Дистилляция данных

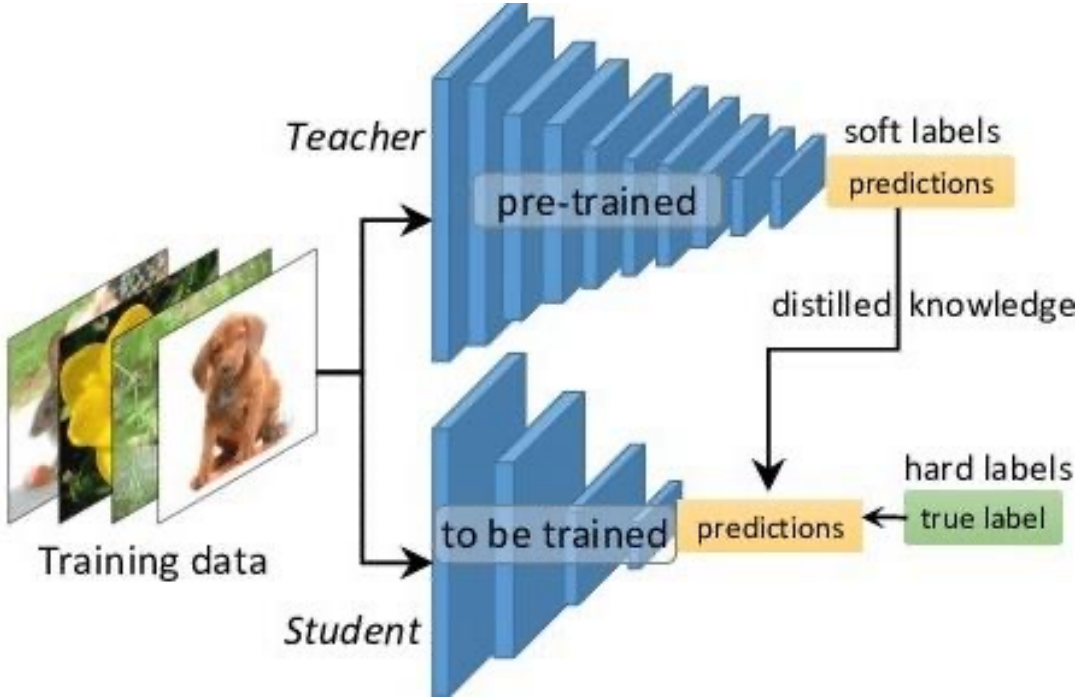
Тряпицын Саша, 192

https://arxiv.org/pdf/1912.07768.pdf

Виды дистилляции

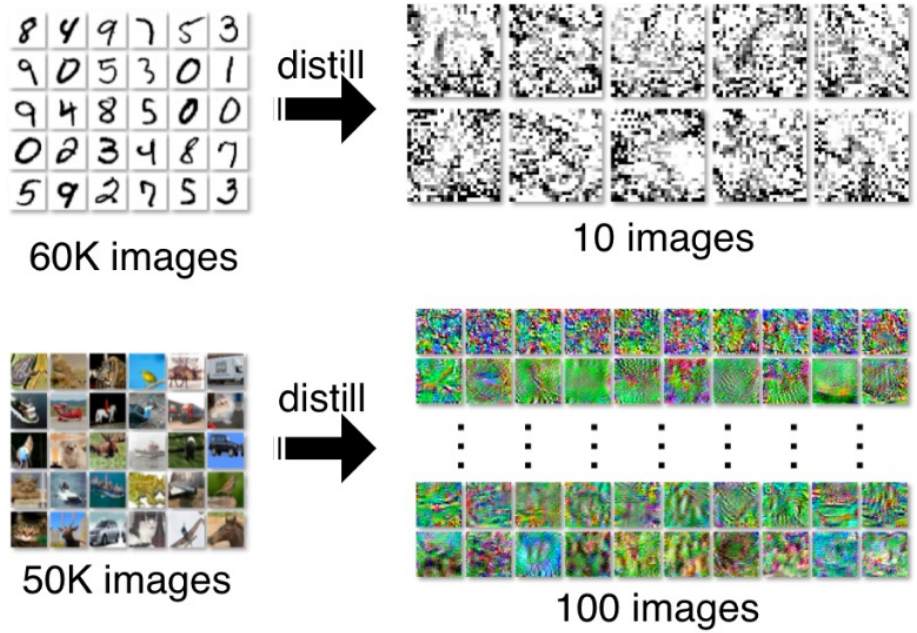
Дистилляция модели

- Уменьшает размер модели
- Ускоряет инференс



Дистилляция данных

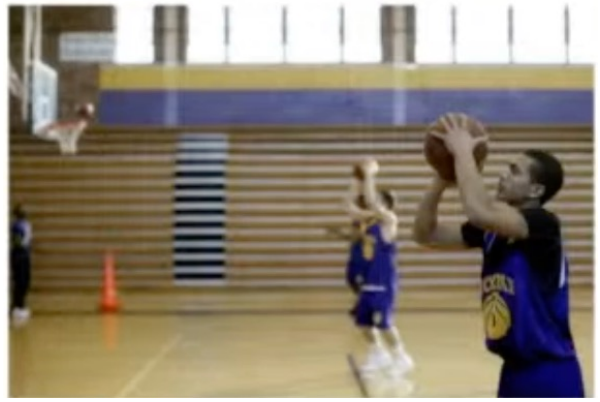
- Уменьшает объем данных
- Ускоряет обучение

