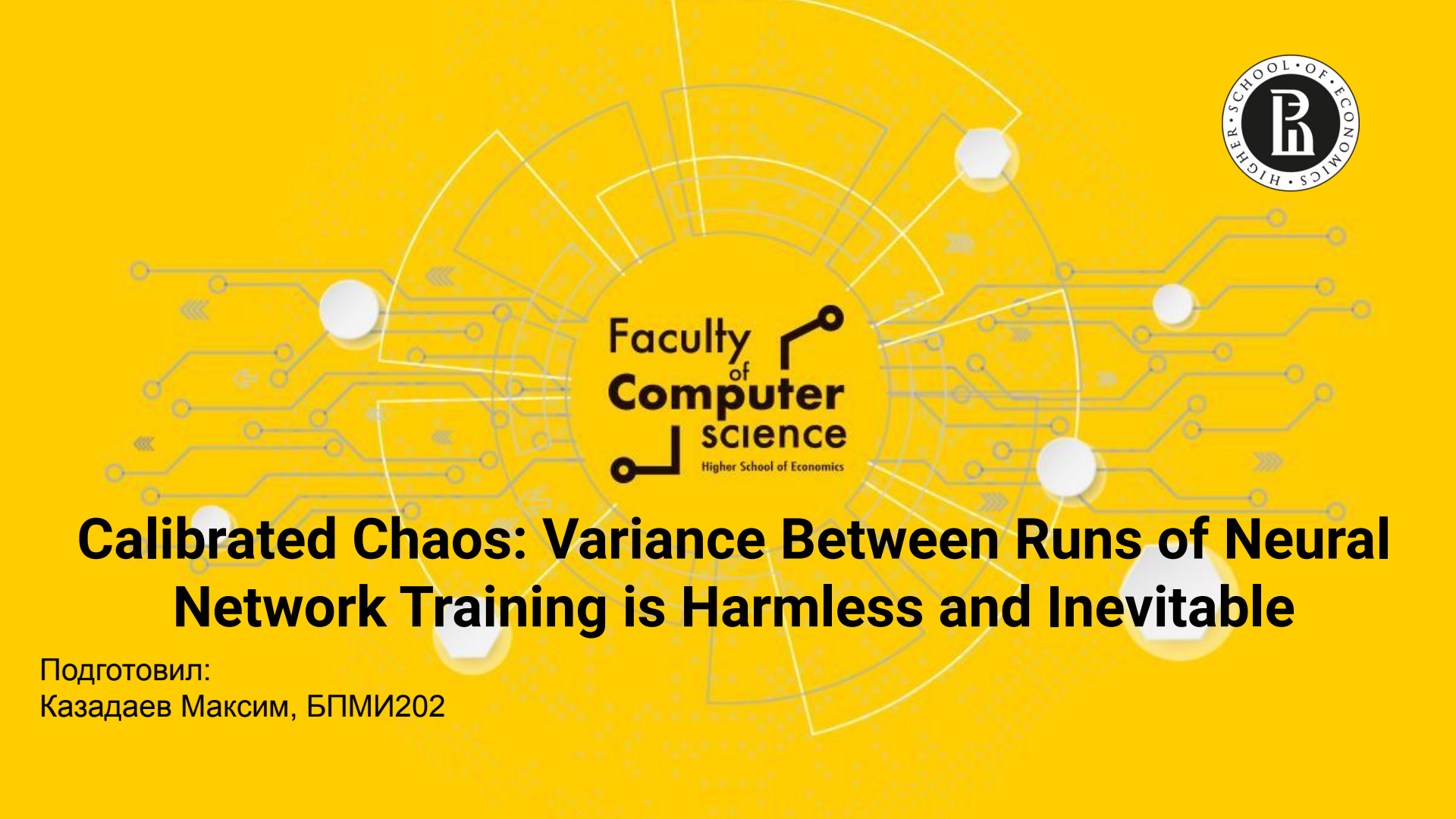
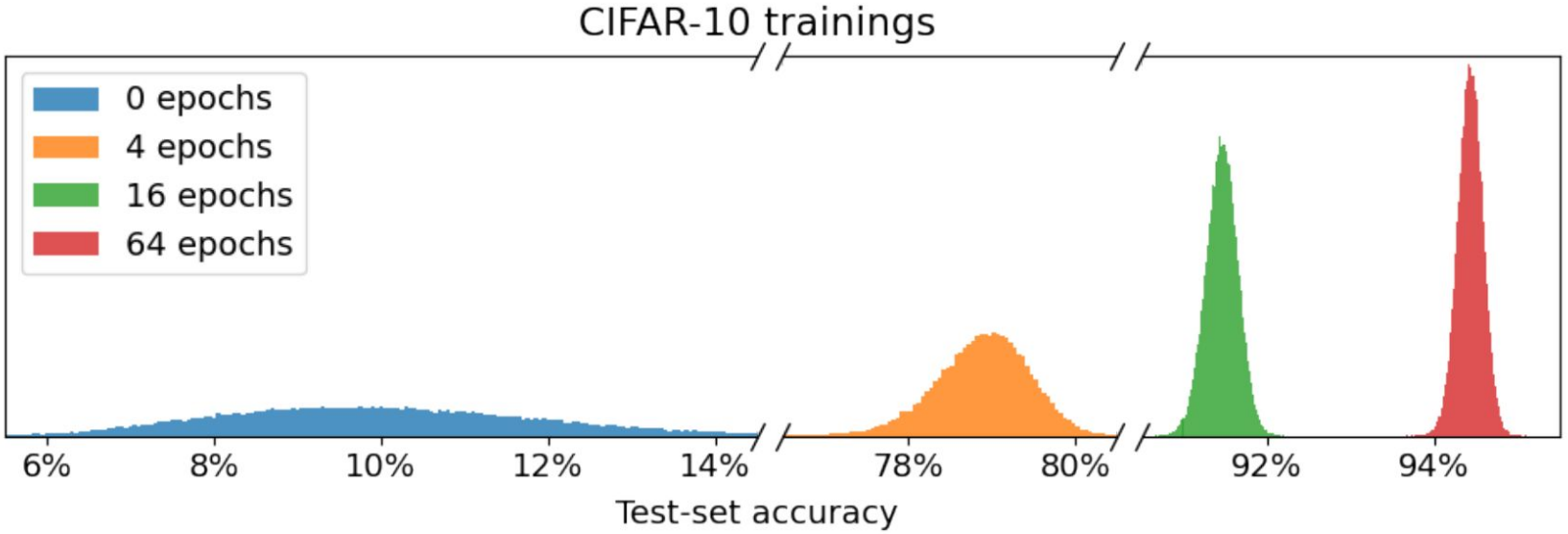


