки, и аугментации снижают степень перепараметризации, поэтому точная подгонка обучающей выборки перестает достигаться, а вместе с ней теряет смысл и использованный выше критерий проведения границ, естественной заменой которому становится выход функции потерь на плато. Вопрос о том, как при этом сместятся сами границы между режимами, остается открытым.

Наконец, современные акустические модели, упомянутые в предыдущем разделе (в частности, wav2vec 2.0 и HuBERT), сначала предобучаются на большом аудиокорпусе без разметки, а затем дообучаются с CTC-функцией потерь на целевой задаче. Открытым остается вопрос о том, воспроизводится ли таксономия трех режимов PLR применительно к темпу обучения на стадии дообучения таких моделей, когда начальная точка задается не случайной инициализацией, а обученным без учителя представлением.

Заключение

Эмпирическая методология работы [27] расширена на задачу распознавания речи: сверточная акустическая модель Jasper обучалась на подвыборке LJSpeech с использованием функции потерь CTC. Полученные результаты подтверждают, что подрежим 2A наблюдается в существенно более далекой от оригинальной работы постановке: как при переходе от классификации изображений к распознаванию речи, так и при отказе от масштабно-инвариантного обучения на сфере в пользу стандартного SGD.

Наиболее уверенно переносятся свойства, связанные с обобщающей способностью и геометрией пространства весов; картина выучиваемых признаков воспроизводится в более мягкой форме. Тем самым эффекты 2A не сводятся ни к сферической постановке, ни к задаче классификации изображений, хотя вопрос об их универсальности потребует проверки на других корпусах и архитектурах.

Помимо непосредственного воспроизведения результатов получено несколько неочевидных наблюдений.

Различные диагностические признаки подрежима 2A перестают выполняться при разных значениях PLR, поэтому границу между 2A и 2B естественнее рассматривать как совокупность независимых переходов, а не как единую точку. Это наблюдение указывает на пользу расписаний темпа обучения, последовательно проходящих весь диапазон 2A, вместо фиксации одного «оптимального» значения.

Обнаружено также, что даже в практической постановке масштабно-инвариантная динамика проявляется неявно: при наибольших PLR раздутая норма весов подавляет эффективный темп обучения на стадии дообучения.

Наконец, предложенное усечение мел-частотного кепстра показало, что модели, предобученные выше порога сходимости, лучше сохраняют качество распознавания речи при сокращении входа до нескольких первых кепстральных коэффициентов, хотя подрежимы 2A и 2B по этому признаку не различаются.

Маскирование компонент мел-спектрограммы дополнительно локализовало преимущество 2A в нижней части банка мел-фильтров, что дает проверяемое предсказание для узкополосного аудио.

Систематическая проверка перечисленных наблюдений на более крупных корпусах, современных архитектурах и с использованием аугментаций остается предметом дальнейшего исследования.

Список использованных источников

[1] Заковряшин А. С., Малинин П. В., Лепендин А. А. *Применение распределений мел-частотных кепстральных коэффициентов для голосовой идентификации личности* // Известия Алтайского государственного университета.– 2014.– № 1-1 (81).– С. 156–160. ↑334

[2] Krizhevsky A., Sutskever I., Hinton G. E. *ImageNet classification with deep convolutional neural networks* // Communications of the ACM.– 2017.– Т. **60**.– № 6.– С. 84–90. ↑316

[3] Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. *An image is worth 16x16 words: transformers for image recognition at scale* // *9th International Conference on Learning Representations*, ICLR 2021 (Virtual Event, Austria, 3–7 May 2021).– Curran Associates, Inc..– 2021.– ISBN 9798331321949.– 22 с. arXiv 2010.11929 ↑316

[4] Hinton G., Deng L., Yu D., Dahl G. E., Mohamed A.-r., Jaitly N., Senior A., Vanhoucke V., Nguyen P., Sainath T. N., Kingsbury B. *Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups* // IEEE Signal Processing Magazine.– 2012.– Т. **29**.– № 6.– С. 82–97. ↑316

[5] Radford A., Kim J. W., Xu T., Brockman G., McLeavey C., Sutskever I. *Robust speech recognition via large-scale weak supervision* // *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA, 23–29 July 2023), PMLR.– т. **202**.– ICML.– ISBN 9781713889182.– С. 28492–28518. arXiv 2212.04356 ↑316

[6] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, Ch. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, Ch. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei *Language models are few-shot learners* // *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 6–12, 2020), Advances in Neural Information Processing Systems.– т. **33**, Red Hook: Curran Associates Inc..– 2020.– ISBN 978-1-7138-2954-6.– С. 1877–1901. arXiv 2005.14165 ↑316

[7] Touvron H., Lavril T., Izacard G., Martinet X., Lachaux M.-A., Lacroix T., B. Rozière, Goyal N., Hambro E., Azhar F., Rodriguez A., Joulin A., Grave E., Lample G. *LLaMA: open and efficient foundation language models*.– 2023.– 27 с. arXiv 2302.13971 ↑316

[8] Jumper J., Evans R., Pritzel A., Green T., Figurnov M., Ronneberger O., Tunyasuvunakool K., Bates R., A. Žídek, Potapenko A., Bridgland A., Meyer C., Kohl S. A. A., Ballard A. J., Cowie A., Romera-Paredes B., Nikolov S., Jain R., Adler J., Back T., Petersen S., Reiman D., Clancy E., Zielinski M., Steinegger M., Pacholska M., Berghammer T., Bodenstein S., Silver D., Vinyals O., Senior A. W., Kavukcuoglu K., Kohli P., Hassabis D. *Highly accurate protein structure prediction with AlphaFold* // Nature.– 2021.– Т. **596**.– № 7873.– С. 583–589. ↑316

[9] LeCun Y., Bengio Y., Hinton G. *Deep learning* // Nature.– 2015.– Т. **521**.– № 7553.– С. 436–444. doi ↑316

[10] L. Zheng, K. Zhou, P. Liang *On the opportunities and risks of foundation models.*– 2021.– 214 с. arXiv 2108.07258 doi ↑316

[11] Rudin C. *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead* // Nature Machine Intelligence.– 2019.– Т. **1**.– № 5.– С. 206–215. doi ↑316

[12] Rudin C., Chen C., Chen Z., Huang H., Semenova L., Zhong C. *Interpretable machine learning: fundamental principles and 10 grand challenges* // Statistics Surveys.– 2022.– Vol. **16**.– Pp. 1–85. doi ↑316