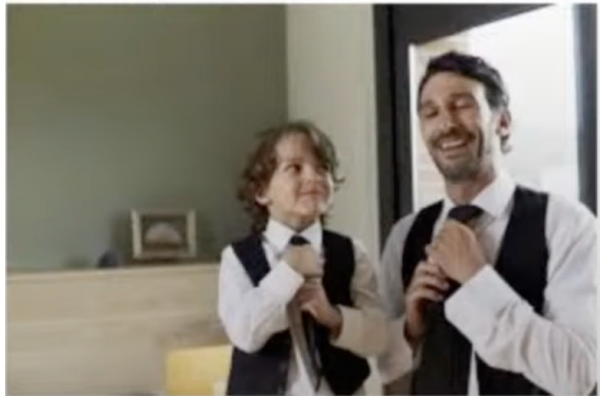


Как учится человек

- Методом проб и ошибок (традиционное обучение)
- Наблюдая за экспертом (дистилляция модели)
- Читая книгу (дистилляция датасета)

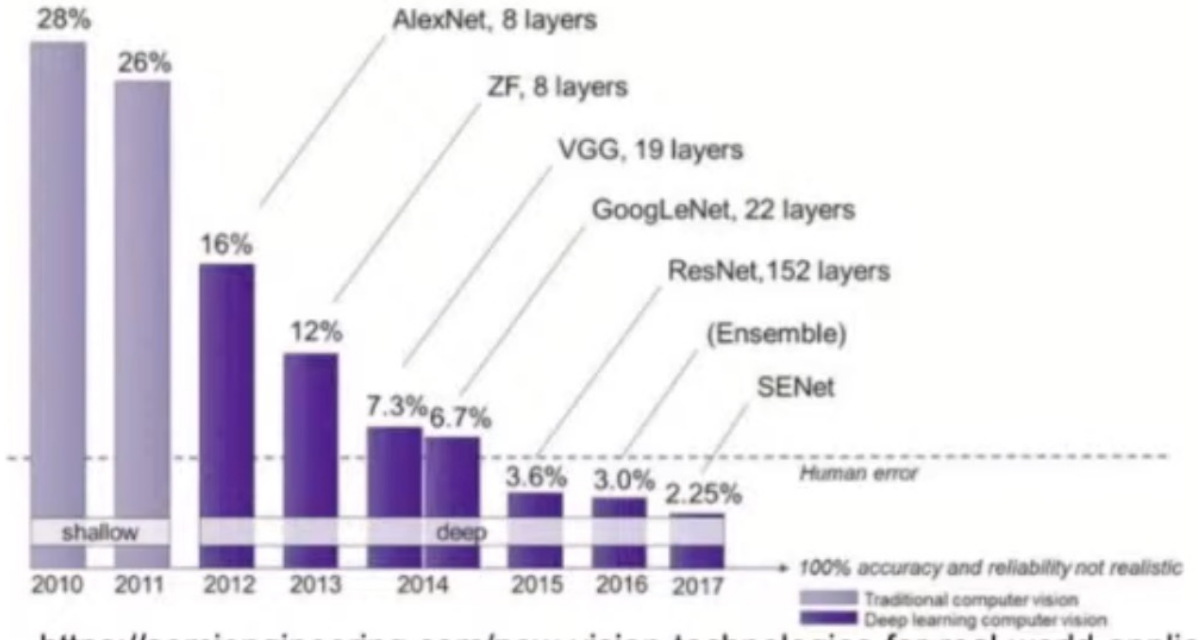
Магия чтения в том, что умение читать дает возможность приобрести любой навык, не связанный с чтением.





