Deep Nets Don’t Learn via Memorization

Our work
- Focuses on differences in learning noise/data
- Conclude DNNs don’t just memorize real data
- Training time is more sensitive to capacity and examples on noise
- Regularization can target memorization

Learning rate * number of iterations
- Affects capacity, and each other.
- Each parameter affects capacity, and each other.
- Capacity is limited by (1) learning algorithm and (2) regularization,
- Fitting noise requires more effective capacity.
- On real data, some examples are always/never fit immediately, and some examples have more/less impact on training (not so for noise).

Fig. 2. Change in normalized time to convergence as a function of dataset size, with capacity fixed at 4096 units. Because there are patterns underlying real data, increasing dataset size doesn’t increase training time for real data as much as it does for noise.

Fig. 3.2 Gin coefficient (a measure of roughness/disparity over categories) of the average loss-sensitivity over the course of training, on a 1000-example real dataset (~#parameters) is limited by (1) learning algorithm and (2) regularization, which impacts function complexity measured by critical sample ratio, function complexity increases very rapidly for noise data (red), increases eventually to almost the same level, for noise labels (green).

Fig. 4.1 Critical sample ratio for randomly chosen examples over the course of training on CIFAR-10, for noise input (randX, red) and noise labels (randY, green), and real data (blue). As measured by critical sample ratio, function complexity increases very rapidly for noise data (red), increases eventually to almost the same level, for noise labels (green).

What is memorization? Behaviour on random noise is a useful operational definition of memorization.

Deep nets can achieve 0 training error on datasets of random noise; does this mean their learning strategy is to memorize everything?

We perform a thorough empirical investigation of behaviour on real vs. noise data, and show this is not the case.

What we show that for deep nets:
- Fitting noise requires more effective capacity
- Training on noise gets harder, faster, when the dataset grows
- On real data, some examples are always/never fit immediately, and some examples have more/less impact on training (not so for noise).
- Simple patterns are learned first, before memorizing
- Regularization can effectively reduce memorization

Related work & Conclusions
Zhang et al. [1] raise questions about memorization and generalization in deep networks. We address these questions by providing insight on learning behaviour of deep nets. Comparing our work with [1]:
- Our work
  - Focuses on differences in learning noise/data
  - Conclude DNNs don’t just memorize real data
  - Training time is more sensitive to capacity and examples on noise
  - Regularization can target memorization

Goodfellow et al. [2] explain that a model's representational capacity (~parameters) is limited by (1) learning algorithm and (2) regularization, to become the effective capacity, and suggest learning rate*iterations as a measure. They note understanding effective capacity is difficult without understanding non-convex optimization.

We demonstrate that the data distribution is also an important consideration, which our proposed of critical sample ratio depends on. Understanding generalization requires thinking about how data, learning, and regularization affect capacity, and each other.

References