Calibrated Chaos: Variance Between Runs of Neural Network Training is Harmless and Inevitable

Подготовил:
Казадаев Максим, БПМИ202

Насколько разное качество на тесте можно получить?



- Обучаем около **60’000 моделей** ResNet-18 на выборке CIFAR-10 (классификация)
- Замеряем их качество на тестовой выборке
- **Почему и насколько сильно** зависит качество на тестовой выборке?

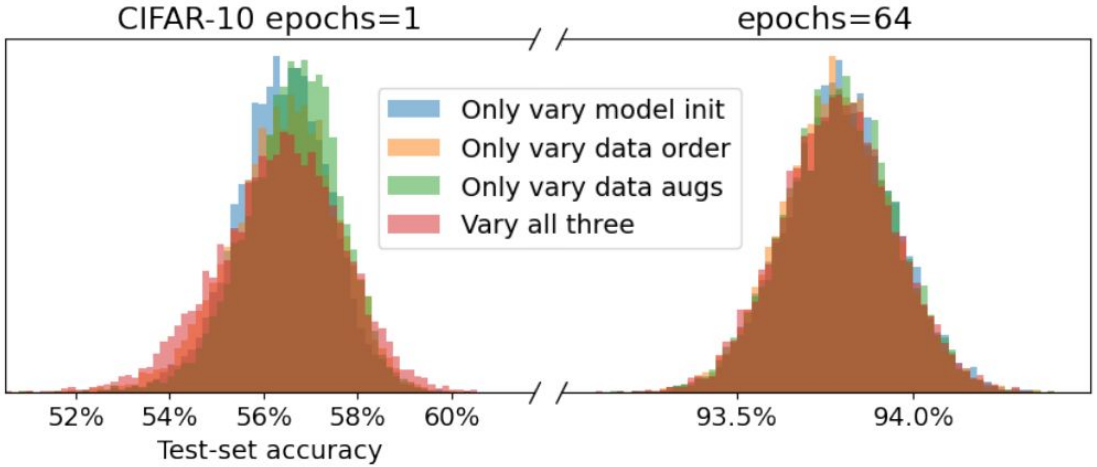
Higher School of Economics
Faculty of Computer science

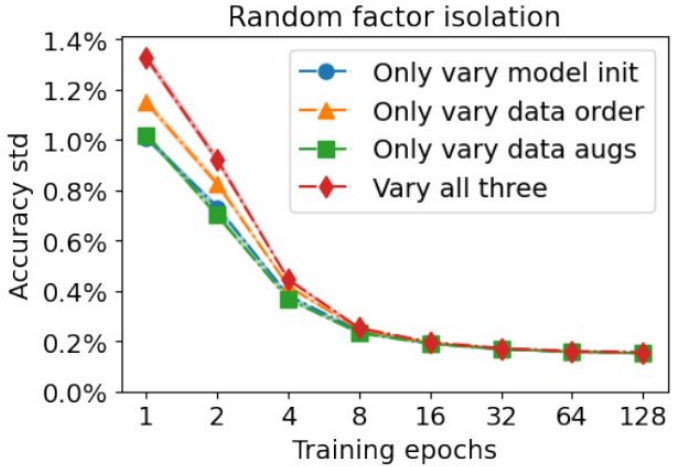
План

- Какие факторы случайности у нас есть?
- Почему сети получаются разными?
- Есть ли лучшие сиды?
- Как объяснить дисперсию качества на тестовой выборке?
- Насколько велика дисперсия качества на новых данных?

