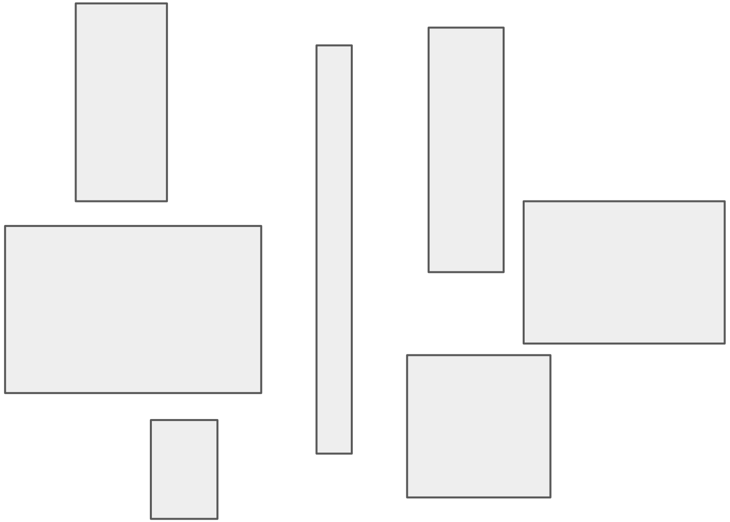
A Closer Look at Memorization in Deep Networks

Devansh Arpit*, Stanislaw Jastrzębski*, Nicolas Ballas*, David Krueger*, Maxinder Kanwal, Tegan Maharaj, Emmanuel Bengio, Asja Fischer, Aaron Courville, Yoshua Bengio, Simon Lacoste-Julien

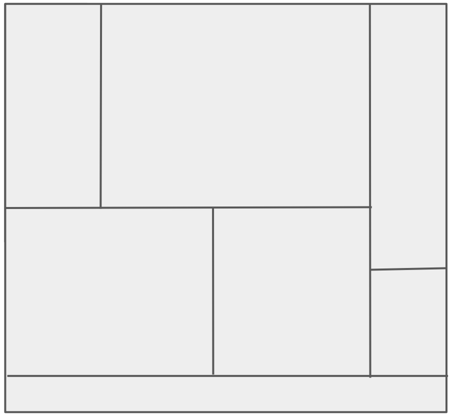
MILA

What is memorization?

Rote learning (memorization)



Meaningful learning (pattern-based)



- Memorization doesn't capitalize on patterns in data **(content agnostic)**
- Operational definition: **behaviour of DNNs trained on random data**

Context: “Understanding Deep Learning Requires Rethinking Generalization” - *Zhang et al. 2017 [1]*

- Shows: DNNs can fit random labels
  … so are DNNs using “brute-force memorization”?

Context: "Understanding Deep Learning Requires Rethinking Generalization" - *Zhang et al. 2017 [1]*

- Shows: DNNs can fit random labels
  … so are DNNs using "brute-force memorization"?
- My main take-away:
  We need data-dependent explanations of DNN generalization ability (...and recent work [2] provides this!)

[2] "Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data" Dziugaite and Roy (2017)

Compare and Contrast

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

Zhang et al. [1]

- Focuses on **similarities**
- **Suggests** DNNs might use memorization to fit data
- Training time increases by a constant factor on noise
- Regularization doesn't explain generalization

Overview of experiments:

1. Qualitative differences in fitting noise vs. real data
2. Deep networks learn simple patterns first
3. Regularization can reduce memorization

Notation:

1. randX - random inputs (i.i.d. Gaussian)
2. randY - random labels

Experiments (1a): Differences in fitting noise vs. real data

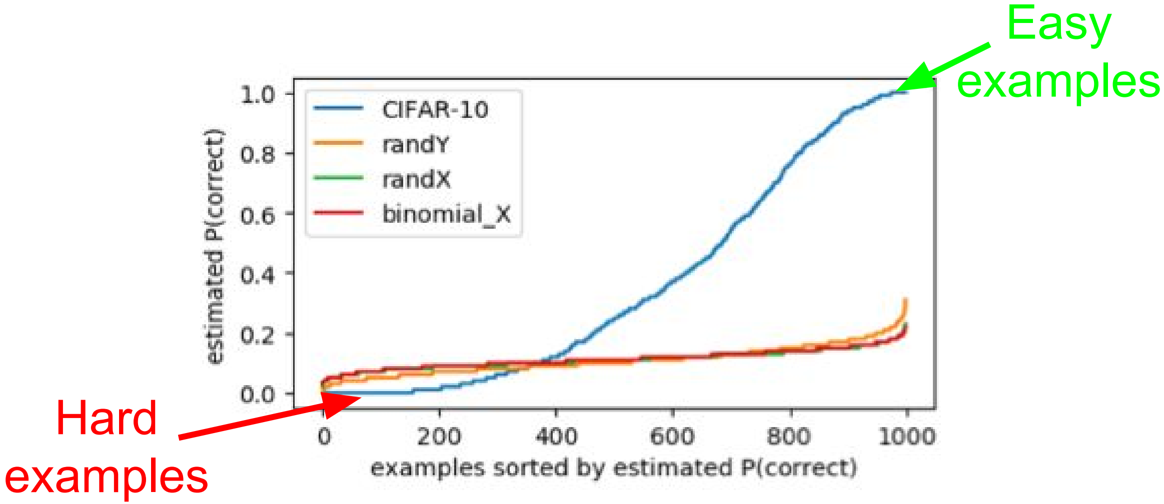


Figure 1. Average (over 100 experiments) misclassification rate for each of 1000 examples after one epoch of training.

Interpretation:

In real data, **easy** examples match **underlying patterns** of the data distribution; hard examples are **exceptions to the patterns.**

In random data, examples are all ~equally hard: learning is **content agnostic**

Experiments (1b): Differences in fitting noise vs. real data

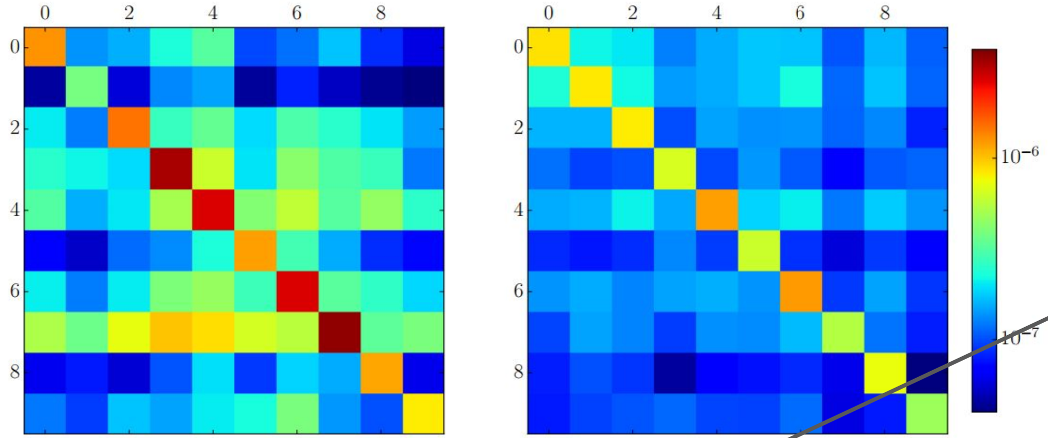
Interpretation:

Meaningful features can be learned by predicting noise

(see also: [3] "Unsupervised Learning by Predicting Noise." Bojanowski, P. and Joulin, A. ICML 2017)

Figure 2. Filters from first layer of network trained on CIFAR10 (left) and randY (right).

Experiments (1c): Differences in fitting noise vs. real data

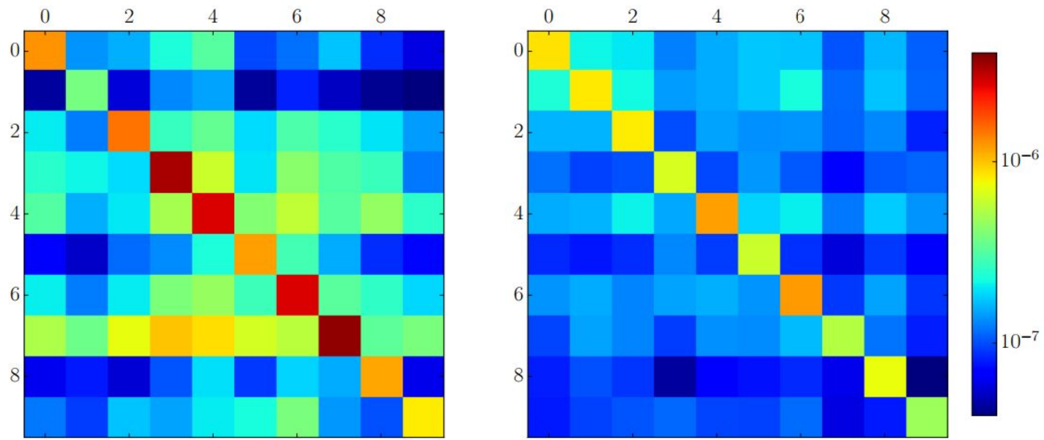


$$\bar{g_x} = \frac{\sum_t \left\| \partial \mathcal{L} / \partial x \right\|_1}{t}$$