[13] Dauphin Y. N., Pascanu R., Gulcehre C., Cho K., Ganguli S., Bengio Y. *Identifying and attacking the saddle point problem in high-dimensional non-convex optimization* // *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems.*– V. 2 (Montreal, Canada, 8–13 December 2014), Cambridge: MIT Press.– 2014.– ISBN 9781510800410.– Pp. 2933–2941. doi arXiv 1406.2572 ↑316

[14] Li H., Xu Z., Taylor G., Studer C., Goldstein T. *Visualizing the loss landscape of neural nets* // *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montreal, Canada, 3–8 December 2018), Advances in Neural Information Processing Systems.– vol. **31**, Red Hook: Curran Associates, Inc..– 2018.– ISBN 9781510884472.– Pp. 6391–6401. doi arXiv 1712.09913 ↑316

[15] S. Jastrzębski, Kenton Z., Arpit D., Ballas N., Fischer A., Bengio Y., Storkey A. *Three factors influencing minima in SGD.*– 2017.– 14 pp. arXiv 1711.04623 doi ↑316, 318, 321

[16] Lewkowycz A., Bahri Y., Dyer E., Sohl-Dickstein J., Gur-Ari G. *The large learning rate phase of deep learning: the catapult mechanism.*– 2020.– 25 pp. arXiv 2003.02218 doi ↑316, 318, 329, 333

[17] Cohen J. M., Kaur S., Li Y., Kolter J. Z., Talwalkar A. *Gradient descent on neural networks typically occurs at the edge of stability* // *9th International Conference on Learning Representations*, ICLR 2021 (Virtual Event, Austria, 3–7 May 2021).– Curran Associates, Inc..– 2021.– ISBN 979-8-3313-0008-1.– 84 pp. doi ↑316, 318, 319

[18] Bengio Y. *Practical recommendations for gradient-based training of deep architectures* // *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science.– vol. **7700**, 2nd ed., Berlin–Heidelberg: Springer.– 2012.– ISBN 978-3-642-35288-1.– Pp. 437–478. doi ↑316

[19] Smith S. L., Le Q. V. *A Bayesian perspective on generalization and stochastic gradient descent* // *6th International Conference on Learning Representations*, ICLR 2018 (Vancouver, BC, Canada, April 30–May 3, 2018).– 2018.– ISBN 9798331321963.– 13 pp. doi ↑316, 318

[20] Smith L. N. *Cyclical learning rates for training neural networks* // *2017 IEEE Winter Conference on Applications of Computer Vision*, WACV 2017 (Santa Rosa, CA, USA, 24–31 March 2017).– IEEE.– 2017.– ISBN 9781509048236.– Pp. 464–472. doi ↑316, 318

[21] Goyal P., P. Dollár, Girshick R., Noordhuis P., Wesolowski L., Kyrola A., Tulloch A., Jia Y., He K. *Accurate, large minibatch SGD: training ImageNet in 1 hour.*– 2017.– 12 pp. arXiv 1706.02677 doi ↑316, 318

[22] Smith L. N., Topin N. *Super-convergence: very fast training of neural networks using large learning rates*, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (Baltimore, Maryland, USA, 15–17 April 2019), Proc. SPIE.– vol. **11006**.– SPIE.– 2019.– ISBN 9781510626775.– id. 1100612. ↑316, 318

[23] Keskar N. S., Mudigere D., Nocedal J., Smelyanskiy M., Tang P. T. P. *On large-batch training for deep learning: generalization gap and sharp minima // 5th International Conference on Learning Representations*, ICLR 2017 (Toulon, France, April 24–26, 2017).– 2017.– ISBN 978-1-7138-7271-9.– 16 pp. ↑316, 318, 321

[24] Kodryan M., Lobacheva E., Nakhodnov M., Vetrov D. *Training scale-invariant neural networks on the sphere can happen in three regimes // NIPS'22: 36th International Conference on Neural Information Processing Systems* (New Orleans, Louisiana, USA, 28 November–9 December 2022), Advances in Neural Information Processing Systems.– vol. **35**, Red Hook: Curran Associates Inc..– 2022.– ISBN 978-1-7138-7108-8.– Pp. 14058–14070. arXiv 2209.03695 ↑316, 318, 319, 320, 327, 329

[25] Находнов М. С., Кодрян М. С., Лобачева Е. М., Ветров Д. С. *Динамика и ландшафт функции потерь для глубоких нейронных сетей при обучении с квадратичной функцией потерь* // Докл. РАН. Матем., информ., проц. упр..– 2022.– Т. **508**.– С. 50–69. ↑316, 318

[26] Li Y., Wei C., Ma T. *Towards explaining the regularization effect of initial large learning rate in training neural networks // NIPS'19: 33rd International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 8–14, 2019), Advances in Neural Information Processing Systems.– vol. **32**, Red Hook: Curran Associates, Inc..– 2019.– ISBN 9781713807933.– Pp. 11674–11685. arXiv 1907.04595 ↑316, 318

[27] Sadrtdinov I., Kodryan M., Pokonechny E., Lobacheva E., Vetrov D. *Where do large learning rates lead us?NIPS'24: 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 10–15, 2024), Advances in Neural Information Processing Systems.– vol. **37**, Red Hook: Curran Associates, Inc..– 2024.– ISBN 979-8-3313-1438-5.– Pp. 58445–58479. ↑316, 317, 318, 320, 322, 324, 328, 329, 330, 331, 333, 334, 336, 339, 340, 342, 343, 345

[28] Ioffe S., Szegedy C. *Batch normalization: accelerating deep network training by reducing internal covariate shift // ICML'15: Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, July 6–11, 2015), PMLR.– vol. **37**.– 2015.– ISBN 9781510810587.– Pp. 448–456. arXiv 1502.03167 ↑317

[29] He K., Zhang X., Ren S., Sun J. *Deep residual learning for image recognition // 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2016 (Las Vegas, NV, USA, 27–30 June 2016).– IEEE.– 2016.– ISBN 9781467388528.– Pp. 770–778. ↑317

[30] Ba J. L., Kiros J. R., Hinton G. E. *Layer normalization.*– 2016.– 14 pp. arXiv 1607.06450 ↑317

[31] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. *Attention is all you need // NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA, December 4–9, 2017), Advances in Neural Information Processing Systems.– vol. **30**, Red Hook: Curran Associates Inc..– 2017.– ISBN 978-1-5108-6096-4.– Pp. 6000–6010. arXiv 1706.03762 ↑317