- Влияние аугментаций
- Влияние Learning Rate
- Влияние Distribution Shifts
- Влияние размера модели

Факторы случайности

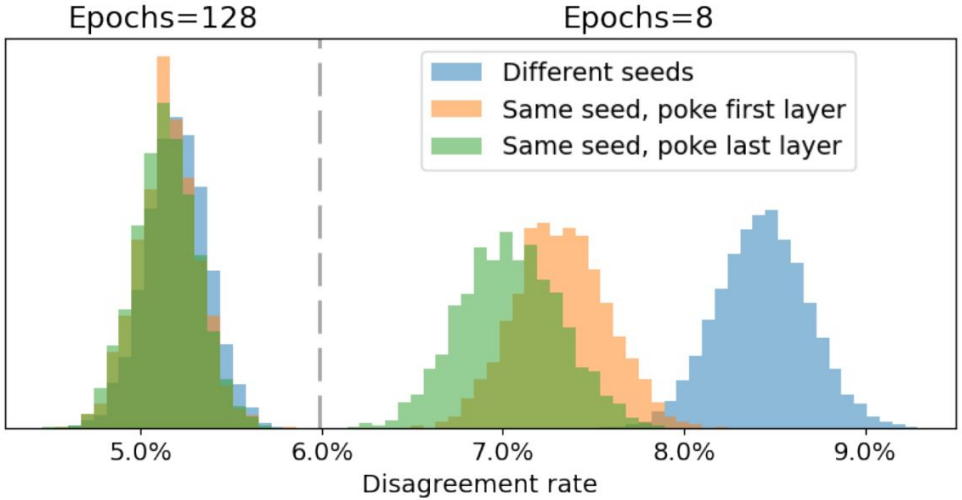


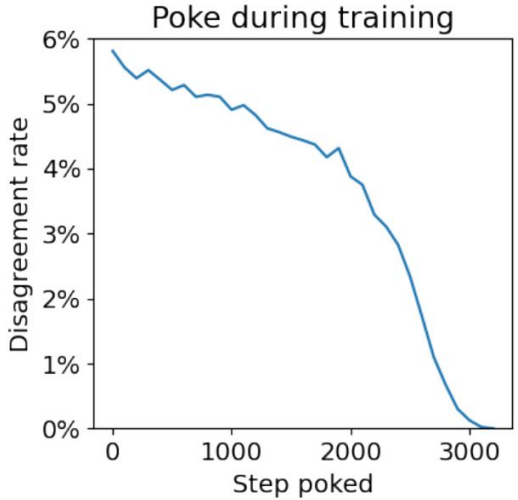


- random seed для инициализации
- random seed для порядка батчей
- random seed для аугментаций

Вывод: При большом числе эпох стандартное отклонение accuracy на тесте **не зависит от типа случайности**.

А если поменять всего 1 вес?

