Understanding Data Augmentation in Deep Learning and Beyond

Shuxiao Chen, Edgar Dobriban, Jane Lee

University of Pennsylvania

July 30, 2019

arxiv.org/abs/1907.10905. Slides: github.com/dobriban.

Overview

Deep learning and invariance

A framework for data augmentation

Summary

Appendix

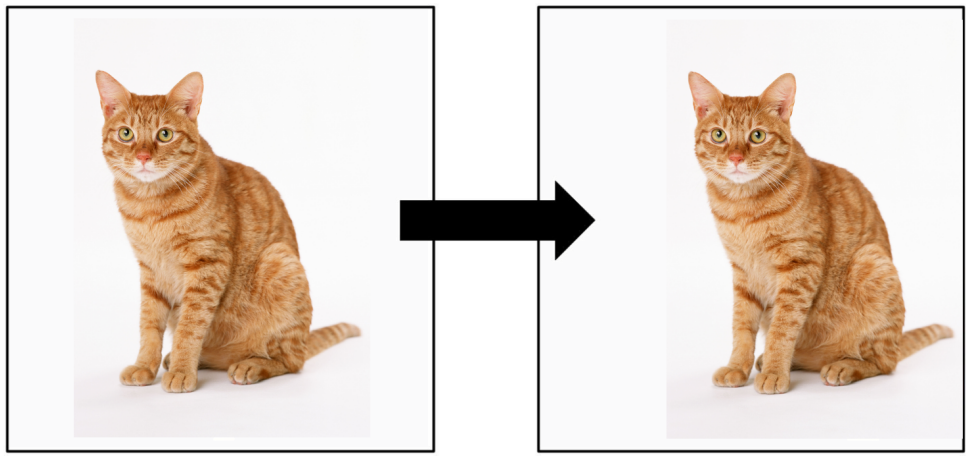
Overview

Deep learning

- Deep learning has achieved great successes in many areas
  - Vision: Image Classification, Image/Video segmentation, ...
  - Language processing: Translation, Voice recognition, ...
  - Reinforcement learning: Game playing, Robot navigation, ...
  - Many others...

Why has deep learning been successful?

- Main reasons: *Data* and *Compute*
- Also: Inductive bias (prior information, structure)
- e.g., Images are invariant to translations and rotations.



- We have a lot of tools for imposing structure:
  - Architecture design: convolutional neural networks (CNNs)
  - Data augmentation
  - Geometric constraints