[32] Kohler J., Daneshmand H., Lucchi A., Hofmann T., Zhou M., Neymeyr K. *Exponential convergence rates for batch normalization: the power of length-direction decoupling in non-convex optimization // Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS 2019 (Naha, Okinawa, Japan, 16–18 April 2019), PMLR.– vol. **89**.– 2019.– Pp. 806–815. ↑317

[33] Bjorck N., Gomes C. P., Selman B., Weinberger K. Q. *Understanding batch normalization // NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems.– vol. **31**, Red Hook: Curran Associates Inc..– 2018.– ISBN 9781510884472.– Pp. 7705–7716. arXiv 1806.02375 ↑317

[34] Santurkar S., Tsipras D., Ilyas A., Madry A. *How does batch normalization help optimization?NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems.– vol. **31**, Red Hook: Curran Associates Inc..– 2018.– ISBN 9781510884472.– Pp. 2488–2498. arXiv 1805.11604 ↑317

[35] Arora S., Li Z., Lyu K. *Theoretical analysis of auto rate-tuning by batch normalization // 7th International Conference on Learning Representations*, ICLR 2019 (New Orleans, LA, USA, May 6–9, 2019).– ICLR.– 2019.– ISBN 978-1-7138-7273-3.– 22 pp. arXiv 1812.03981 ↑317, 318

[36] Hoffer E., Banner R., Golan I., Soudry D. *Norm matters: efficient and accurate normalization schemes in deep networks // NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems.– vol. **31**, Red Hook: Curran Associates Inc..– 2018.– ISBN 9781510884472.– Pp. 2164–2174. arXiv 1803.01814 ↑317

[37] van Laarhoven T. *L2 regularization versus batch and weight normalization.*– 2017.– 9 pp. arXiv 1706.05350 ↑317

[38] Zhang G., Wang C., Xu B., Grosse R. *Three mechanisms of weight decay regularization // 7th International Conference on Learning Representations*, ICLR 2019 (New Orleans, LA, USA, May 6–9, 2019).– 2019.– ISBN 978-1-7138-7273-3.– 16 pp. ↑317

[39] Li Z., Arora S. *An exponential learning rate schedule for deep learning // 8th International Conference on Learning Representations*, ICLR 2020 (Addis Ababa, Ethiopia, April 26–30, 2020).– Curran Associates, Inc..– 2020.– ISBN 9798331321987.– 29 pp. ↑317

[40] Li Z., Lyu K., Arora S. *Reconciling modern deep learning with traditional optimization analyses: the intrinsic learning rate // NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 6–12, 2020), Advances in Neural Information Processing Systems.– vol. **33**, Red Hook: Curran Associates Inc..– 2020.– ISBN 978-1-7138-2954-6.– Pp. 14544–14555. arXiv 2010.02916 ↑318

[41] Lobacheva E., Kodryan M., Chirkova N., Malinin A., Vetrov D. *On the periodic behavior of neural network training with batch normalization and weight decay* // *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* (Online, December 6–14, 2021), Advances in Neural Information Processing Systems.– vol. **34**, Red Hook: Curran Associates Inc..– 2021.– ISBN 978-1-7138-4539-3.– Pp. 21545–21556. ↑318

[42] Wang Y., Chen M., Zhao T., Tao M. *Large learning rate tames homogeneity: convergence and balancing effect* // *The Tenth International Conference on Learning Representations*, ICLR 2022 (Virtual Event, April 25–29, 2022).– Curran Associates, Inc..– 2022.– ISBN 9798331321932.– 45 pp. ↑318

[43] Yang N., Tang C., Tu Y. *Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions* // Physical Review Letters.– 2023.– Vol. **130**.– No. 23.– id. 237101. arXiv 2206.01246 ↑318

[44] Kleinberg B., Li Y., Yuan Y. *An alternative view: when does SGD escape local minima?Proceedings of the 35th International Conference on Machine Learning*, ICML 2018 (Stockholm, Sweden, July 10–15, 2018), PMLR.– vol. **80**.– 2018.– ISBN 9781510867963.– Pp. 2698–2707. arXiv 1802.06175 ↑318

[45] Mohtashami A., Jaggi M., Stich S. U. *Special properties of gradient descent with large learning rates* // *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA, 23–29 July 2023), PMLR.– vol. **202**.– ICML.– 2023.– ISBN 9781713889182.– Pp. 25082–25104. arXiv 2205.15142 ↑318

[46] Andriushchenko M., Varre A. V., Pillaud-Vivien L., Flammarion N. *SGD with large step sizes learns sparse features* // *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA, 23–29 July 2023), PMLR.– vol. **202**.– ICML.– 2023.– ISBN 9781713889182.– Pp. 903–925. arXiv 2210.05337 ↑318, 319, 336

[47] Chen F., Kunin D., Yamamura A., Ganguli S. *Stochastic collapse: how gradient noise attracts SGD dynamics towards simpler subnetworks* // Journal of Statistical Mechanics: Theory and Experiment.– October 2024.– Vol. **2024**.– id. 104023.– 47 pp. ↑318, 319, 336

[48] Ahn K., Bubeck S., Chewi S., Lee Y. T., Suarez F., Zhang Y. *Learning threshold neurons via the "edge of stability"* // *NIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA, December 10–16, 2023), Advances in Neural Information Processing Systems.– vol. **36**, Red Hook: Curran Associates Inc..– 2023.– ISBN 9781713899921.– Pp. 19540–19569. arXiv 2212.07469 ↑318, 319

[49] Nacson M. S., Ravichandran K., Srebro N., Soudry D. *Implicit bias of the step size in linear diagonal neural networks* // *39th International Conference on Machine Learning*, ICML 2022 (Baltimore, Maryland, USA, 17–23 July 2022), PMLR.– vol. **162**.– ICML.– 2022.– ISBN 9781713871019.– Pp. 16270–16295. URL ↑318, 319, 336

[50] Rosenfeld E., Risteski A. *Outliers with opposing signals have an outsized effect on neural network optimization* // *The Twelfth International Conference on Learning Representations*, ICLR 2024 (Vienna, Austria, May 7–11, 2024).– 2024.– ISBN 978-1-7138-9865-8.– 54 pp. ↑318, 319

[51] Robbins H., Monro S. *A stochastic approximation method* // The Annals of Mathematical Statistics.– 1951.– Vol. **22**.– No. 3.– Pp. 400–407. ↑318

[52] Du S. S., Zhai X., Poczos B., Singh A. *Gradient descent provably optimizes over-parameterized neural networks* // *7th International Conference on Learning Representations*, ICLR 2019 (New Orleans, LA, USA, May 6–9, 2019).– ICLR.– 2019.– ISBN 978-1-7138-7273-3.– 19 pp. ↑319