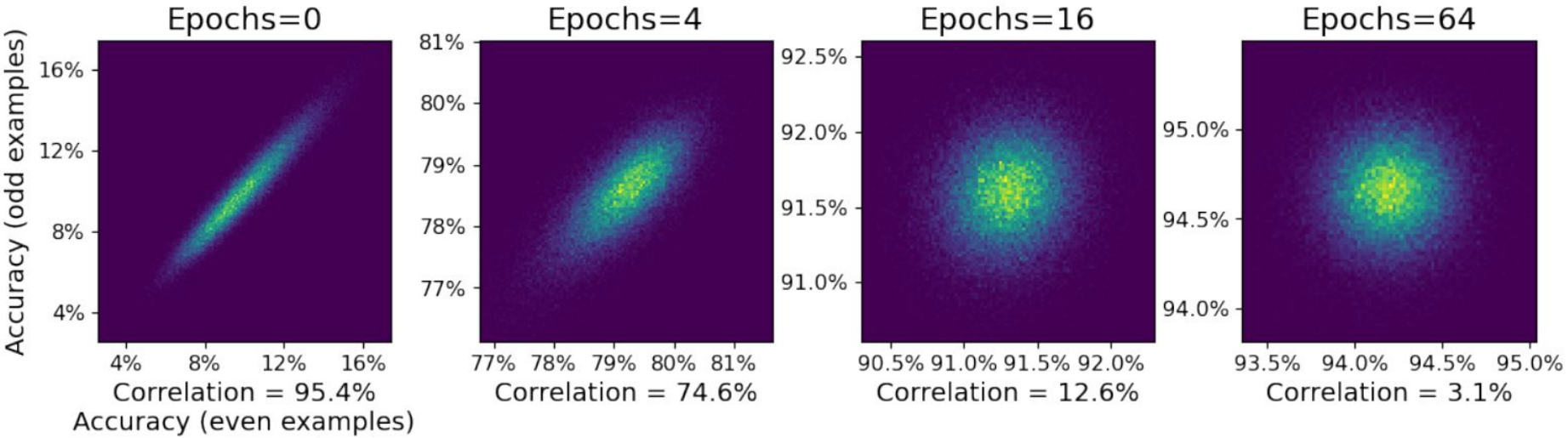




- Зафиксируем все параметры сети. Умножим случайный вес на 1.001 или представляем первый батч в f16
- Рассчитаем долю предсказаний, где нейросети выдают разные ответы

Вывод: Случайность вносится самим **процессом обучения**, а не конкретными факторами.

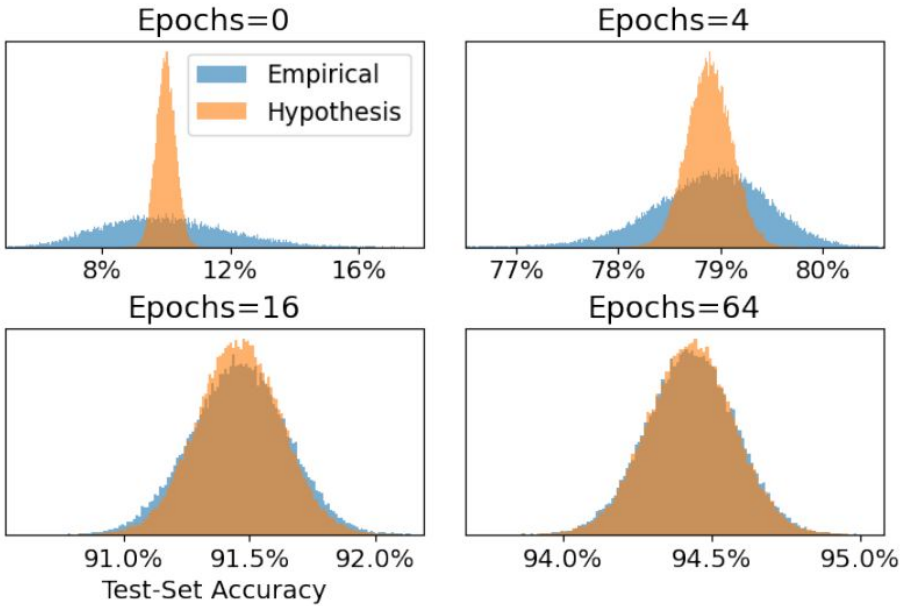
Есть ли лучшие сиды?

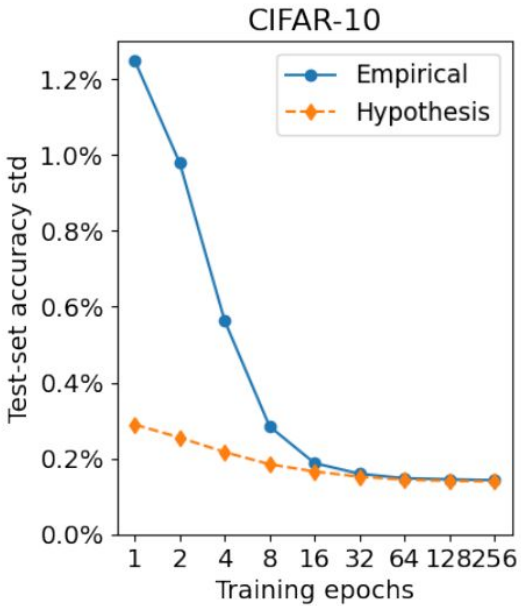


- Перебираем один гиперпараметр – random seed
- Делим тестовую выборку на 2 части: первую часть используем как валидацию для перебора random seed. Измеряем качество на обеих выборках

Вывод: Перебирать сид как гиперпараметр не стоит

Как объяснить дисперсию качества?