Architecture design

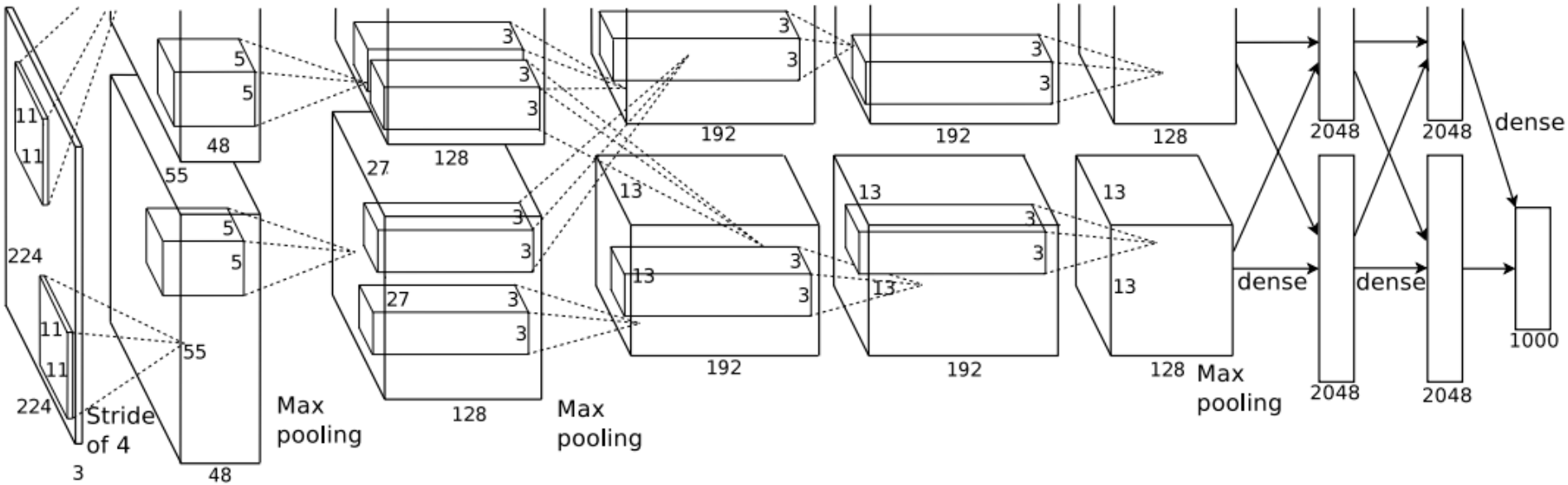


Figure: AlexNet. Krizhevsky et al (2012, NIPS), Won ILSVRC 2012.

Data Augmentation

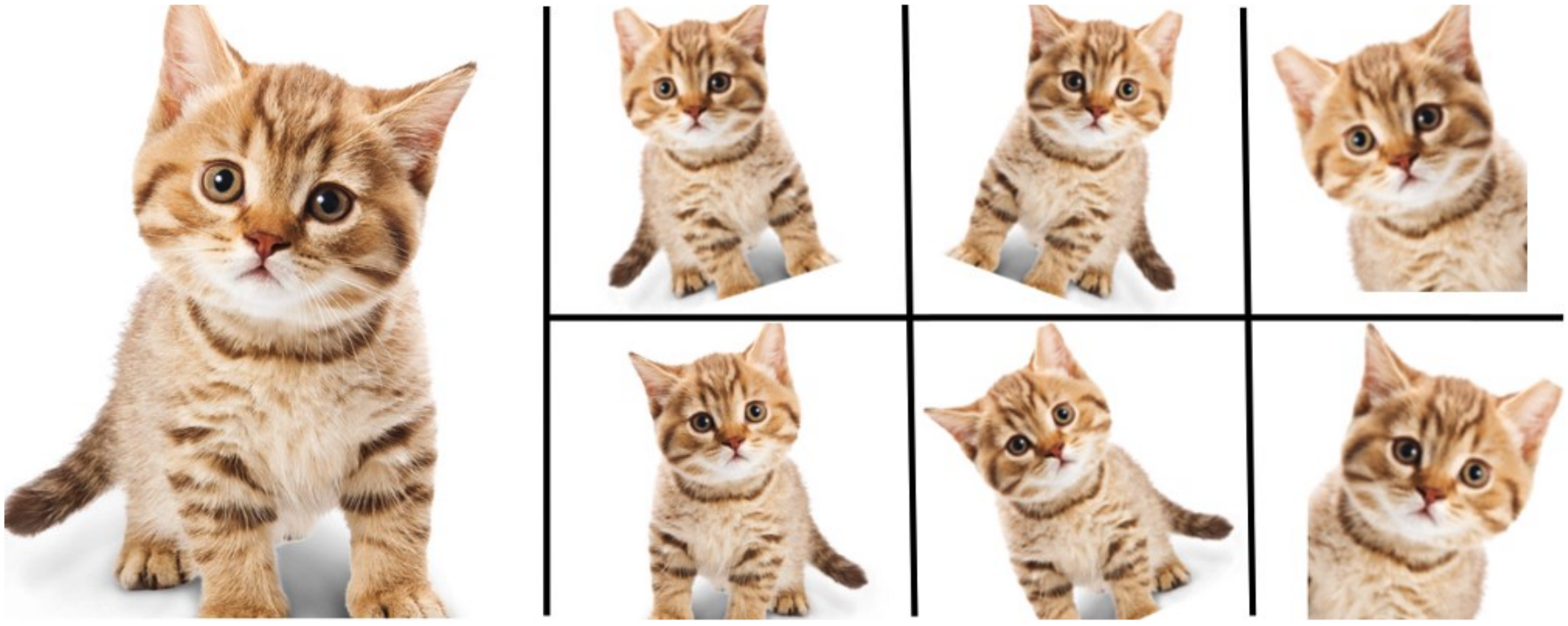


Figure: Examples of Data Augmentation: flipping, rotation, cropping.