Применение дистилляции данных

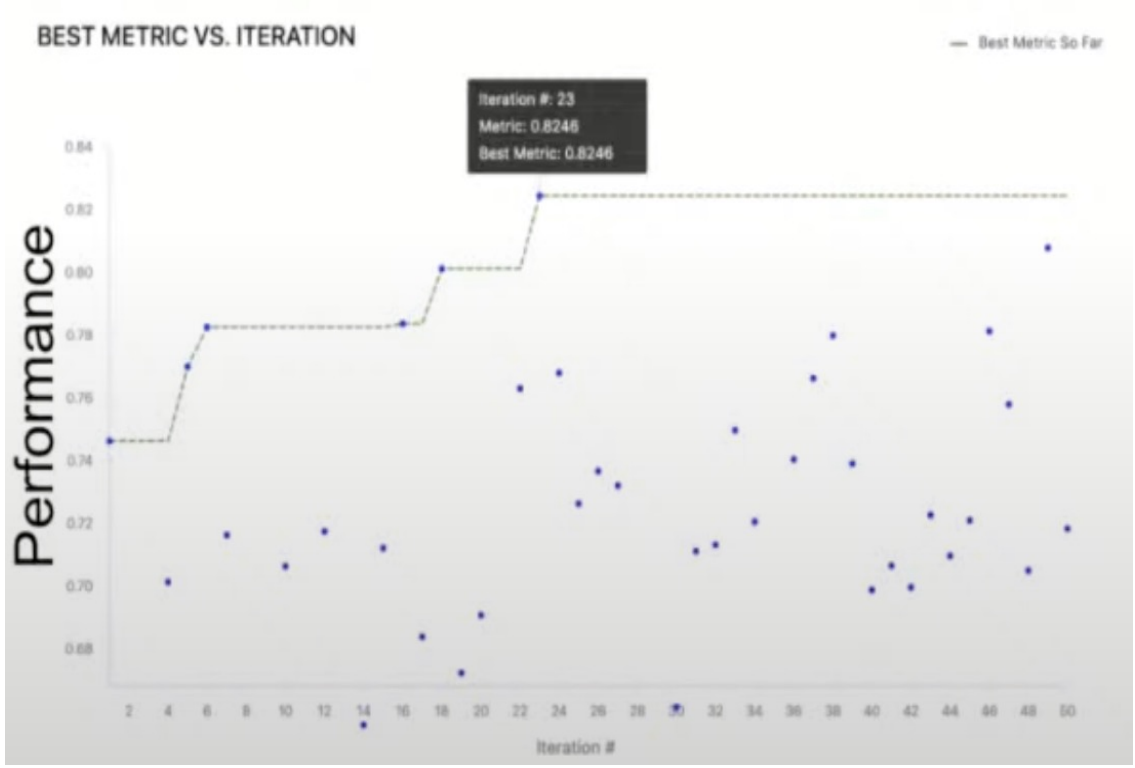
- Качество ML моделей растет с каждым годом за счет новых архитектур
- Появляется необходимость подбирать конфигурацию новых сложных архитектур под конкретные задачи