[53] Zou D., Cao Y., Zhou D., Gu Q. *Stochastic gradient descent optimizes over-parameterized deep ReLU networks*.– 2018.– 54 pp. arXiv 1811.08888 ↑319

[54] Ma S., Bassily R., Belkin M. *The power of interpolation: understanding the effectiveness of SGD in modern over-parametrized learning* // *Proceedings of the 35th International Conference on Machine Learning*, ICML 2018 (Stockholm, Sweden, July 10–15, 2018), PMLR.– vol. **80**.– 2018.– ISBN 9781510867963.– Pp. 3325–3334. ↑319

[55] Sadrtdinov I., Lobacheva E., Klimov I., Burtsev M., Katsnelson M. I., Vetrov D. *Can training dynamics of scale-invariant neural networks be explained by the thermodynamics of an ideal gas?*– 2025.– 33 pp.; *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI 2026 (Bremen, Germany, 15-21 August 2026).– IJCAI.– ISBN 978-1-956792-09-6.– Pp. 4759–4767 arXiv 2511.07308 ↑319

[56] Izmailov P., Podoprikhin D., Garipov T., Vetrov D., Wilson A. G. *Averaging weights leads to wider optima and better generalization* // *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*.– V. 1, UAI 2018 (Monterey, California, USA, 6–10 August 2018), Red Hook: Curran Associates, Inc..– 2018.– ISBN 978-1-5108-7160-1.– Pp. 876–885. ↑320, 326

[57] Garipov T., Izmailov P., Podoprikhin D., Vetrov D. P., Wilson A. G. *Loss surfaces, mode connectivity, and fast ensembling of DNNs* // *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems.– vol. **31**, Red Hook: Curran Associates Inc..– 2018.– ISBN 9781510884472.– Pp. 8803–8812. arXiv 1802.10026 ↑320, 331

[58] Frankle J., Dziugaite G. K., Roy D. M., Carbin M. *Linear mode connectivity and the lottery ticket hypothesis* // *ICML'20: Proceedings of the 37th International Conference on Machine Learning* (Online, 13–18 July 2020), PMLR.– vol. **119**.– 2020.– ISBN 9781713821120.– Pp. 3259–3269. arXiv 1912.05671 ↑320, 331

[59] Entezari R., Sedghi H., Saukh O., Neyshabur B. *The role of permutation invariance in linear mode connectivity of neural networks* // *The Tenth International Conference on Learning Representations*, ICLR 2022 (Virtual Event, April 25–29, 2022).– Curran Associates, Inc..– 2022.– ISBN 9798331321932.– 24 pp. ↑320, 331

[60] Huang W., Liu X., Wang X., Yamagishi J., Qian Y. *From sharpness to better generalization for speech deepfake detection* // *Proceedings of Interspeech 2025* (Rotterdam, The Netherlands, 17–21 August 2025).– ISCA.– 2025.– Pp. 5338–5342. ↑321

[61] Gaido M., Papi S., Bentivogli L., Brutti A., Cettolo M., Gretter R., Matassoni M., Nabih M., Negri M. *The warmup dilemma: how learning rate strategies impact speech-to-text model convergence* // *Proceedings of the 22nd International Conference on Spoken Language Translation*, IWSLT 2025 (Vienna, Austria, 31 July–1 August 2025).– ACL.– 2025.– ISBN 979-8-89176-272-5.– Pp. 47–55. arXiv 2505.23420 ↑321

[62] Vaessen N., van Leeuwen D. A. *The effect of batch size on contrastive self-supervised speech representation learning.*– 2024.– 14 pp. arXiv 2402.13723 ↑321

[63] Zeghidour N., Xu Q., Liptchinsky V., Usunier N., Synnaeve G., Collobert R. *Fully convolutional speech recognition.*– 2018.– 5 pp. arXiv 1812.06864 ↑321

[64] Nagamine T., Seltzer M. L., Mesgarani N. *Exploring how deep neural networks form phonemic categories // Proceedings of Interspeech 2015.*– V. 1 (Dresden, Germany, 6–10 September 2015).– ISCA.– 2015.– ISBN 9781510817906.– Pp. 1912–1916. ↑321

[65] Nagamine T., Seltzer M. L., Mesgarani N. *On the role of nonlinear transformations in deep neural network acoustic models // Proceedings of Interspeech 2016.*– V. 2 (San Francisco, California, USA, 8–12 September 2016).– ISCA.– 2016.– ISBN 978-1-5108-3313-5.– Pp. 803–807. ↑321

[66] Pallett D. S. *Performance assessment of automatic speech recognizers* // Journal of Research of the National Institute of Standards and Technology.– 1985.– Vol. **90**.– No. 5.– Pp. 371–387. ↑323

[67] Li J., Lavrukhin V., Ginsburg B., Leary R., Kuchaiev O., Cohen J. M., Nguyen H., Gadde R. T. *Jasper: an end-to-end convolutional neural acoustic model // Proceedings of Interspeech 2019.*– V. 1 (Graz, Austria, 15–19 September 2019).– ISCA.– 2019.– ISBN 978-1-5108-9683-3.– Pp. 71–75. ↑323

[68] Srivastava N., Hinton G., Krizhevsky A., Sutskever I., Salakhutdinov R. *Dropout: a simple way to prevent neural networks from overfitting* // Journal of Machine Learning Research.– 2014.– Vol. **15**.– No. 1.– Pp. 1929–1958. ↑323

[69] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., Killeen T., Lin Z., Gimelshein N., Antiga L., Desmaison A., A. Köpf, Yang E., DeVito Z., Raison M., Tejani A., Chilamkurthy S., Steiner B., Fang L., Bai J., Chintala S. *PyTorch: An imperative style, high-performance deep learning library // NIPS'19: 33rd International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 8–14, 2019), Advances in Neural Information Processing Systems.– vol. **32**, Red Hook: Curran Associates, Inc..– 2019.– ISBN 9781713807933.– Pp. 8026–8037. arXiv 1912.01703 ↑324

[70] Graves A., S. Fernández, Gomez F., Schmidhuber J. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks // ICML'06: Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA, June 25–29, 2006), New York: ACM.– 2006.– ISBN 978-1-59593-383-6.– Pp. 369–376. ↑325

[71] Поляк Б. Т. *О некоторых способах ускорения сходимости итерационных методов* // Ж. вычисл. матем. и матем. физ.– 1964.– Т. **4**.– № 5.– С. 791–803. ↑325

