Can Neural Network Memorization Be Localized?

Докладчик: Борзыкин Валерий

Введение

- Что такое запоминание (memorization)?
- Запоминание не связано с целыми слоями
- Можно ли локализовать запоминание в отдельных нейронах?
  - Можно, и за это отвечает только небольшое количество нейронов (или каналов)
- Новый метод дропаута: Example-Tied Dropout
  - Точность на запомненных примерах падает со 100% до 3%
  - Обобщение увеличивается

Два этапа исследования и новый метод дропаута

- На уровне слоев
  - Исследование нормы градиента
  - Откат одного из слоев до состояния из предыдущих эпох
  - Инициализация одного из слоев стартовым значением и дообучение только на чистых данных
- На уровне нейронов
  - Последовательное удаление важных для примеров нейронов для поиска минимального количества, которые наиболее важны
- Example-Tied Dropout

Кратко про сетап

- Чтобы работать с зашумленными данными, случайно поменяем метку класса у 10% примеров от всего датасета
- Для тестов выбраны:
  - модели ResNet-9, ResNet-50 и ViT
  - датасеты для классификация изображений CIFAR-10, MNIST и SVHN
- Для ResNet-50 и ViT были объединены значения для нескольких слоев, чтобы упростить анализ и визуализацию

Исследования: сравнение градиентов

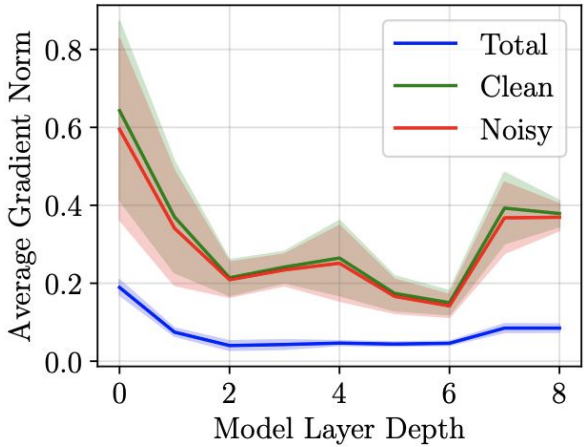


Figure 1. Gradient norm contribution from noisy examples closely follows that for clean examples even when they constitute only 10% of the dataset. Results depicted for epochs 15-20 for a ResNet-9 model trained on the CIFAR-10 dataset with 10% label noise.

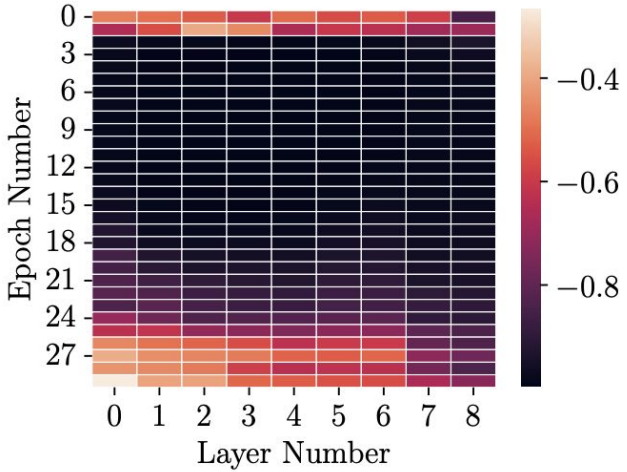
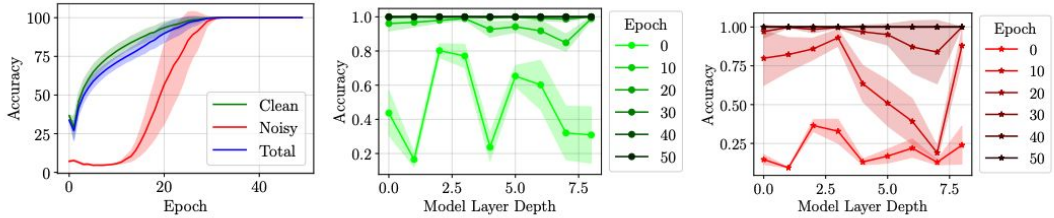
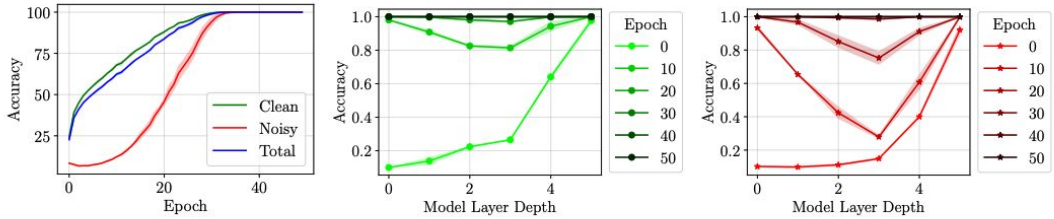


Figure 2. Cosine similarity between the average gradients of clean and mislabeled examples per layer, per epoch for ResNet9 model on the CIFAR10 dataset with 10% label noise. The memorization of mislabeled examples happens between epochs 10–30 (Figure 3).