Experiment

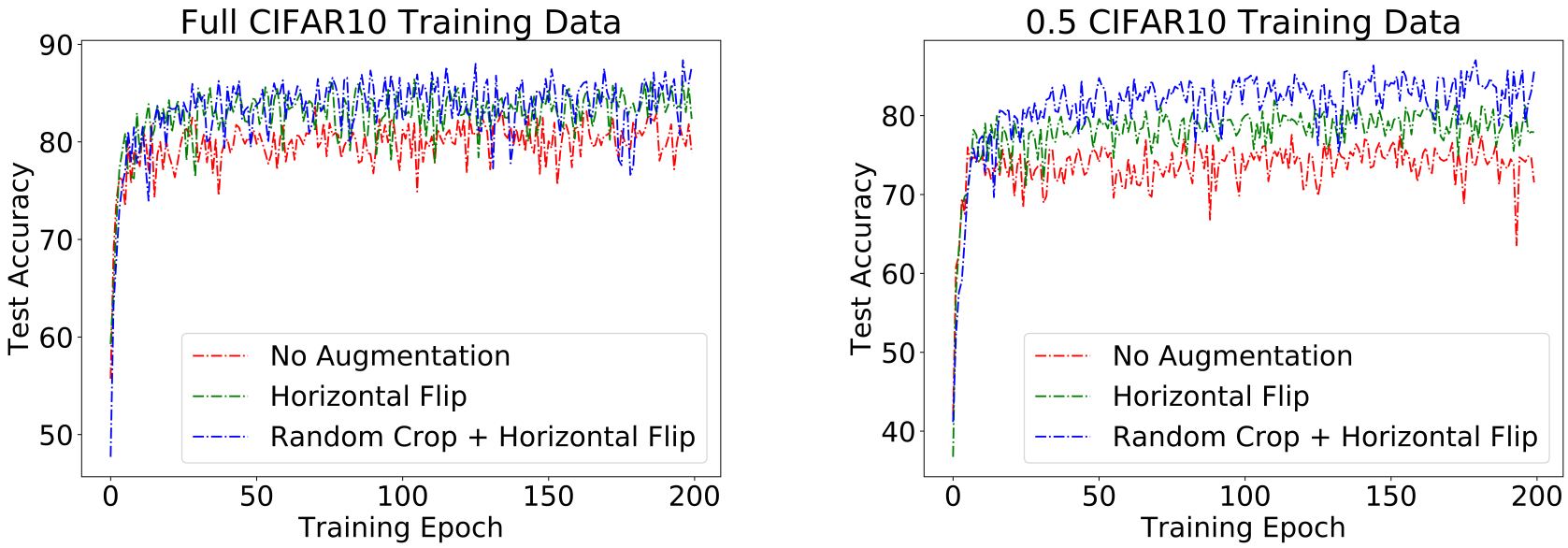


Figure: Benefits of data augmentation: A comparison of test accuracy across training epochs of ResNet18 (He et al., 2016) (1) without data augmentation, (2) horizontally flipping the image with 0.5 probability, and (3) a composition of randomly cropping a 32×32 portion of the image and random horizontal flip. Left: full CIFAR10 training data; Right half training data.

Outlook

- A general framework for understanding data augmentation is missing
- Without this, we need expensive and time-consuming experimentation to get it right
- Cannot see how to extend applicability beyond the classical ones (images and sound)

Overview

Our contributions

- We develop a general framework for data augmentation
- "averaging over group orbits"
- "reduces variance"
- Enables us to use data augmentation for new problems, where other approaches are traditionally used.