Применение дистилляции данных

- Чтобы протестировать новую конфигурацию, нужно обучить с нуля свежую модель

- Знания о данных никак не передаются между моделями из разных итераций

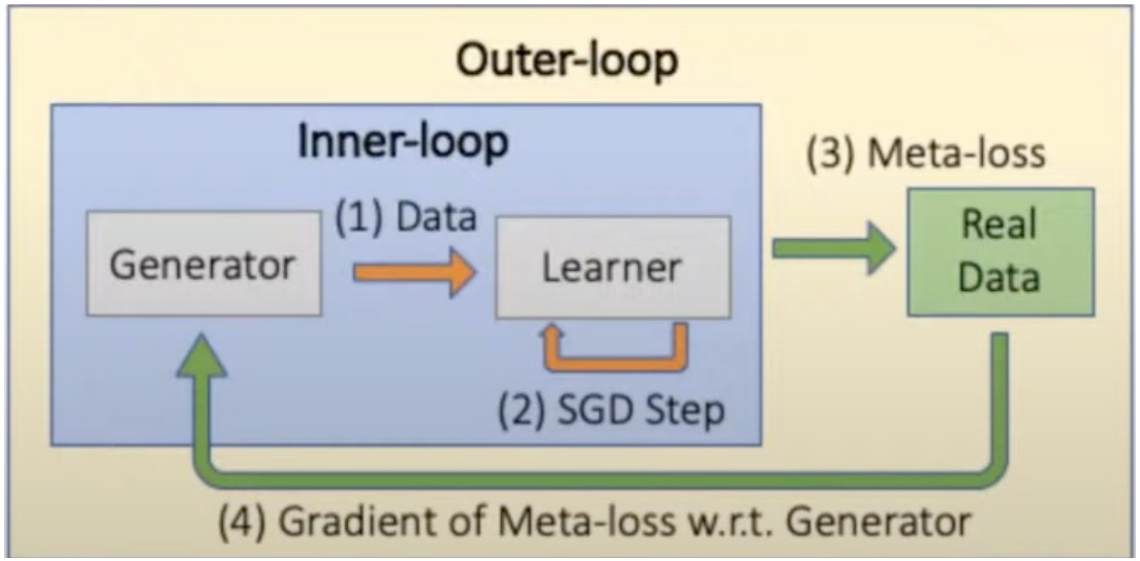


Способ дистиллировать данные

Раньше: найти самую «полезную» подвыборку данных

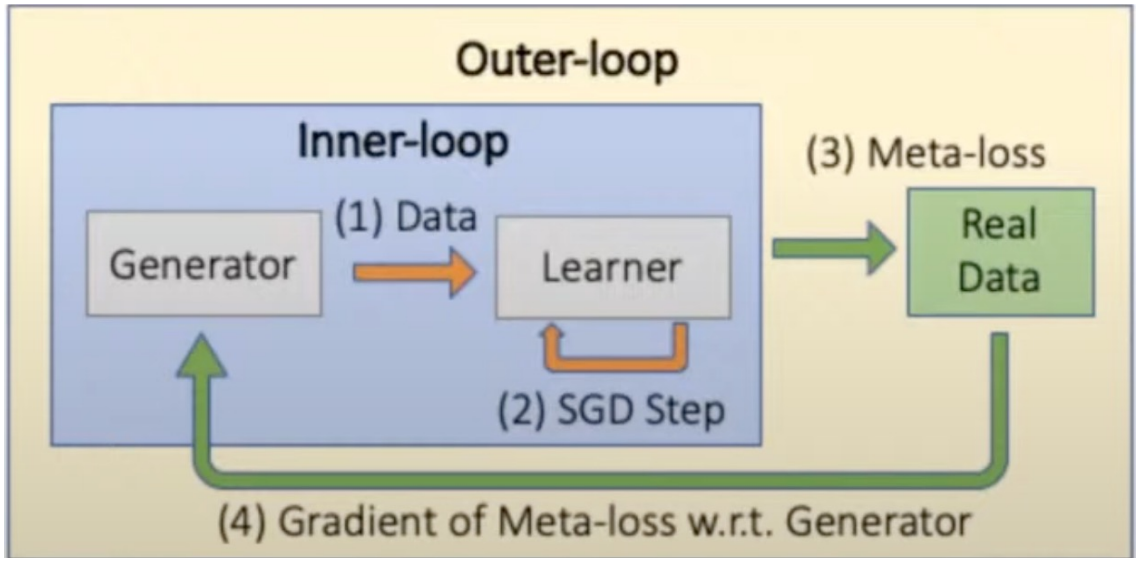
Сейчас: GTN = Generative Teaching Networks