[72] Gulati A., Qin J., Chiu C.-C., Parmar N., Zhang Y., Yu J., Han W., Wang S., Zhang Z., Wu Y., Pang R. *Conformer: Convolution-augmented Transformer for speech recognition // Proceedings of Interspeech 2020.*– V. 1 (Shanghai, China, 25–29 October 2020).– ICSA.– 2020.– ISBN 978-1-7138-2069-7.– Pp. 5036–5040. ↑343

[73] Baevski A., Zhou H., Mohamed A., Auli M. *wav2vec 2.0: A framework for self-supervised learning of speech representations // NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 6–12, 2020), Advances in Neural Information Processing Systems.– vol. **33**, Red Hook: Curran Associates Inc..– 2020.– ISBN 978-1-7138-2954-6.– Pp. 12449–12460. arXiv 2006.11477 ↑343

[74] Hsu W.-N., Bolte B., Tsai Y.-H. H., Lakhotia K., Salakhutdinov R., Mohamed A. *HuBERT: Self-supervised speech representation learning by masked prediction of hidden units* // IEEE/ACM Transactions on Audio, Speech, and Language Processing.– 2021.– Vol. **29**.– Pp. 3451–3460. ↑343

[75] Panayotov V., Chen G., Povey D., Khudanpur S. *LibriSpeech: An ASR corpus based on public domain audio books* // *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2015 (South Brisbane, QLD, Australia, 19–24 April 2015).– IEEE.– 2015.– ISBN 978-1-4673-6997-8.– Pp. 5206–5210. ↑344

[76] Park D. S., Chan W., Zhang Y., Chiu C.-C., Zoph B., Cubuk E. D., Le Q. V. *SpecAugment: A simple data augmentation method for automatic speech recognition* // *Proceedings of Interspeech 2019*.– V. 1 (Graz, Austria, 15–19 September 2019).– ISCA.– 2019.– ISBN 978-1-5108-9683-3.– Pp. 2613–2617. ↑344

Поступила в редакцию 22.07.2026;
одобрена после рецензирования 26.08.2026;
принята к публикации 12.09.2026;
опубликована онлайн 20.09.2026.

Рекомендовал к публикации *к.т.н. С. А. Амелькин*

Информация об авторе:

Ильдус Рустемович Садртдинов

аспирант национального исследовательского университета «Высшая школа экономики». Научные интересы включают глубинное обучение, анализ динамики обучения нейронных сетей и интерпретация их восприятия признаков.

0009-0000-1295-6091

e-mail: *isadrtdinov@hse.ru*

Декларация об отсутствии личной заинтересованности: *благополучие автора не зависит от результатов исследования.*

EN RU PROGRAM SYSTEMS: THEORY AND APPLICATIONS ISSN 2079-3316
Research Article ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

UDC 004.934:004.032.26
10.25209/2079-3316-2026-17-3-315-362

The effect of initial learning rate on properties of neural networks in automatic speech recognition

Ildus Rustemovich **Sadrtdinov**
HSE University, Moscow, Russia
✉ *isadrtdinov@hse.ru*

Abstract. The initial learning rate of stochastic gradient descent plays a decisive role in the quality of the resulting solution and in the structure of features learned by a neural network. Recent works on image classification have established a taxonomy of three regimes of the initial learning rate and identified a narrow range of high values that yields optimal quality after subsequent fine-tuning with a low learning rate or weight averaging.

This work extends this methodology to speech recognition, where the regime taxonomy is reproduced in its main features, and the identified range of high values retains both its superior final quality after fine-tuning and weight averaging and the linear mode connectivity of the resulting solutions in weight space. The picture of learned features is reproduced in a weaker form: as the learning rate grows, the models shift their reliance toward the lower cepstral coefficients and the low-frequency part of the mel spectrogram, yet within the high-value region this shift no longer singles out the range of best quality. In addition, truncation of the mel-frequency cepstrum is proposed as a portable tool for analyzing learned features, applicable to any model operating on spectrograms.

The obtained results show that the identified regime is a general property of stochastic optimization of neural networks rather than an artifact of the image classification task or of training on a sphere. Therefore, the identified regime is not an artifact of the image classification task and successfully transfers to speech recognition. (*In Russian*).

Key words and phrases: learning rate, stochastic gradient descent, speech recognition, linear mode connectivity, feature learning

2020 *Mathematics Subject Classification*: 68T07; 68T10, 68T20

For citation: Ildus R. Sadrtdinov. *The effect of initial learning rate on properties of neural networks in automatic speech recognition*. Program Systems: Theory and Applications, 2026, **17**:3(72), pp. 315–362. (*In Russ.*).
https://psta.psiras.ru/read/psta2026_3_315-362.pdf

© Sadrtdinov I. R. 2026 CC BY

References

[1] A. S. Zakovryashin, P. V. Malinin, A. A. Lependin. "Speaker recognition using mel-frequency cepstral coefficient distributions", *Izvestiya Altajskogo gosudarstvennogo universiteta*, 2014, no. 1-1 (81), pp. 156–160 (in Russian).

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton. "ImageNet classification with deep convolutional neural networks", *Communications of the ACM*, **60**:6 (2017), pp. 84–90.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby. "An image is worth 16x16 words: transformers for image recognition at scale", *9th International Conference on Learning Representations*, ICLR 2021 (Virtual Event, Austria, 3–7 May 2021), Curran Associates, Inc., 2021, ISBN 9798331321949, 22 pp. arXiv 2010.11929

[4] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. -r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, B. Kingsbury. "Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups", *IEEE Signal Processing Magazine*, **29**:6 (2012), pp. 82–97.

[5] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever. "Robust speech recognition via large-scale weak supervision", *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA, 23–29 July 2023), PMLR, vol. **202**, ICML, ISBN 9781713889182, pp. 28492–28518. arXiv 2212.04356

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A., Askell S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, Ch. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, Ch. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei. "Language models are few-shot learners", *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 6–12, 2020), Advances in Neural Information Processing Systems, vol. **33**, Curran Associates Inc., Red Hook, 2020, ISBN 978-1-7138-2954-6, pp. 1877–1901. arXiv 2005.14165

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. -A. Lachaux, T. Lacroix, Rozière B., N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample. *LLaMA: open and efficient foundation language models*, 2023, 27 pp. arXiv 2302.13971

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, Žídek A., A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis. "Highly accurate protein structure prediction with AlphaFold", *Nature*, **596**:7873 (2021), pp. 583–589.