Similar to [4] "Understanding black-box predictions via influence functions." Koh and Liang (ICML 2017)

Per-class **loss-sensitivity (g)**; a cell i,j represents the average loss-sensitivity of examples of class i w.r.t. training examples of class j. **Left** is real data, **right** is random data. Loss-sensitivity is more highly class-correlated for random data.

Experiments (1c): Differences in fitting noise vs. real data

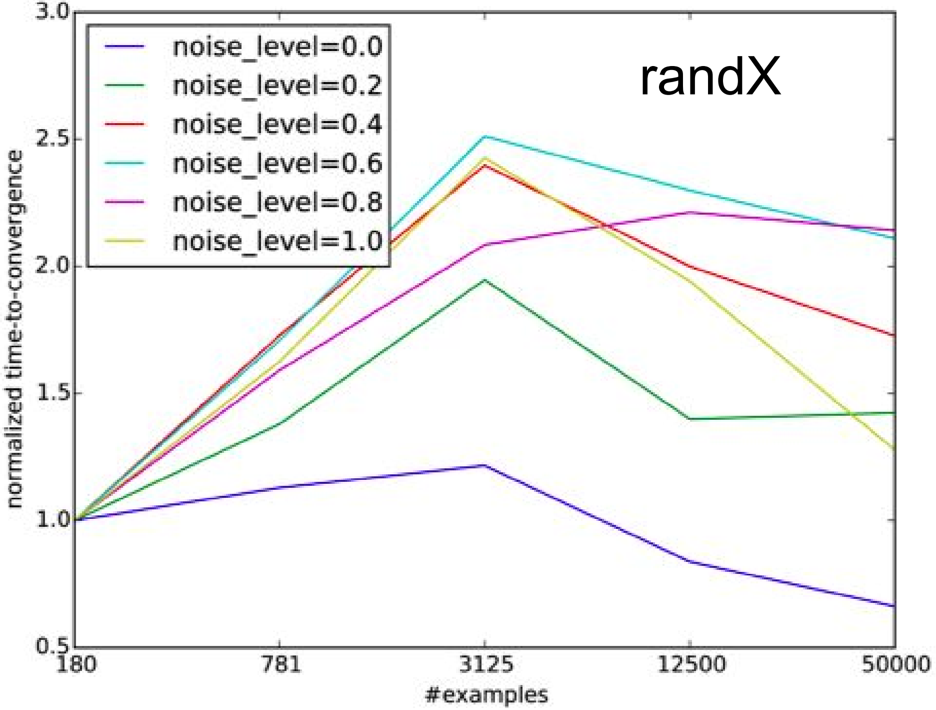


Interpretation:
On real data, more patterns (e.g. low-level features) are shared across classes.

(This is a selling-point of deep distributed representations!)

Per-class **loss-sensitivity (g)**; a cell i,j represents the average loss-sensitivity of examples of class i w.r.t. training examples of class j. **Left** is real data, **right** is random data. Loss-sensitivity is more highly class-correlated for random data.

Experiments (1d): Differences in fitting noise vs. real data



Interpretation:

Fitting more real data examples is easier because they follow meaningful patterns

(Note that this contradicts Zhang et al., who claim a constant factor slow-down on noise data!)

