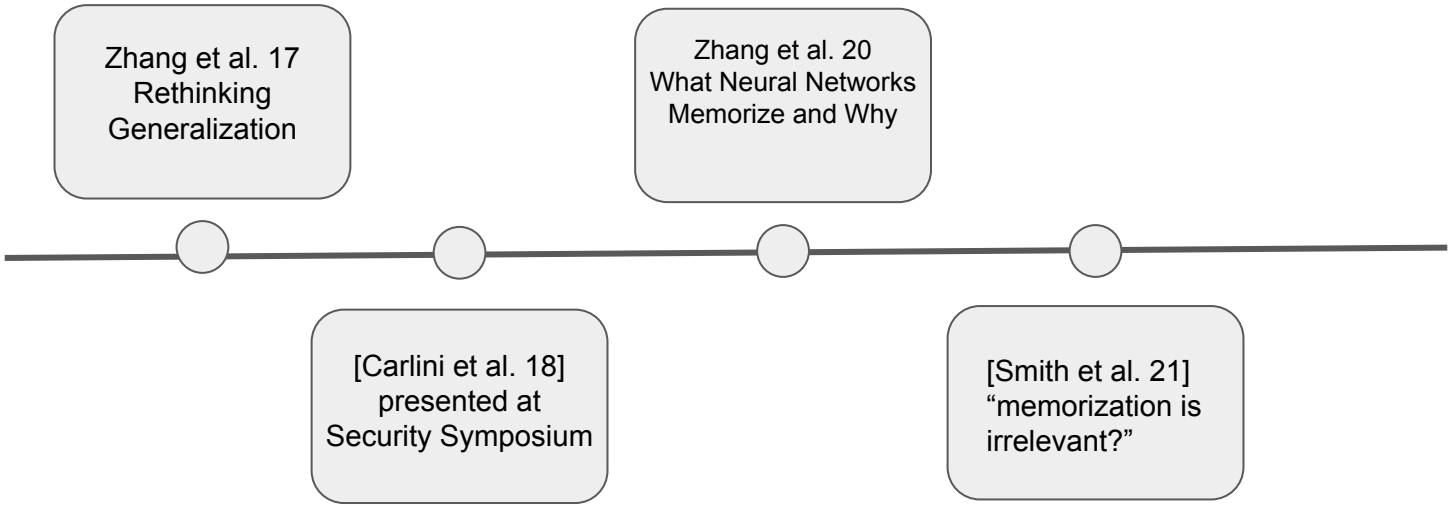
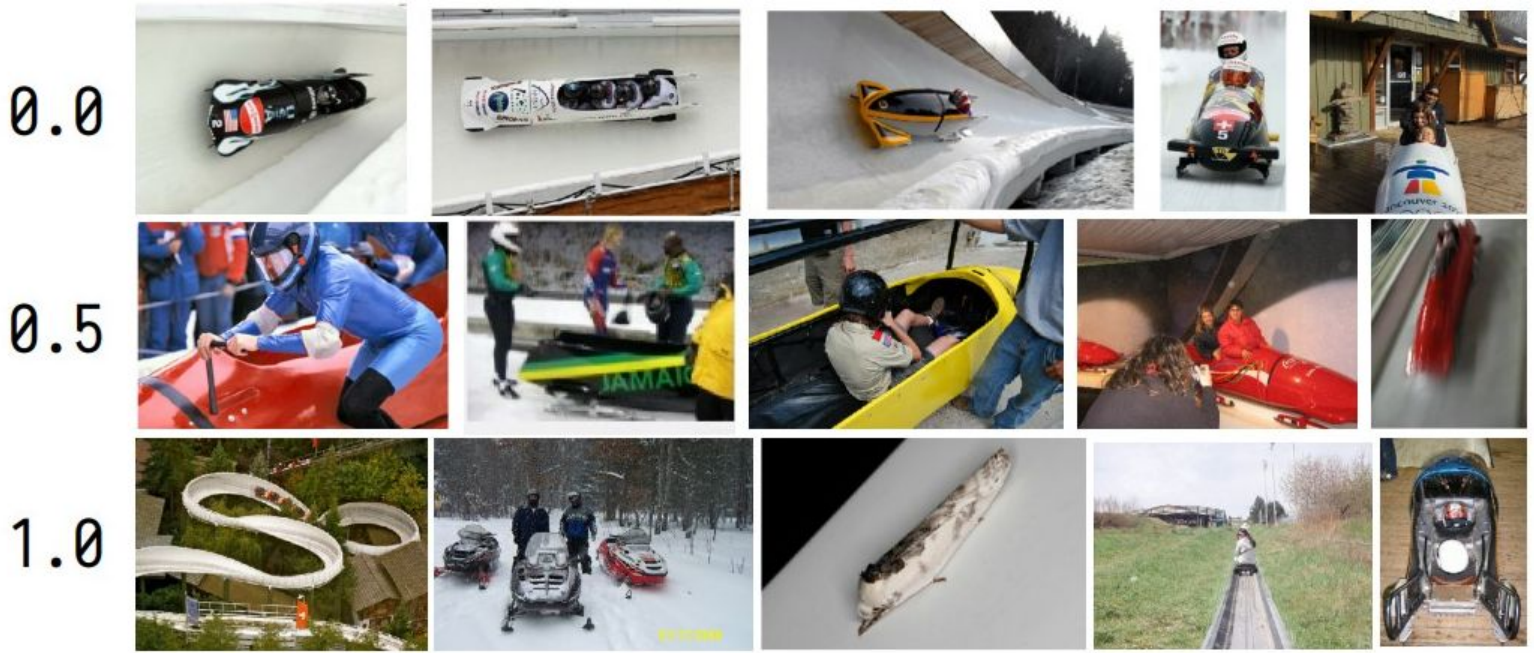
What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation



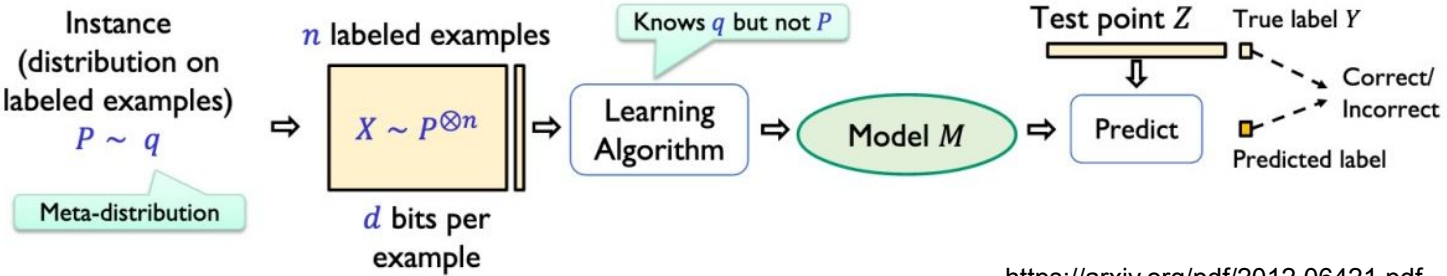
6 ноября, 2023
Александр Демин

Сеттинг



https://arxiv.org/pdf/2008.03703.pdf

Сеттинг

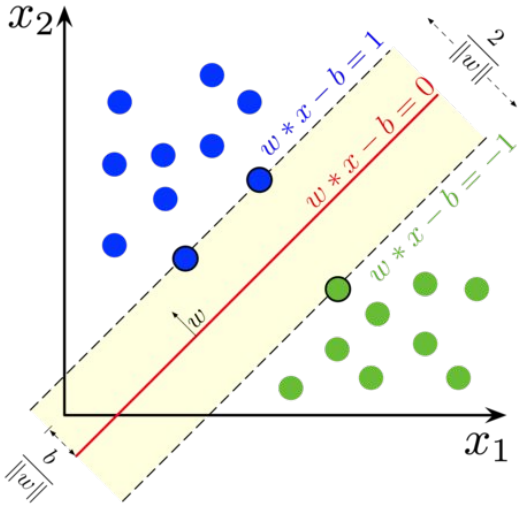


https://arxiv.org/pdf/2012.06421.pdf

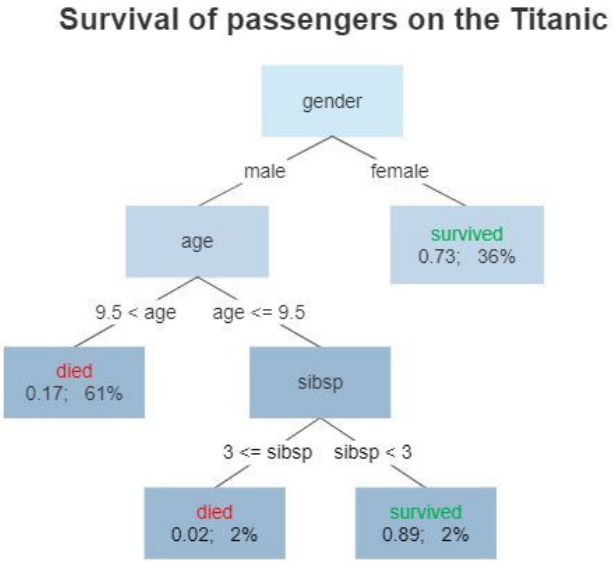
Меморизация: параметры обученной модели *M* содержат информацию о датасете *X*

- Когда меморизация полезна / вредна?
- Формализация меморизации

Явная “меморизация”