Гипотеза: $C_{x_i,y_i} \perp\!\!\!\perp C_{x_j,y_j} \quad \forall\, i \neq j$

Индикатор правильного ответа

Вывод: при большом числе эпох нейросеть ошибается на примерах независимо

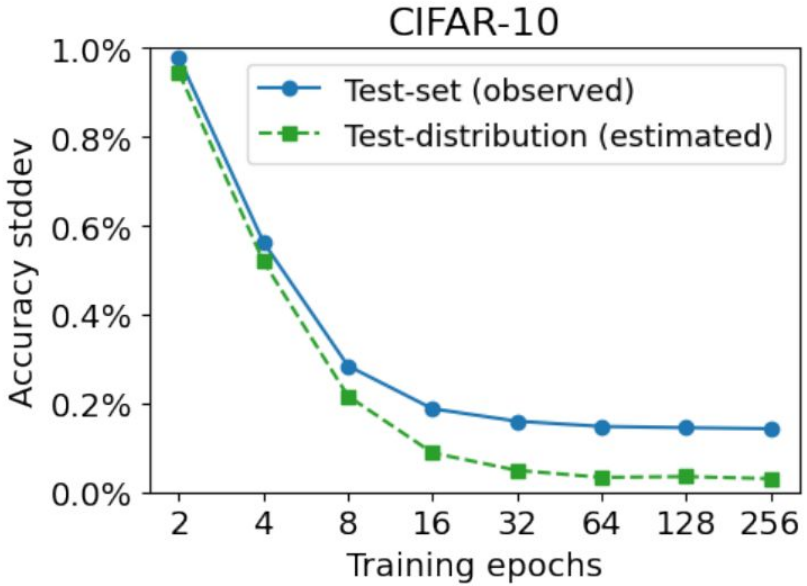
Дисперсия на распределении

$$\underset{\theta\sim\mathcal{A}}{\mathrm{Var}}(A(\theta)) = \frac{n}{n-1}\cdot \underset{S\sim\mathcal{D}^n}{\mathbb{E}}\left[\underset{\theta\sim\mathcal{A}}{\mathrm{Var}}(A_S(\theta)) - \frac{1}{n^2}\sum_{i=1}^{n}\overline{C}_{x_i,y_i}(1-\overline{C}_{x_i,y_i})\right]$$

Случайная тестовая выборка

Случайная нейросеть

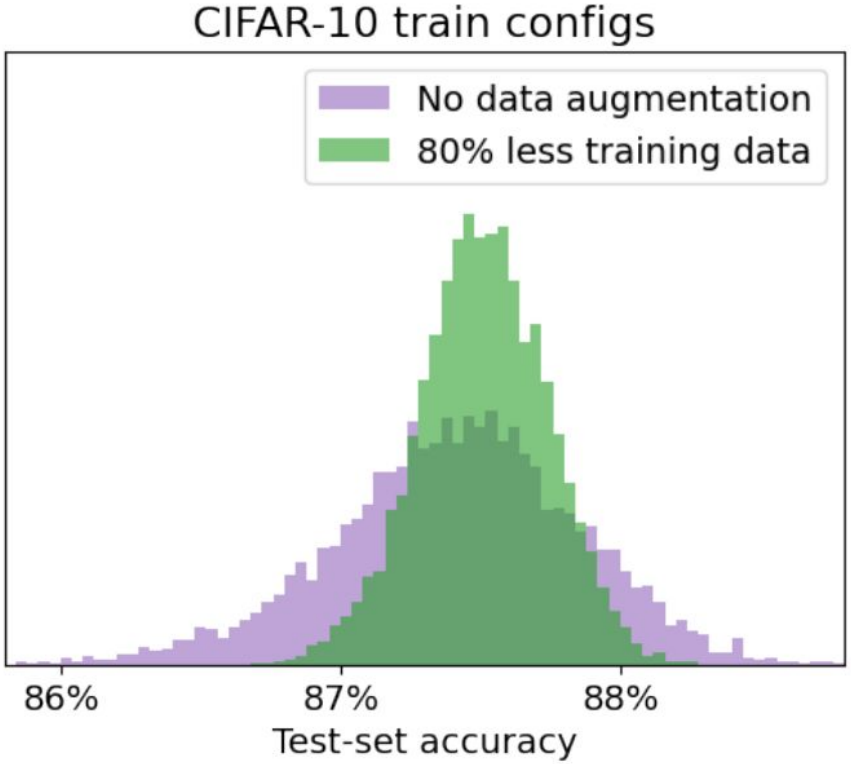
Средняя точность на объекте



Вывод:

- Оценка дисперсии качества между разными нейросетями на тестовой выборке **выше**, чем на настоящем распределении