Experiments (2a): DNNs learn simple patterns first

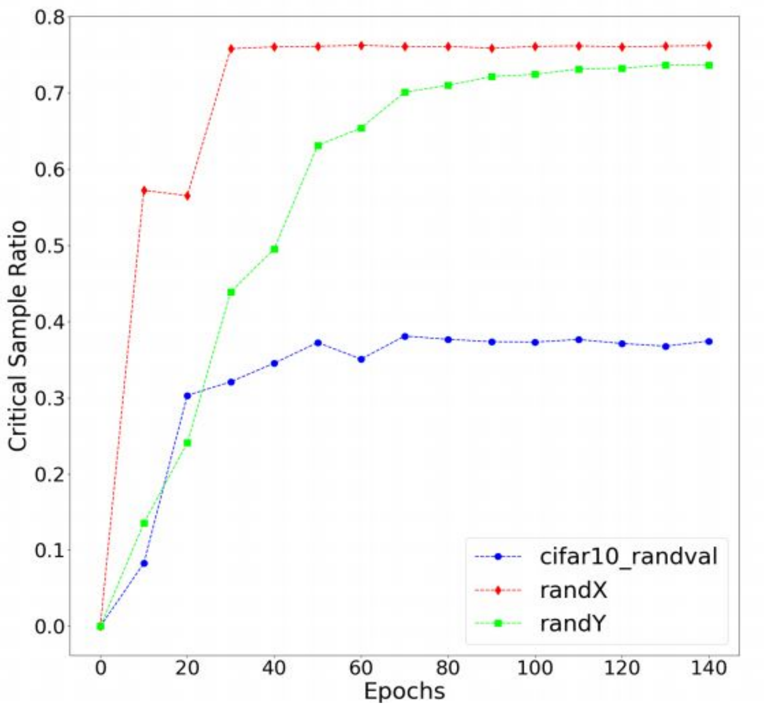
Critical sample ratio: how many data-points have an adversarial example nearby?

$$\arg\max_i f_i(\mathbf{x}) \neq \arg\max_j f_j(\hat{\mathbf{x}})$$

Interpretation:

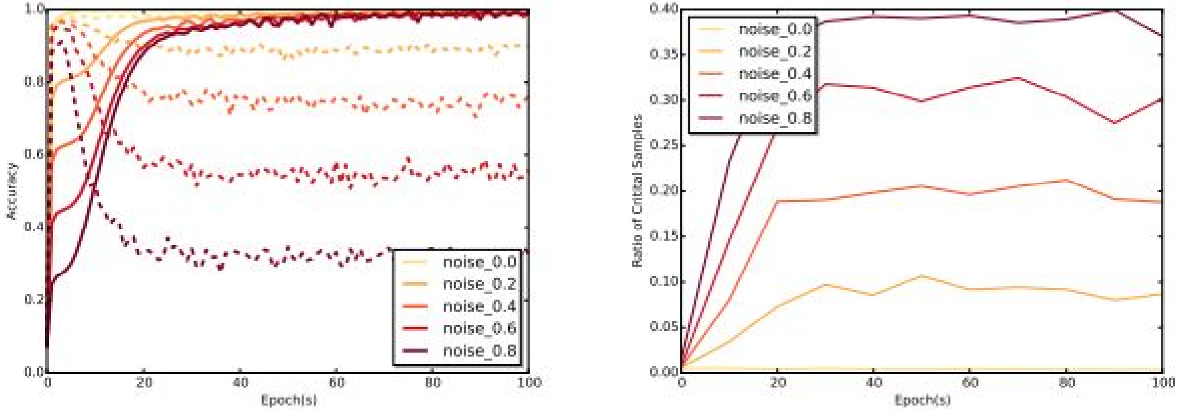
Learned hypotheses are less complex for real data

See [5] "Robust large margin deep neural networks." Sokolic et al.



Experiments (2b): DNNs learn simple patterns first

SOLID: trainset dashed: valid (real data only)



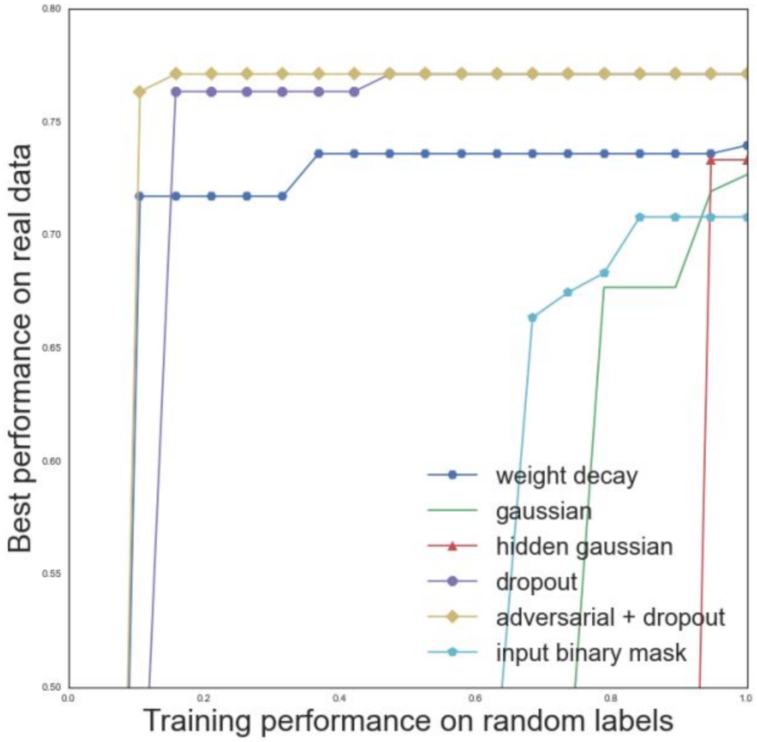
(b) Noise added on classification labels.