[9] Y. LeCun, Y. Bengio, G. Hinton. "Deep learning", *Nature*, **521**:7553 (2015), pp. 436–444.
[10] Yu. Zhang, L. Zheng, K. Zhou, P. Liang. *On the opportunities and risks of foundation models*, 2021, 214 pp. arXiv 2108.07258
[11] C. Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead", *Nature Machine Intelligence*, **1**:5 (2019), pp. 206–215.
[12] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, C. Zhong. "Interpretable machine learning: fundamental principles and 10 grand challenges", *Statistics Surveys*, **16** (2022), pp. 1–85.
[13] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, Y. Bengio. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization", *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems*. V. 2 (Montreal, Canada, 8–13 December 2014), Advances in Neural Information Processing Systems, vol. **27**, MIT Press, Cambridge, 2014, ISBN 9781510800410, pp. 2933–2941. arXiv 1406.2572
[14] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein. "Visualizing the loss landscape of neural nets", *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montreal, Canada, 3–8 December 2018), Advances in Neural Information Processing Systems, vol. **31**, Curran Associates, Inc., Red Hook, 2018, ISBN 9781510884472, pp. 6391–6401. arXiv 1712.09913
[15] Jastrzębski S., Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, A. Storkey. *Three factors influencing minima in SGD*, 2017, 14 pp. arXiv 1711.04623
[16] A. Lewkowycz, Y. Bahri, E. Dyer, J. Sohl-Dickstein, G. Gur-Ari. *The large learning rate phase of deep learning: the catapult mechanism*, 2020, 25 pp. arXiv 2003.02218
[17] J. M. Cohen, S. Kaur, Y. Li, J. Z. Kolter, A. Talwalkar. "Gradient descent on neural networks typically occurs at the edge of stability", *9th International Conference on Learning Representations*, ICLR 2021 (Virtual Event, Austria, 3–7 May 2021), Curran Associates, Inc., 2021, ISBN 979-8-3313-0008-1, 84 pp.
[18] Y. Bengio. "Practical recommendations for gradient-based training of deep architectures", *Neural Networks: Tricks of the Trade*, Lecture Notes in Computer Science, vol. **7700**, 2nd ed., Springer, Berlin–Heidelberg, 2012, ISBN 978-3-642-35288-1, pp. 437–478.
[19] S. L. Smith, Q. V. Le. "A Bayesian perspective on generalization and stochastic gradient descent", *6th International Conference on Learning Representations*, ICLR 2018 (Vancouver, BC, Canada, April 30–May 3, 2018), 2018, ISBN 9798331321963, 13 pp.
[20] L. N. Smith. "Cyclical learning rates for training neural networks", *2017 IEEE Winter Conference on Applications of Computer Vision*, WACV 2017 (Santa Rosa, CA, USA, 24–31 March 2017), IEEE, 2017, ISBN 9781509048236, pp. 464–472.

[21] P. Goyal, Dollár P., R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, K. He. *Accurate, large minibatch SGD: training ImageNet in 1 hour*, 2017, 12 pp. arXiv 1706.02677

[22] L. N. Smith, N. Topin. "Super-convergence: very fast training of neural networks using large learning rates", Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications (Baltimore, Maryland, USA, 15–17 April 2019), Proc. SPIE, vol. **11006**, SPIE, 2019, ISBN 9781510626775, id. 1100612.

[23] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. T. P. Tang. "On large-batch training for deep learning: generalization gap and sharp minima", *5th International Conference on Learning Representations*, ICLR 2017 (Toulon, France, April 24–26, 2017), 2017, ISBN 978-1-7138-7271-9, 16 pp.

[24] M. Kodryan, E. Lobacheva, M. Nakhodnov, D. Vetrov. "Training scale-invariant neural networks on the sphere can happen in three regimes", *NIPS'22: 36th International Conference on Neural Information Processing Systems* (New Orleans, Louisiana, USA, 28 November–9 December 2022), Advances in Neural Information Processing Systems, vol. **35**, Curran Associates Inc., Red Hook, 2022, ISBN 978-1-7138-7108-8, pp. 14058–14070. arXiv 2209.03695

[25] M. S. Nakhodnov, M. S. Kodryan, E. M. Lobacheva, D. S. Vetrov. "Loss function dynamics and landscape for deep neural networks trained with quadratic loss", *Dokl. RAN. Matem., inform., proc. upr.*, **508** (2022), pp. 50–69.

[26] Y. Li, C. Wei, T. Ma. "Towards explaining the regularization effect of initial large learning rate in training neural networks", *NIPS'19: 33rd International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 8–14, 2019), Advances in Neural Information Processing Systems, vol. **32**, Curran Associates, Inc., Red Hook, 2019, ISBN 9781713807933, pp. 11674–11685. arXiv 1907.04595

[27] I. Sadrtdinov, M. Kodryan, E. Pokonechny, E. Lobacheva, D. Vetrov. Where do large learning rates lead us? *NIPS'24: 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 10–15, 2024), Advances in Neural Information Processing Systems, vol. **37**, Curran Associates, Inc., Red Hook, 2024, ISBN 979-8-3313-1438-5, pp. 58445–58479.

[28] S. Ioffe, C. Szegedy. "Batch normalization: accelerating deep network training by reducing internal covariate shift", *ICML'15: Proceedings of the 32nd International Conference on Machine Learning* (Lille, France, July 6–11, 2015), PMLR, vol. **37**, 2015, ISBN 9781510810587, pp. 448–456. arXiv 1502.03167

[29] K. He, X. Zhang, S. Ren, J. Sun. "Deep residual learning for image recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR 2016 (Las Vegas, NV, USA, 27–30 June 2016), IEEE, 2016, ISBN 9781467388528, pp. 770–778.

[30] J. L. Ba, J. R. Kiros, G. E. Hinton. *Layer normalization*, 2016, 14 pp. arXiv 1607.06450

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin. "Attention is all you need", *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA, December 4–9, 2017), Advances in Neural Information Processing Systems, vol. **30**, Curran Associates Inc., Red Hook, 2017, ISBN 978-1-5108-6096-4, pp. 6000–6010. arXiv 1706.03762

[32] J. Kohler, H. Daneshmand, A. Lucchi, T. Hofmann, M. Zhou, K. Neymeyr. "Exponential convergence rates for batch normalization: the power of length-direction decoupling in non-convex optimization", *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, AISTATS 2019 (Naha, Okinawa, Japan, 16–18 April 2019), PMLR, vol. **89**, 2019, pp. 806–815.