Влияние аугментаций



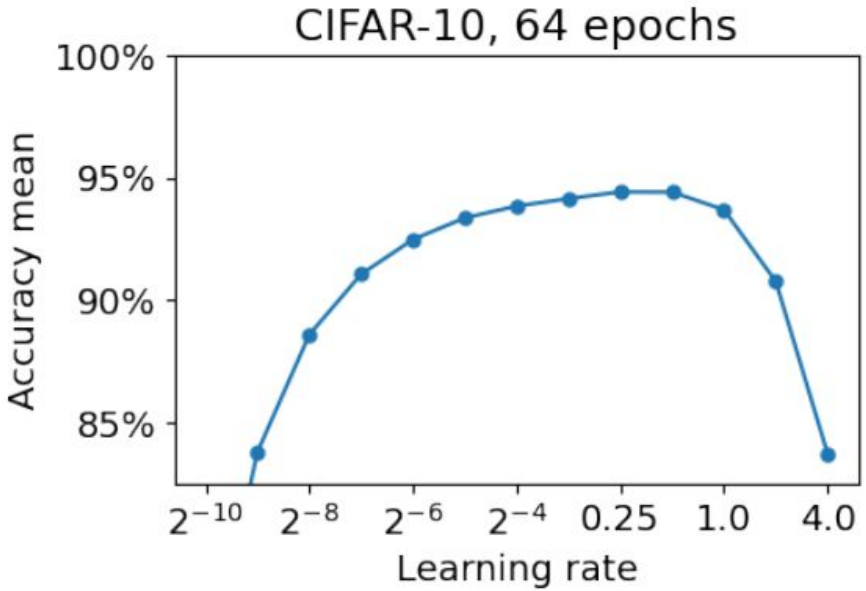
Исследуем 2 ситуации:

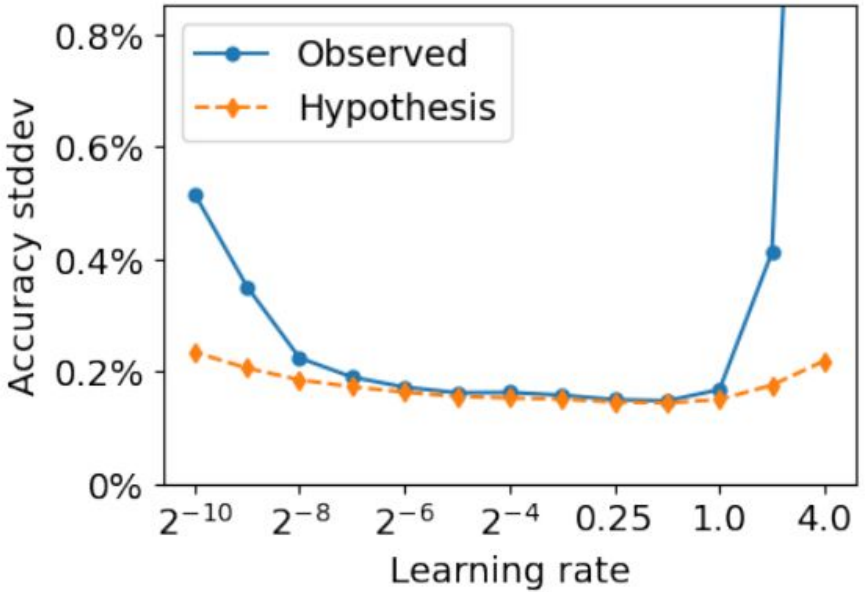
1. Не делаем аугментаций (random flipping и random resized crop)
2. Используем только 20% данных, но делаем аугментации

Вывод:

- Аугментации существенно снижают дисперсию качества на тестовой выборке

Влияние Learning Rate



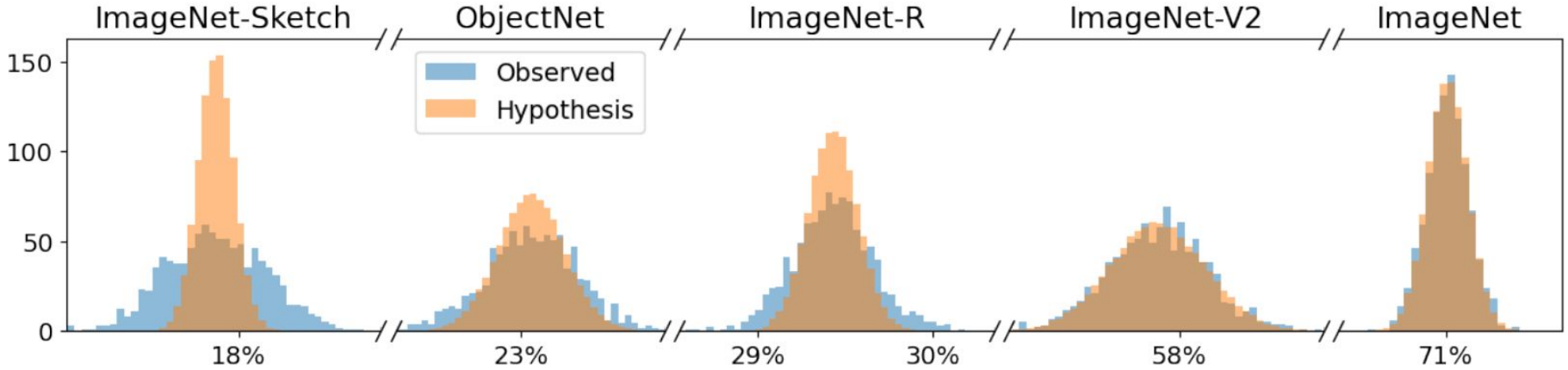


Сравниваем *среднее* качество и его *стандартное отклонение* по моделям на тестовой выборке