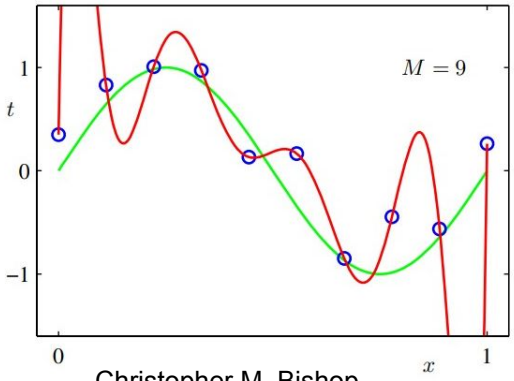


https://en.wikipedia.org/wiki/Support_vector_machine

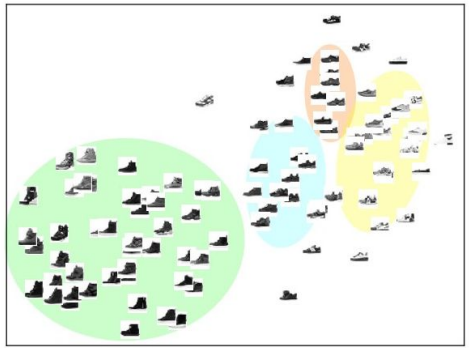


https://en.wikipedia.org/wiki/Decision_tree_learning

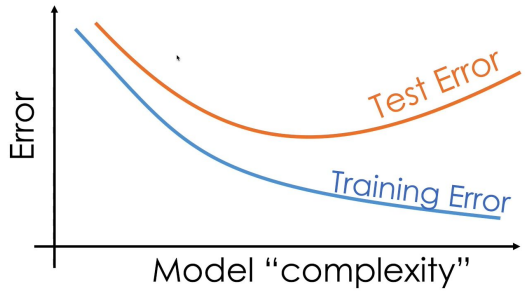
Меморизация и переобучение



Christopher M. Bishop,
Pattern Recognition and Machine Learning

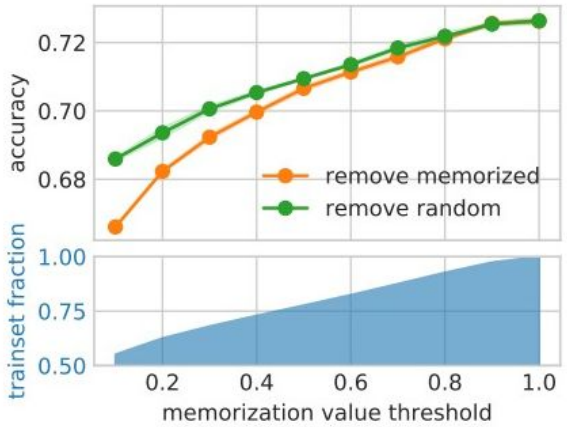


(a) **Memorized exceptions** in the Fashion-MNIST “sneaker” class.