- GTN генерирует батч данных для обучения модели
- GTN учится выдавать лучшие батч данных для целевой задачи



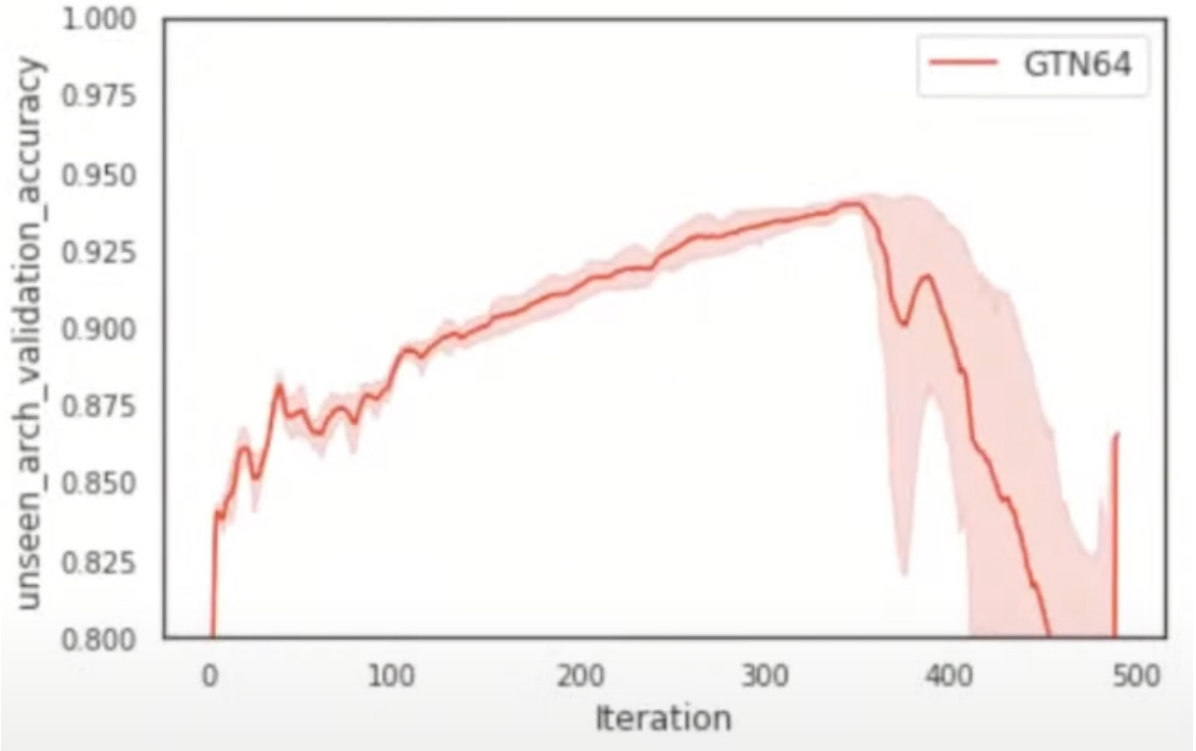
Как дистиллировать данные

- GTN по некоторому шуму генерирует данные
- Модель обучается на данных от GTN
- Считается мета-лосс модели на реальных данных
- GTN оптимизируется через мета-лосс