[33] N. Bjorck, C. P. Gomes, B. Selman, K. Q. Weinberger. "Understanding batch normalization", *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems, vol. **31**, Curran Associates Inc., Red Hook, 2018, ISBN 9781510884472, pp. 7705–7716. arXiv 1806.02375

[34] S. Santurkar, D. Tsipras, A. Ilyas, A. Madry. How does batch normalization help optimization? *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems, vol. **31**, Curran Associates Inc., Red Hook, 2018, ISBN 9781510884472, pp. 2488–2498. arXiv 1805.11604

[35] S. Arora, Z. Li, K. Lyu. "Theoretical analysis of auto rate-tuning by batch normalization", *7th International Conference on Learning Representations*, ICLR 2019 (New Orleans, LA, USA, May 6–9, 2019), ICLR, 2019, ISBN 978-1-7138-7273-3, 22 pp. arXiv 1812.03981

[36] E. Hoffer, R. Banner, I. Golan, D. Soudry. "Norm matters: efficient and accurate normalization schemes in deep networks", *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems, vol. **31**, Curran Associates Inc., Red Hook, 2018, ISBN 9781510884472, pp. 2164–2174. arXiv 1803.01814

[37] T. van Laarhoven. *L2 regularization versus batch and weight normalization*, 2017, 9 pp. arXiv 1706.05350

[38] G. Zhang, C. Wang, B. Xu, R. Grosse. "Three mechanisms of weight decay regularization", *7th International Conference on Learning Representations*, ICLR 2019 (New Orleans, LA, USA, May 6–9, 2019), 2019, ISBN 978-1-7138-7273-3, 16 pp.

[39] Z. Li, S. Arora. "An exponential learning rate schedule for deep learning", *8th International Conference on Learning Representations*, ICLR 2020 (Addis Ababa, Ethiopia, April 26–30, 2020), Curran Associates, Inc., 2020, ISBN 9798331321987, 29 pp.

[40] Z. Li, K. Lyu, S. Arora. "Reconciling modern deep learning with traditional optimization analyses: the intrinsic learning rate", *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 6–12, 2020), Advances in Neural Information Processing Systems, vol. **33**, Curran Associates Inc., Red Hook, 2020, ISBN 978-1-7138-2954-6, pp. 14544–14555. arXiv 2010.02916

[41] E. Lobacheva, M. Kodryan, N. Chirkova, A. Malinin, D. Vetrov. "On the periodic behavior of neural network training with batch normalization and weight decay", *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* (Online, December 6–14, 2021), Advances in Neural Information Processing Systems, vol. **34**, Curran Associates Inc., Red Hook, 2021, ISBN 978-1-7138-4539-3, pp. 21545–21556.

[42] Y. Wang, M. Chen, T. Zhao, M. Tao. "Large learning rate tames homogeneity: convergence and balancing effect", *The Tenth International Conference on Learning Representations*, ICLR 2022 (Virtual Event, April 25–29, 2022), Curran Associates, Inc., 2022, ISBN 9798331321932, 45 pp.

[43] N. Yang, C. Tang, Y. Tu. "Stochastic gradient descent introduces an effective landscape-dependent regularization favoring flat solutions", *Physical Review Letters*, **130**:23 (2023), id. 237101. arXiv 2206.01246

[44] B. Kleinberg, Y. Li, Y. Yuan. An alternative view: when does SGD escape local minima? *Proceedings of the 35th International Conference on Machine Learning*, ICML 2018 (Stockholm, Sweden, July 10–15, 2018), PMLR, vol. **80**, 2018, ISBN 9781510867963, pp. 2698–2707. arXiv 1802.06175

[45] A. Mohtashami, M. Jaggi, S. U. Stich. "Special properties of gradient descent with large learning rates", *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA, 23–29 July 2023), PMLR, vol. **202**, ICML, 2023, ISBN 9781713889182, pp. 25082–25104. arXiv 2205.15142

[46] M. Andriushchenko, A. V. Varre, L. Pillaud-Vivien, N. Flammarion. "SGD with large step sizes learns sparse features", *ICML'23: Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA, 23–29 July 2023), PMLR, vol. **202**, ICML, 2023, ISBN 9781713889182, pp. 903–925. arXiv 2210.05337

[47] F. Chen, D. Kunin, A. Yamamura, S. Ganguli. "Stochastic collapse: how gradient noise attracts SGD dynamics towards simpler subnetworks", *Journal of Statistical Mechanics: Theory and Experiment*, **2024** (October 2024), id. 104023, 47 pp.

[48] K. Ahn, S. Bubeck, S. Chewi, Y. T. Lee, F. Suarez, Y. Zhang. "Learning threshold neurons via the "edge of stability"", *NIPS'23: Proceedings of the 37th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA, December 10–16, 2023), Advances in Neural Information Processing Systems, vol. **36**, Curran Associates Inc., Red Hook, 2023, ISBN 9781713899921, pp. 19540–19569. arXiv 2212.07469

[49] M. S. Nacson, K. Ravichandran, N. Srebro, D. Soudry. "Implicit bias of the step size in linear diagonal neural networks", *39th International Conference on Machine Learning*, ICML 2022 (Baltimore, Maryland, USA, 17–23 July 2022), PMLR, vol. **162**, ICML, 2022, ISBN 9781713871019, pp. 16270–16295. URL

[50] E. Rosenfeld, A. Risteski. "Outliers with opposing signals have an outsized effect on neural network optimization", *The Twelfth International Conference on Learning Representations*, ICLR 2024 (Vienna, Austria, May 7–11, 2024), 2024, ISBN 978-1-7138-9865-8, 54 pp. doi

[51] H. Robbins, S. Monro. "A stochastic approximation method", *The Annals of Mathematical Statistics*, **22**:3 (1951), pp. 400–407. doi

[52] S. S. Du, X. Zhai, B. Poczos, A. Singh. "Gradient descent provably optimizes over-parameterized neural networks", *7th International Conference on Learning Representations*, ICLR 2019 (New Orleans, LA, USA, May 6–9, 2019), ICLR, 2019, ISBN 978-1-7138-7273-3, 19 pp. doi

[53] D. Zou, Y. Cao, D. Zhou, Q. Gu. *Stochastic gradient descent optimizes over-parameterized deep ReLU networks*, 2018, 54 pp. arXiv 1811.08888 doi

[54] S. Ma, R. Bassily, M. Belkin. "The power of interpolation: understanding the effectiveness of SGD in modern over-parametrized learning", *Proceedings of the 35th International Conference on Machine Learning*, ICML 2018 (Stockholm, Sweden, July 10–15, 2018), PMLR, vol. **80**, 2018, ISBN 9781510867963, pp. 3325–3334. doi