Setup

- Datapoints $X_1, \ldots, X_n$ (e.g., images)
- Consider a group G acting on the data space. (e.g., rotations)
- We assume that distribution of the data X is invariant under the group action G. For any $g \in G$:

$$X =_d gX.$$

similar results can be derived under approximate invariance – $\mathbb{E}W_d(X, gX)$ small

Setup - Example

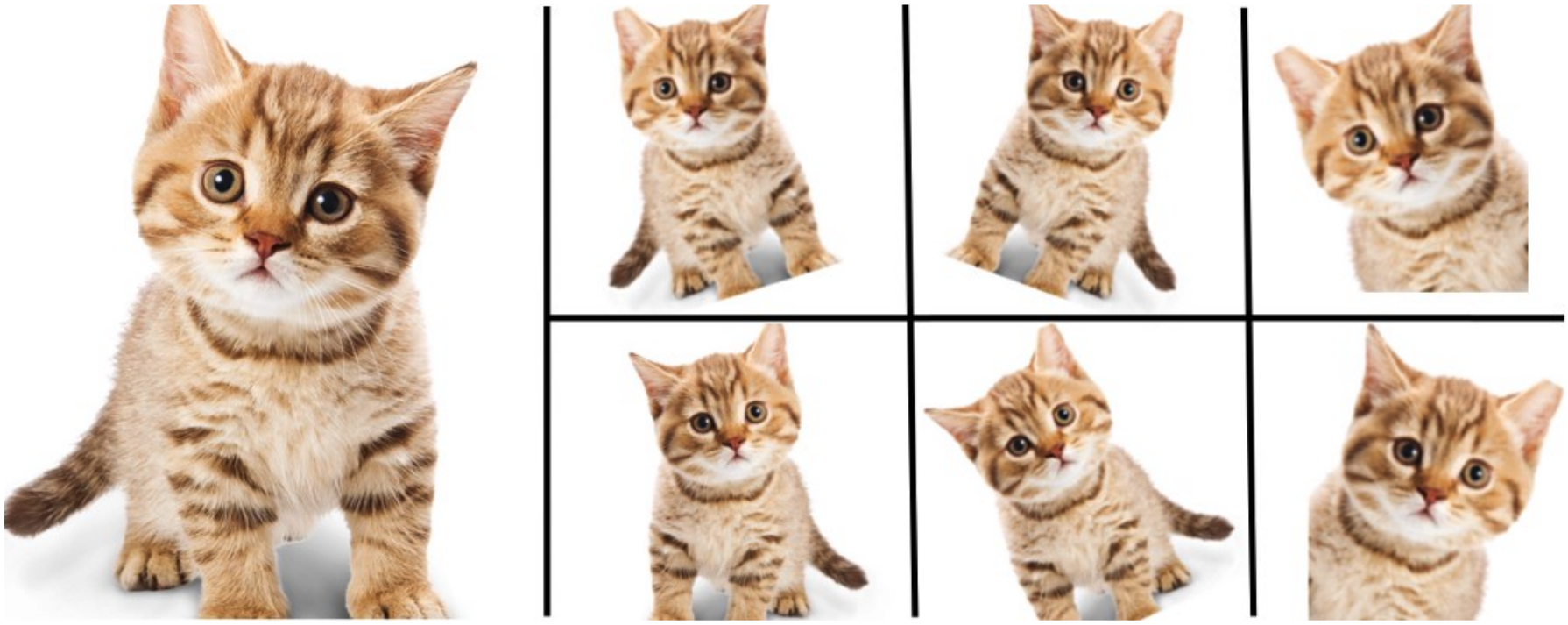


Figure: Example of $X \approx_d gX$.

How to do data augmentation?

- Intuitively, expand training set: train on all datapoints gX_i, $g \in G$
- If G is finite, this is actually doable (but expensive)
- In practice G is usually infinite
- How to handle that?