Проблема нестабильности

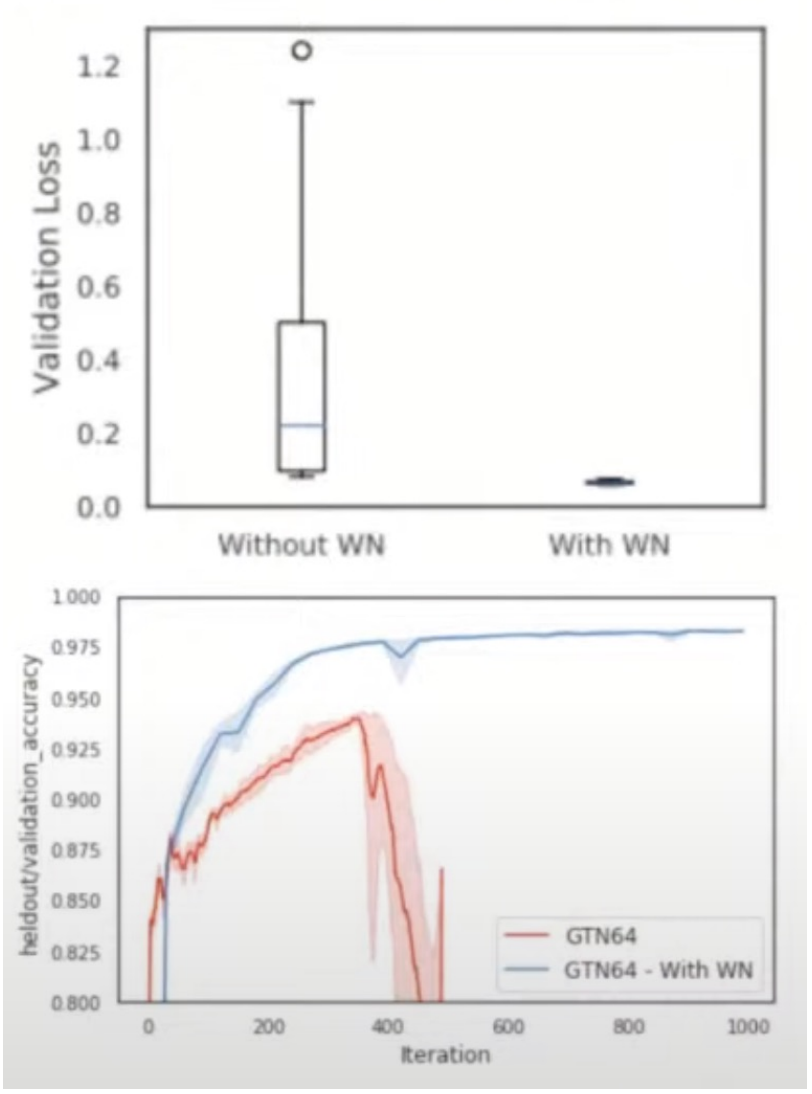
- Обучение GTN нестабильно
- Проблема похожа на взрыв градиентов