[55] I. Sadrtdinov, E. Lobacheva, I. Klimov, M. Burtsev, M. I. Katsnelson, D. Vetrov. *Can training dynamics of scale-invariant neural networks be explained by the thermodynamics of an ideal gas?* 2025, 33 pp.; *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI 2026 (Bremen, Germany, 15-21 August 2026), IJCAI, ISBN 978-1-956792-09-6, pp. 4759–4767 arXiv 2511.07308 doi doi

[56] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson. "Averaging weights leads to wider optima and better generalization", *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*. V. 1, UAI 2018 (Monterey, California, USA, 6–10 August 2018), Curran Associates, Inc., Red Hook, 2018, ISBN 978-1-5108-7160-1, pp. 876–885. doi

[57] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, A. G. Wilson. "Loss surfaces, mode connectivity, and fast ensembling of DNNs", *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada, December 3–8, 2018), Advances in Neural Information Processing Systems, vol. **31**, Curran Associates Inc., Red Hook, 2018, ISBN 9781510884472, pp. 8803–8812. arXiv 1802.10026 doi

[58] J. Frankle, G. K. Dziugaite, D. M. Roy, M. Carbin. "Linear mode connectivity and the lottery ticket hypothesis", *ICML'20: Proceedings of the 37th International Conference on Machine Learning* (Online, 13–18 July 2020), PMLR, vol. **119**, 2020, ISBN 9781713821120, pp. 3259–3269. doi arXiv 1912.05671

[59] R. Entezari, H. Sedghi, O. Saukh, B. Neyshabur. "The role of permutation invariance in linear mode connectivity of neural networks", *The Tenth International Conference on Learning Representations*, ICLR 2022 (Virtual Event, April 25–29, 2022), Curran Associates, Inc., 2022, ISBN 9798331321932, 24 pp.

[60] W. Huang, X. Liu, X. Wang, J. Yamagishi, Y. Qian. "From sharpness to better generalization for speech deepfake detection", *Proceedings of Interspeech 2025* (Rotterdam, The Netherlands, 17–21 August 2025), ISCA, 2025, pp. 5338–5342.

[61] M. Gaido, S. Papi, L. Bentivogli, A. Brutti, M. Cettolo, R. Gretter, M. Matassoni, M. Nabih, M. Negri. "The warmup dilemma: how learning rate strategies impact speech-to-text model convergence", *Proceedings of the 22nd International Conference on Spoken Language Translation*, IWSLT 2025 (Vienna, Austria, 31 July–1 August 2025), ACL, 2025, ISBN 979-8-89176-272-5, pp. 47–55. arXiv 2505.23420

[62] N. Vaessen, D. A. van Leeuwen. *The effect of batch size on contrastive self-supervised speech representation learning*, 2024, 14 pp. arXiv 2402.13723

[63] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, R. Collobert. *Fully convolutional speech recognition*, 2018, 5 pp. arXiv 1812.06864

[64] T. Nagamine, M. L. Seltzer, N. Mesgarani. "Exploring how deep neural networks form phonemic categories", *Proceedings of Interspeech 2015*. V. 1 (Dresden, Germany, 6–10 September 2015), ISCA, 2015, ISBN 9781510817906, pp. 1912–1916.

[65] T. Nagamine, M. L. Seltzer, N. Mesgarani. "On the role of nonlinear transformations in deep neural network acoustic models", *Proceedings of Interspeech 2016*. V. 2 (San Francisco, California, USA, 8–12 September 2016), ISCA, 2016, ISBN 978-1-5108-3313-5, pp. 803–807.

[66] D. S. Pallett. "Performance assessment of automatic speech recognizers", *Journal of Research of the National Institute of Standards and Technology*, **90**:5 (1985), pp. 371–387.

[67] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. M. Cohen, H. Nguyen, R. T. Gadde. "Jasper: an end-to-end convolutional neural acoustic model", *Proceedings of Interspeech 2019*. V. 1 (Graz, Austria, 15–19 September 2019), ISCA, 2019, ISBN 978-1-5108-9683-3, pp. 71–75.

[68] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting", *Journal of Machine Learning Research*, **15**:1 (2014), pp. 1929–1958.

[69] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, Köpf A., E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala. "PyTorch: An imperative style, high-performance deep learning library", *NIPS'19: 33rd International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 8–14, 2019), Advances in Neural Information Processing Systems, vol. **32**, Curran Associates, Inc., Red Hook, 2019, ISBN 9781713807933, pp. 8026–8037. arXiv 1912.01703

[70] A. Graves, Fernández S., F. Gomez, J. Schmidhuber. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", *ICML'06: Proceedings of the 23rd International Conference on Machine Learning* (Pittsburgh, Pennsylvania, USA, June 25–29, 2006), ACM, New York, 2006, ISBN 978-1-59593-383-6, pp. 369–376.

[71] B. T. Polyak. "Some methods of speeding up the convergence of iteration methods", *USSR Computational Mathematics and Mathematical Physics*, **4**:5 (1964), pp. 791–803.

[72] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang. "Conformer: Convolution-augmented Transformer for speech recognition", *Proceedings of Interspeech 2020*. V. 1 (Shanghai, China, 25–29 October 2020), ICSA, 2020, ISBN 978-1-7138-2069-7, pp. 5036–5040.

[73] A. Baevski, H. Zhou, A. Mohamed, M. Auli. "wav2vec 2.0: A framework for self-supervised learning of speech representations", *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada, December 6–12, 2020), Advances in Neural Information Processing Systems, vol. **33**, Curran Associates Inc., Red Hook, 2020, ISBN 978-1-7138-2954-6, pp. 12449–12460. arXiv 2006.11477

[74] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, A. Mohamed. "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **29** (2021), pp. 3451–3460.

[75] V. Panayotov, G. Chen, D. Povey, S. Khudanpur. "LibriSpeech: An ASR corpus based on public domain audio books", *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP 2015 (South Brisbane, QLD, Australia, 19–24 April 2015), IEEE, 2015, ISBN 978-1-4673-6997-8, pp. 5206–5210.

[76] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le. "SpecAugment: A simple data augmentation method for automatic speech recognition", *Proceedings of Interspeech 2019*. V. 1 (Graz, Austria, 15–19 September 2019), ISCA, 2019, ISBN 978-1-5108-9683-3, pp. 2613–2617.