Вывод: при оптимальном Learning Rate достигается одновременно и наименьшая дисперсия на тестовой выборке.

Higher School of Economics
Faculty of Computer science

Влияние Distribution Shifts



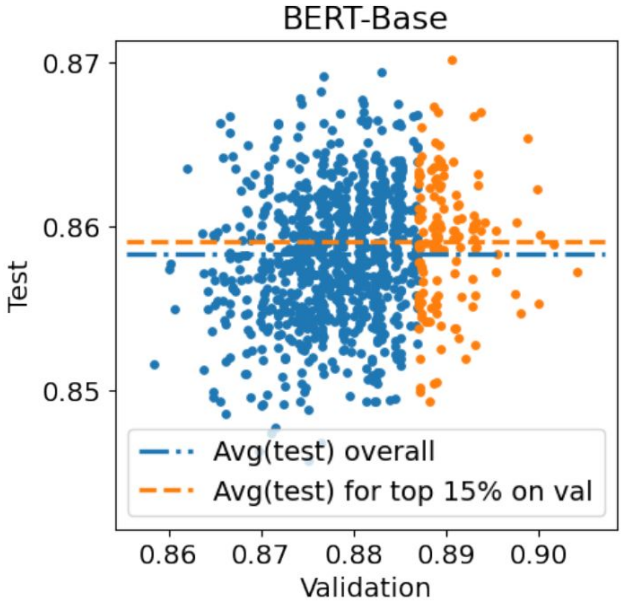
- Протестируем модели на смещенных распределениях
- Построим наблюдаемое распределение качества и гипотетическое при независимости ошибок на наблюдениях

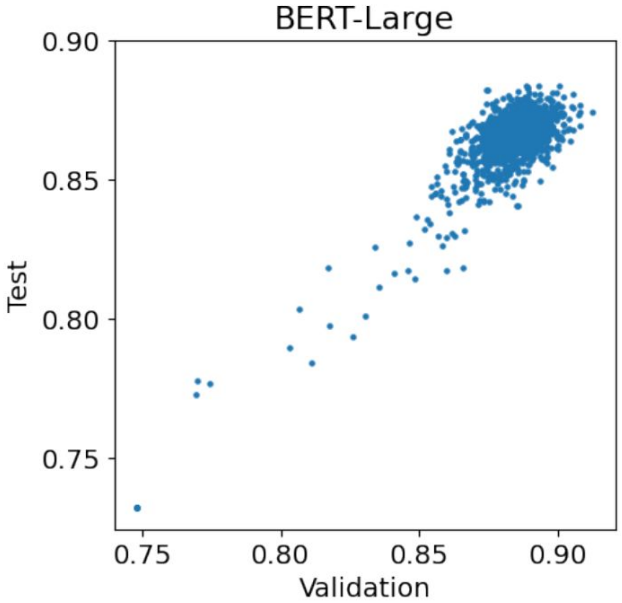
Вывод:

- На похожих distribution shifts распределение качества совпадает с гипотетическим, а на очень далеких оно отличается (как это было с тренировкой на небольшое число эпох)
- Дисперсия качества на смещенных распределениях существенно выше

Higher School of Economics
Faculty of Computer science

Влияние размера модели





- Сравниваем 2 модели: большую и маленькую
- Делаем fine-tuning BERT-а на задачу классификации, являются ли предложения последовательными
- Измеряем качество на двух тестовых выборках

Вывод:

- Для небольшой модели качество на валидации не коррелирует с качеством на тесте
- Большая модель нестабильна, и от перезапуска обучения качество может существенно вырасти

Выводы

- Случайность при обучении сети вносится самим **процессом обучения**: нельзя выделить главный фактор в виде инициализации, порядка батчей и аугментаций.
- Нейросети, дающие высокое качество на одной тестовой выборке не гарантируют такого же высокого качества на другой выборке: **random seed перебирать не стоит.**
- Дисперсия качества может быть объяснена тем, что нейросеть **ошибается на каждом объекте независимо.**
- Дисперсия качества на распределении **ниже**, чем дисперсия качества на тестовой выборке: при обучении одной архитектуры мы будем получать примерно одинаковое качество на новых данных.
- Стоит делать **аугментации** и **подбирать Learning Rate**