Разрешение нестабильности

- Нормализация весов модели

$$\mathbf{w} = \frac{g}{||\mathbf{v}||}\mathbf{v}$$



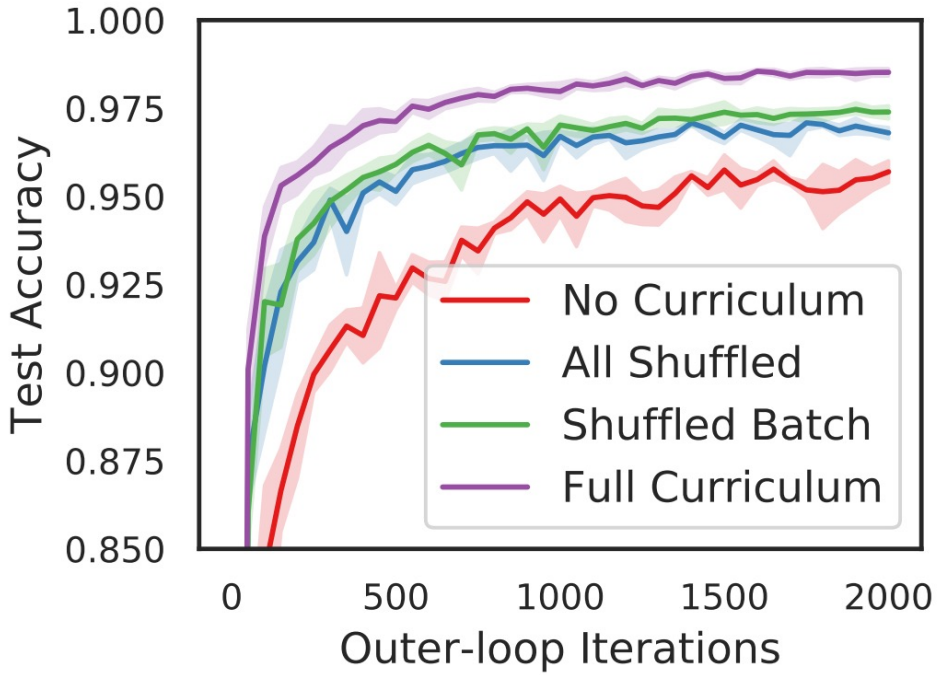
Salimans, Tim, Durk P. Kingma “Weight normalization: a simple reparameterization to accelerate training of deep neural networks”, 2016

Порядок данных от GTN

Интуиция:

- Сначала хочется учить более абстрактные закономерности

- Со временем хочется усложнять и конкретизировать примеры



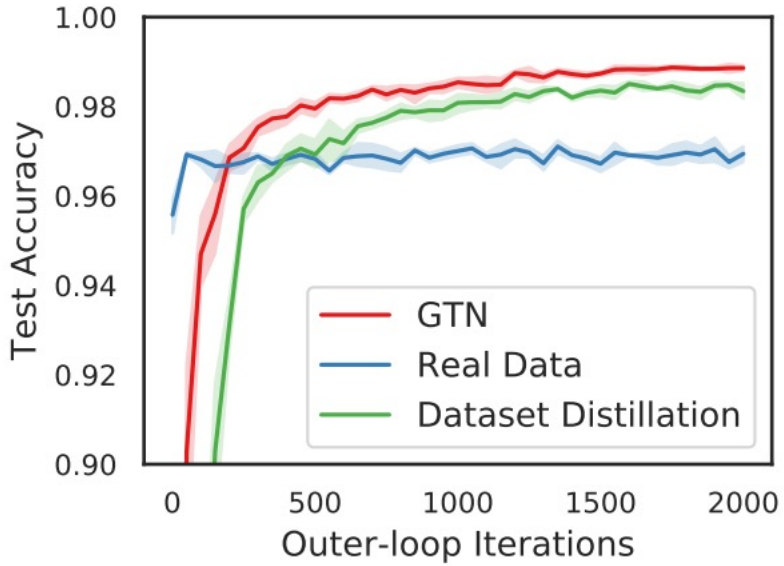
Поиск архитектуры модели

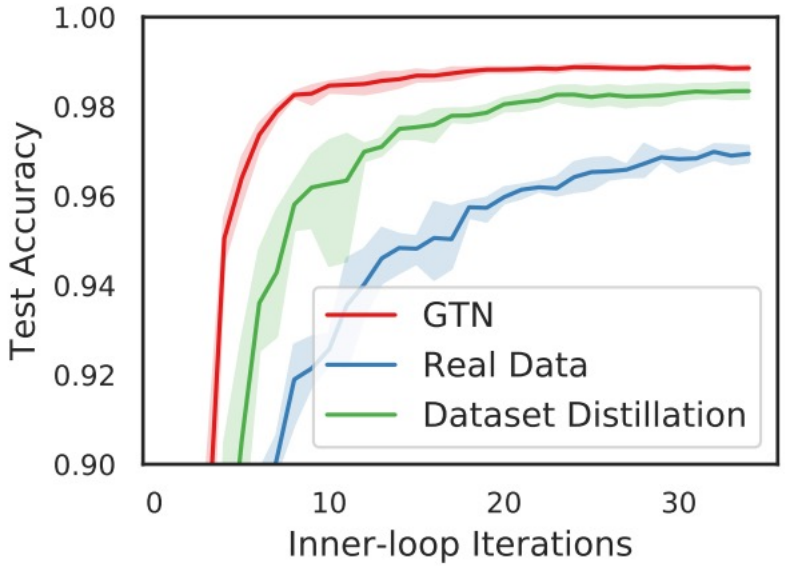
- Есть некоторый генератор архитектур, качество которых мы хотим померить
- Есть некоторый способ измерения качества архитектуры

Хочется ускорить второй пункт

Результаты обучения

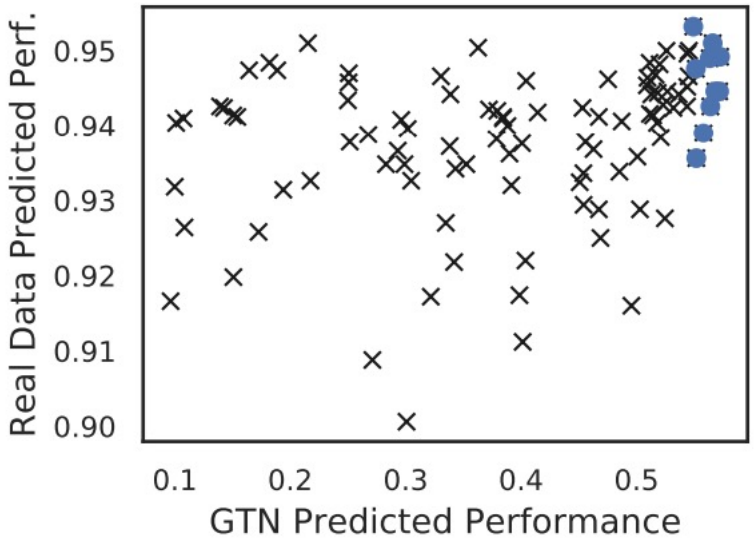
- Важно быстро получать обученную модель
- GTN справляется лучше классической дистилляции данных и отсутствия дистилляции вообще





Результаты обучения

- Результаты обучения на GTN должны коррелировать с результатами обучения на всех данных
- Это позволит выбирать архитектуру, лучшую по обучению на GTN



Заключение

- Дистилляция датасета – целая задача с разными подходами: классический и генеративный
- SOTA Решение решение для поиска архитектуры модели – обучение с помощью GTN

Model	Error(%)	#params	GPU Days
Random Search + GHN (Zhang et al., 2018)	4.3 ± 0.1	5.1M	0.42
Random Search + Weight Sharing (Luo et al., 2018)	3.92	3.9M	0.25
Random Search + Real Data (baseline)	3.88 ± 0.08	12.4M	10
Random Search + GTN (ours)	**3.84** ± 0.06	8.2M	0.67
Random Search + Real Data + Cutout (baseline)	3.02 ± 0.03	12.4M	10
Random Search + GTN + Cutout (ours)	**2.92** ± 0.06	8.2M	0.67
Random Search + Real Data + Cutout (F=128) (baseline)	2.51 ± 0.13	151.7M	10
Random Search + GTN + Cutout (F=128) (ours)	**2.42** ± 0.03	97.9M	0.67