How is data augmentation done in practice?

- Given a loss function $L(\theta, X)$, minimize the empirical risk

$$\min_{\theta} \sum_{i=1}^{n} L(\theta, X_i).$$

- Iteratively over time $t = 1, 2, \ldots$ by stochastic gradient descent (SGD): apply random augmentation g_t to datapoint $X_{i,t}$. Update

$$\theta \leftarrow \theta - \eta_t \nabla L(\theta, g_t X_{i,t}).$$

- This is SGD on an *augmented* empirical risk, where we take an average over all augmentations:

$$\min_{\theta} \sum_{i=1}^{n} \mathbb{E}_G L(\theta, g X_i).$$

Variance reduction for MLE

- Let $X_1, \ldots, X_n \in \mathbb{R}^d \sim_{iid} P_\theta$ from a statistical model $\{P_\theta, \theta \in \Theta\}$
- $\hat{\theta}_n, \hat{\theta}_{n,G}$, approximate ERM, augmented ERM
- V_θ: Hessian of $\theta \mapsto L(\theta, x)$
- **Theorem**: Under regularity, we have

$$\sqrt{n}(\hat{\theta}_n - \theta) \Rightarrow \mathcal{N}(0, V_\theta^{-1}\mathbb{E}[\nabla L(\theta, \cdot)\nabla L(\theta, \cdot)^\mathsf{T}]V_\theta^{-1})$$

$$\sqrt{n}(\hat{\theta}_{n,G} - \theta) \Rightarrow \mathcal{N}\left(0, V_\theta^{-1} J_\theta V_\theta^{-1}\right),$$

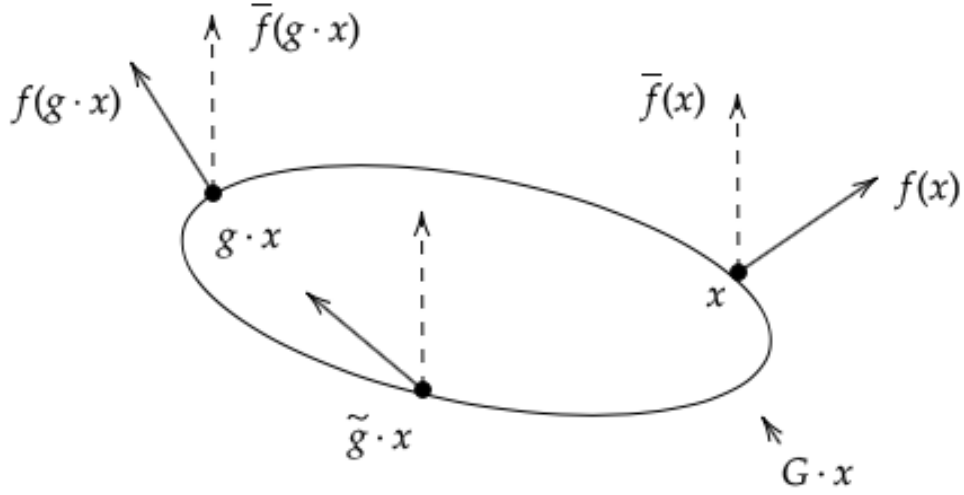
$$J_\theta = \mathbb{E}[\nabla L(\theta, \cdot)\nabla L(\theta, \cdot)^\mathsf{T}] - \mathbb{E}\mathrm{Cov}_G \nabla L(\theta, gX)$$

- The efficiency gain is governed by the covariance of the gradient

$$\mathrm{Cov}_G \nabla L(\theta, gX)$$

over G. "How much variance reduction we get by group averaging."