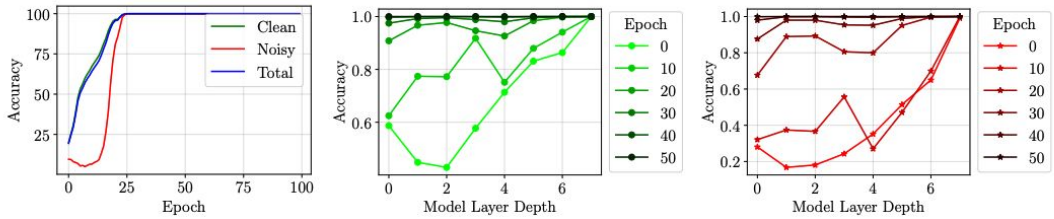
Запоминание на уровне слоев: layer rewinding



(a) Training and Rewinding curves for ResNet-9 Model trained on CIFAR10 dataset



(b) Training and Rewinding curves for ResNet-50 Model trained on CIFAR10 dataset



(c) Training and Rewinding curves for ViT (small) Model trained on CIFAR10 dataset

Запоминание на уровне слоев: layer retraining

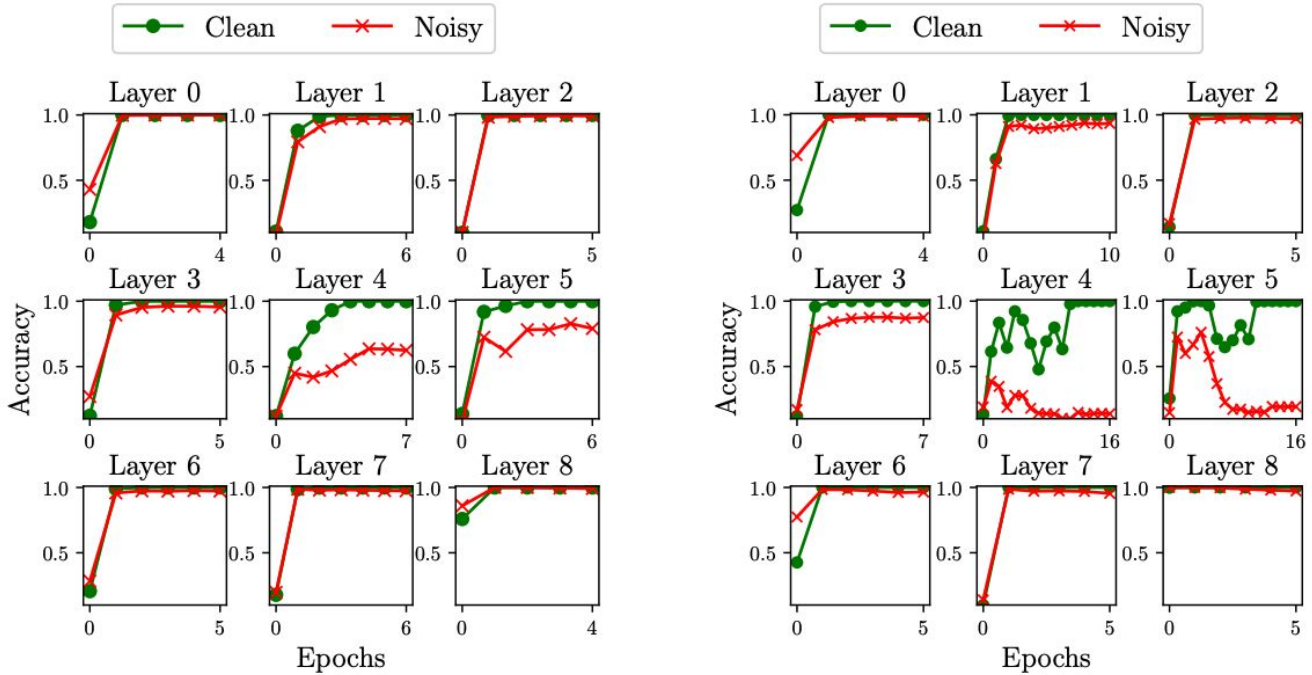


Figure 4. The impact of retraining individual layers from scratch on only clean examples, while keeping the rest of the model frozen. The results shown are for ResNet-9 model on the CIFAR10 (left) and MNIST (right) datasets with 10% random label noise.