Сильные и слабые стороны статьи

Сильные стороны:

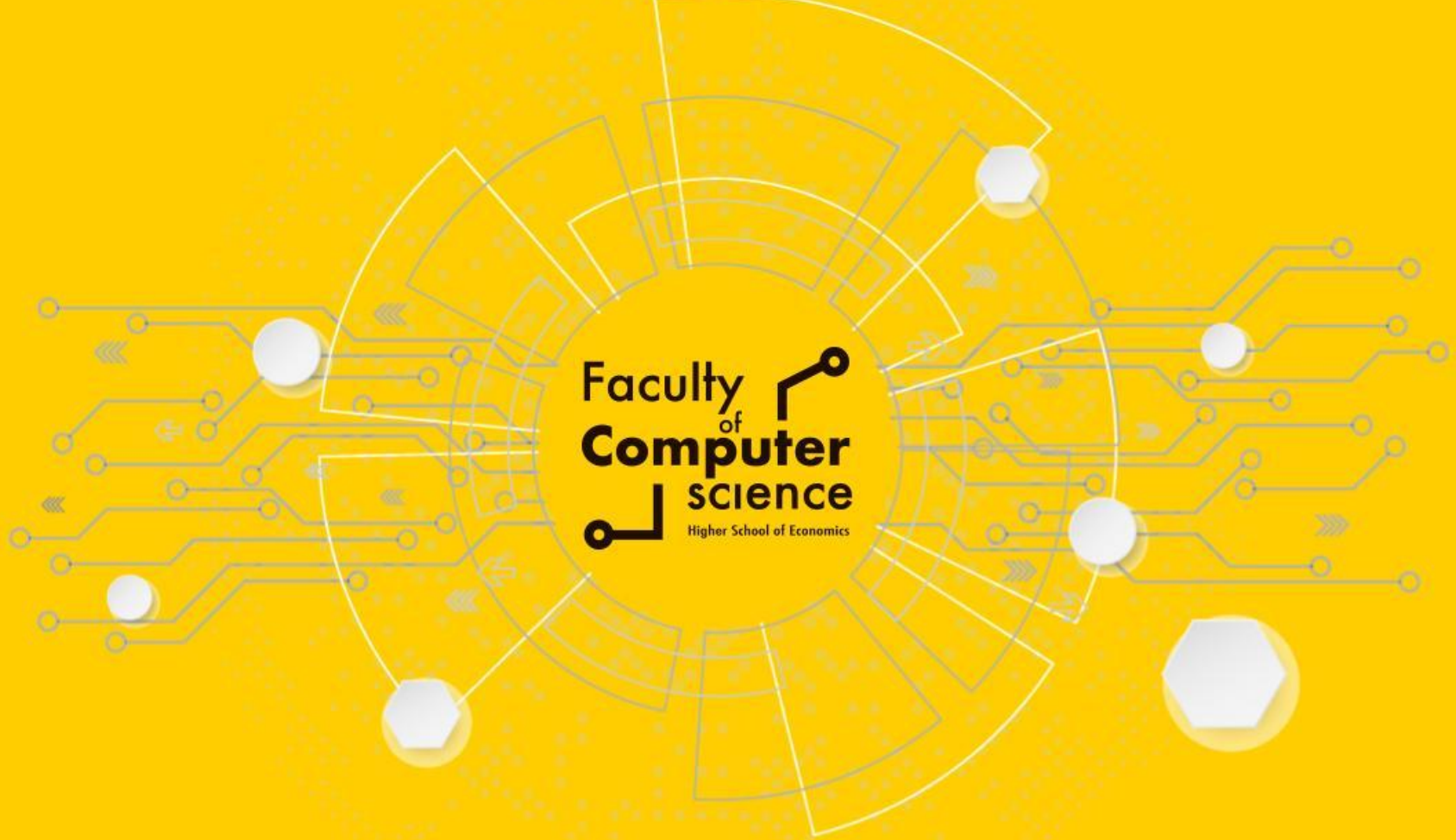
- Есть теоретическое обоснование практических результатов
- Сильный результат про случайность процесса обучения
- Эксперименты с теми вещами, которые не ясно как использовать на практике

Слабые стороны:

- В статье рассматривается только один вид модели: ResNet. Может быть эффект как с BERT-ом, когда размер модели влияет на результаты
- BERT в экспериментах тренируется только 3 эпохи. Обучение на большем числе эпох могло бы дать другие результаты
- Не затрагивается эффект ансамблирования

Материалы

Calibrated Chaos: Variance Between Runs of Neural Network Training is Harmless and Inevitable
https://arxiv.org/abs/2304.01910



mskazadaev@edu.hse.ru