MNIST

Interpretation:

DNNs fit real data-points (which follow patterns) before fitting noise

Experiments (3): Regularization can Reduce Memorization



Interpretation:

We can severely limit memorization without hurting learning!

Adversarial training (+dropout) is particularly effective, supporting use of **critical sample ratio** to measure complexity

Conclusions

1. Qualitative differences in fitting noise vs. real data
2. Deep networks learn simple patterns first
3. Regularization can reduce memorization

QUESTIONS?

[1] “Understanding deep learning requires rethinking generalization.” Zhang, Chiyuan, Bengio, Samy, Hardt, Moritz, Recht, Benjamin, and Vinyals, Oriol. ICLR 2017 **(best paper award)**

[2] “Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data” Dziugaite, Gintaire and Roy, Daniel M. arXiv 2017

[3] “Unsupervised Learning by Predicting Noise.” Bojanowski, P. and Joulin, A. ICML 2017

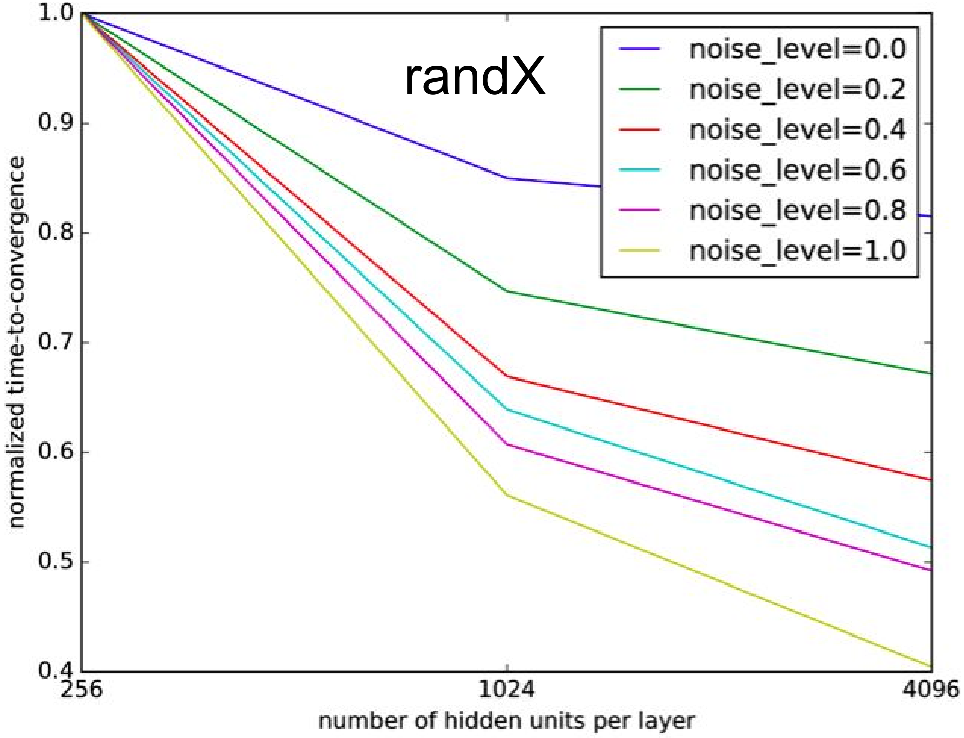
[4] “Understanding black-box predictions via influence functions.” Koh, Pang Wei and Liang, Percy. ICML 2017 **(best paper award)**

[5] “Robust large margin deep neural networks.” Sokolic, Jure, Giryes, Raja, Sapiro, Guillermo, and Rodrigues, Miguel RD. 2016.

[6] “Adversarial examples in the physical world.” Kurakin, Alexey, Goodfellow, Ian, and Bengio, Samy. ICLR 2017

Come to the poster (105) for even more experiments!!

Experiments (1e): Differences btw fitting noise vs. real data



Interpretation:

More effective capacity is needed to fit random data