Запоминание на уровне нейронов

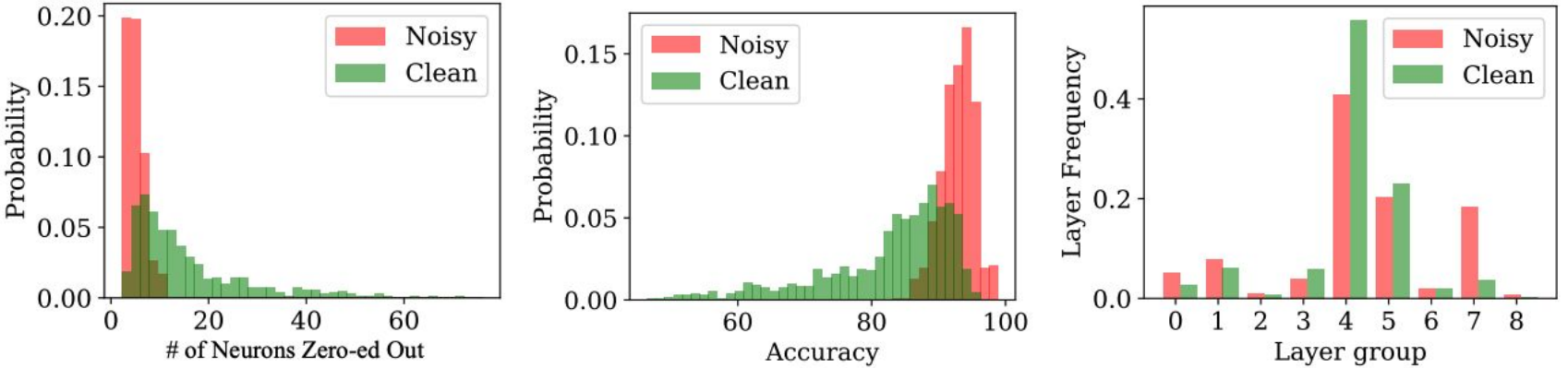


Figure 5. For each example in a subset of 1000 clean and 1000 noisy examples, we iteratively remove the most important neurons from a ResNet-9 model trained on the CIFAR-10 dataset with 10% random label noise, until the example's prediction flips. (a) Memorized examples need fewer neurons to flip their prediction. (b) Upon flipping the prediction, the drop in accuracy on the sample is much lower. (c) The most important neurons are distributed across layers in a similar way for both clean and mislabeled examples.

Example-Tied Dropout

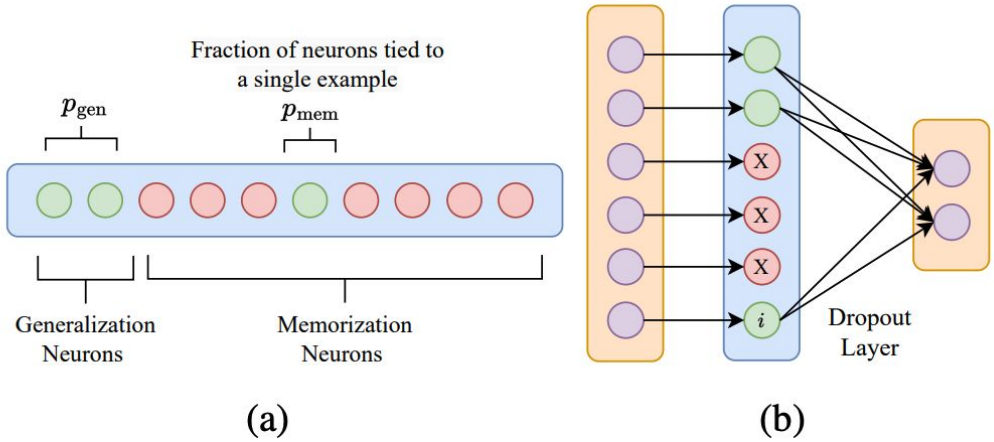


Figure 6. (a) A schematic diagram explaining the difference between the generalization and memorization neurons. At test time, we dropout all the memorization neurons. (b) Forward propagation for input tied to the i^{th} memorization neuron

Example-Tied Dropout

Dataset	Before Dropout			After Dropout		
	Clean	Noisy	Test	Clean	Noisy	Test
CIFAR10	99.9%	99.3%	79.3%	90.8%	3.1%	82.7%
MNIST	100%	100%	99.0%	99.2%	0.1%	99.3%
SVHN	99.9%	99.6%	89.5%	95.8%	1.4%	89.6%

Table 1. Dropping out memorization neurons leads to a sharp drop in accuracy on mislabeled examples with a minor impact on prediction on clean and unseen examples.

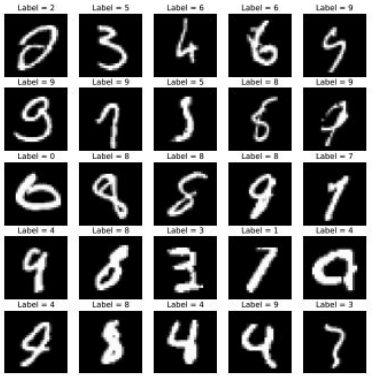


Figure 7. Most of the clean examples that are forgotten when dropping out the neurons responsible for memorization in the case of Example-tied dropout were either mislabeled or inherently ambiguous and unique requiring memorization for correct classification.

Плюсы и минусы

Плюсы:

- Исследовали влияние на запоминание и для уровня слоев, и для уровня нейронов
- Придумали способ находить зашумленные примеры в данных
- Новый метод дропаута Example-Tied Dropout, улучшающий обобщение

Минусы:

- Эксперименты только для датасетов с классификацией картинок
- Качество на тесте увеличивается несильно
- Сравнение с другими дропаутами только по точности предсказания зашумленных данных