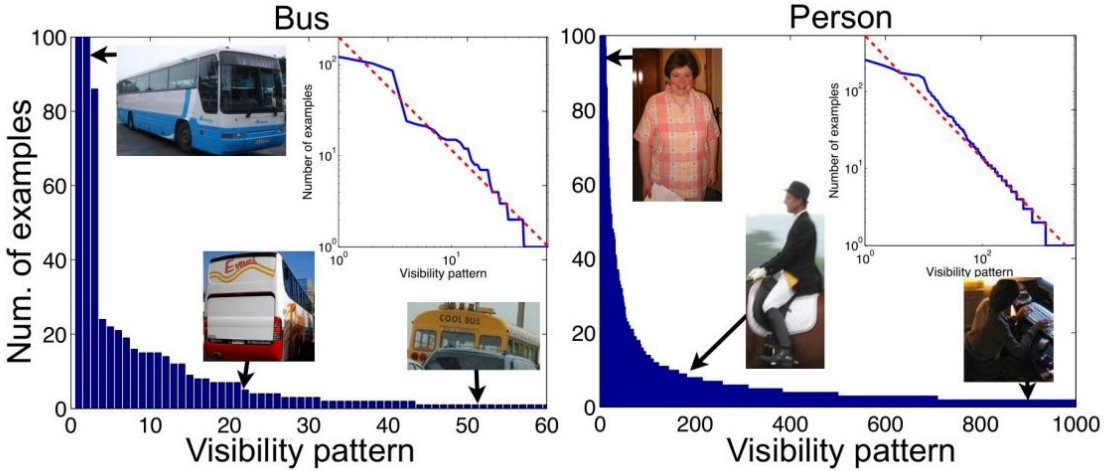
- Мы понимаем меморизацию как способность показывать хорошее качество на выбросах
- Нас интересует случай, в котором модель способна к обобщению
- Регуляризация снижает уровень меморизации [Arpit et el. 17]

Меморизация и метрики



Удаление запоминаемых примеров ухудшает качество

[Carlini et al. 19]



Получена *нижняя оценка* на качество модели [Feldman et al., 19]

–>

Оптимальная обобщаемость возможна лишь при полной меморизации

https://arxiv.org/pdf/2008.03703.pdf

Меморизация и приватность & копирайт

WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

https://xkcd.com/2169/

Artwork by Hollie Mengert (left) vs. images generated with Stable Diffusion DreamBooth in her style (right)
ttps://waxy.org/2022/11/

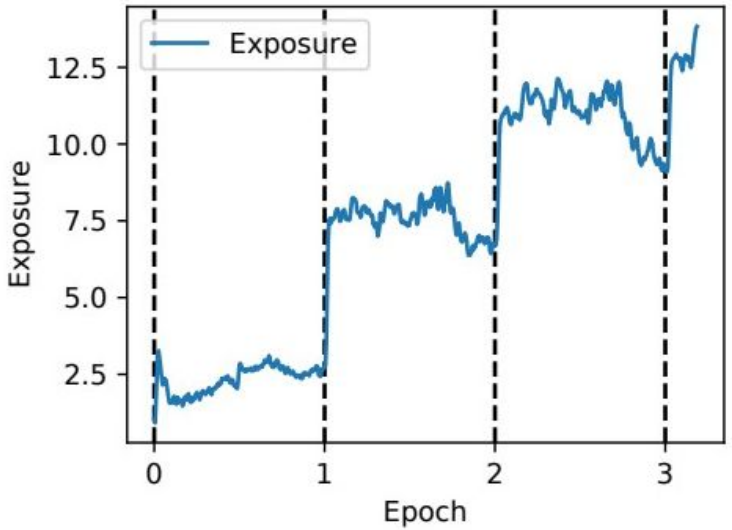
Меморизация и приватность & копирайт

Эксперимент [Carlini et al, 20]:

- датасет Penn Treebank (PTB)
- LSTM, 2 слоя, 600к параметров
- датасет аугментируется предложением *“My social security number is 078-05-1120”*

Результаты:

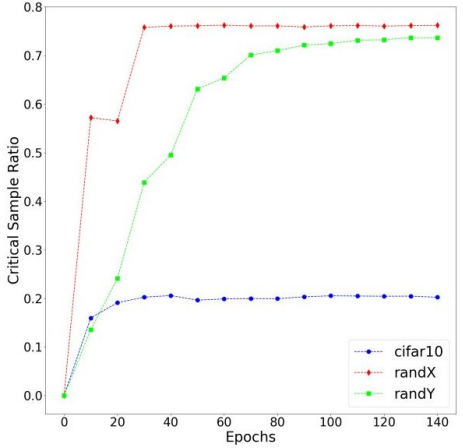
- Модель просят закончить предложение *“My social security number is 078..”*
- Авторы вводят метрику *exposure* для экспериментального изучения меморизации

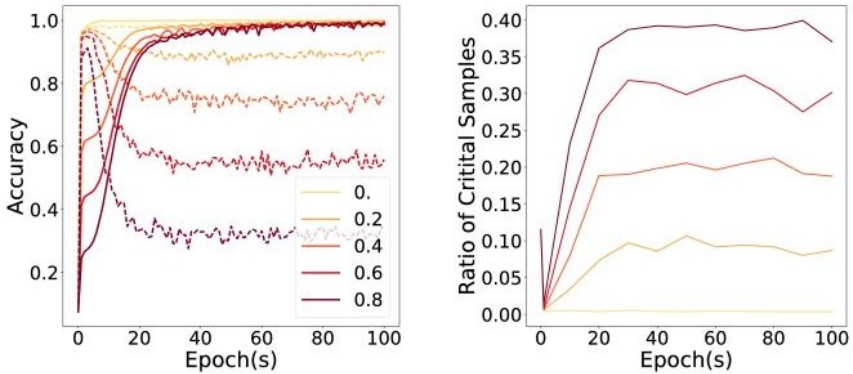


https://www.usenix.org/system/files/sec19-carlini.pdf

Меморизация и обучение

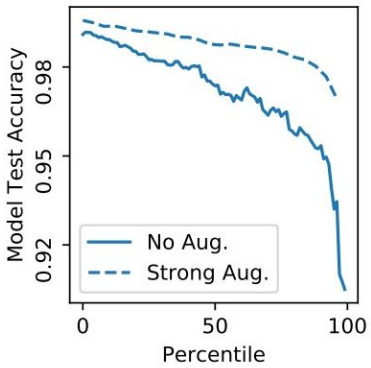
- Авторы вводят метрику *critical sample ratio* [Arpit et al. 17]
- Делается вывод о том, что нейронные сети склонны сначала запоминать общие паттерны в изображениях



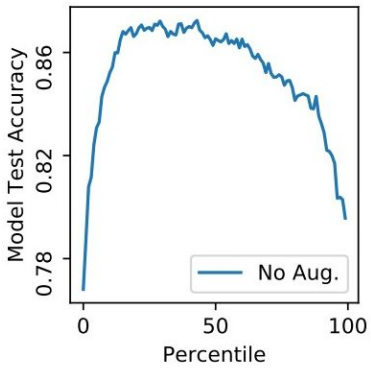


(b) Noise added on classification labels.

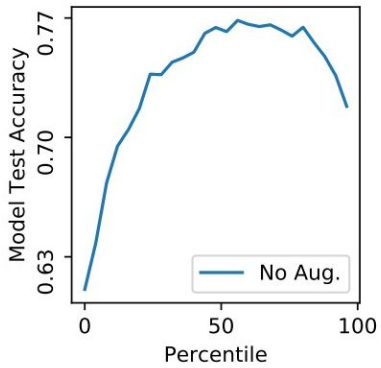
Меморизация и обучение



(a) MNIST



(b) Fashion-MNIST



(c) CIFAR-10

- Авторы вводят 5 (!) метрик, например, *Adversarial Robustness* [Carlini et al. 19]
- Получается оценить лишь меморизацию в общем, не поточечную