Intuition: Orbit Averaging



- Without averaging: Bahadur representation

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{\sqrt{n}} V_\theta^{-1} \sum_{i=1}^{n} \nabla L(\theta, X_i) + o_p(1).$$

- With averaging:

$$\sqrt{n}(\hat{\theta}_{n,G} - \theta) = \frac{1}{\sqrt{n}} V_\theta^{-1} \sum_{i=1}^{n} \nabla \mathbb{E}_G L(\theta, gX_i) + o_p(1).$$

Finite-sample results: Rademacher generalization bound

Proposition

Let $L(\theta, \cdot) \in [0,1]$ be Lipschitz w.r.t. some lower semi-continuous metric d. Let $\hat{\theta}$, $\hat{\theta}_G$ be ERM, augmented ERM, θ population risk minimizer. With probability at least $1-\delta$ over the draw of $X_1, \ldots, X_n$

$$\mathbb{E}L(\hat{\theta}, X) - \mathbb{E}L(\theta, X) \leq 2R_n(L \circ \Theta) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}}$$

$$\mathbb{E}L(\hat{\theta}_G, X) - \mathbb{E}L(\theta, X) \leq 2R_n(L_G \circ \Theta) + \sqrt{\frac{2\log\frac{2}{\delta}}{n}} + 2\|L\|_{\mathrm{Lip},d} \cdot \mathbb{E}_G W_d(X, gX),$$

where the Rademacher complexity $R_n(L_G \circ \Theta)$ of the augmented loss class can be bounded as

$$R_n(L_G \circ \Theta) \leq R_n(L \circ \Theta) + \|L\|_{\mathrm{Lip},d}\mathbb{E}_G W_d(X, gX).$$

In particular, if $X =_d gX$, then $R_n(L_G \circ \Theta) \leq R_n(L \circ \Theta)$, so augmentation decreases the Rademacher complexity.

Notes

- MLE: Can beat the fundamental Fisher information lower bound over non-invariant estimators.
- Also propose how to do data augmentation beyond ERM

Some Related Work

- Great deal of related work, see paper
- Inductive bias: Fukushima (1980), LeCun et al. (1989), ...
- Data augmentation methods: Baird (1992), Mirza and Osindero (2014); Antoniou et al. (2017); Hauberg et al. (2016); Ratner et al. (2017); Tran et al. (2017); Sixt et al. (2018); DeVries and Taylor (2017); Cubuk et al. (2018)
- Neural architecture design: Cohen and Welling (2016a); Cohen et al. (2018a,b); Dieleman et al. (2016); Worrall et al. (2017); Cohen and Welling (2016b); Esteves et al. (2018a,b, 2019)
- Invariant probabilistic models: Bloem-Reddy and Teh (2019)
- Other prominent works: Bengio et al. (2011); Rajput et al. (2019); Engstrom et al. (2017); Javadi et al. (2019); Liu et al. (2019); Hernández-García et al. (2018); Dao et al. (2019)

Example: Nonlinear Least Squares and 2L NN

- Nonlinear Least Squares

$$Y = f(\theta, X) + \varepsilon, \qquad \varepsilon \perp\!\!\!\perp X, \qquad \mathbb{E}\varepsilon = 0,$$

  where $\theta \in \mathbb{R}^p$.
- We have a group G acting on $\mathbb{R}^d \times \mathbb{R}$ *only through* X:

$$g(X, Y) = (gX, Y),$$

  and the invariance is characterized by

$$(gX, Y) =_d (X, Y).$$

Example: Nonlinear Least Squares and 2L NN

- The meaning of invariance is two-fold:
  1. the feature vector X is invariant: $X =_d gX$ for any $g \in G$;
  2. the mean label is invariant: $f(\theta, x) = f(\theta, gx)$ for any $x \in \mathbb{R}^d$ and any $g \in G$.
- Makes sense:
  - P(image) = P(rotated image)
  - P(cat|image)= P(cat|rotated image)

Example ctd

- ERM

$$\sqrt{n}(\hat{\theta}_{ERM} - \theta) \Rightarrow \mathcal{N}\left(0, \sigma^2 I_\theta^{-1}\right).$$

- Augmented ERM:

$$\sqrt{n}(\hat{\theta}_{aERM} - \theta) \Rightarrow \mathcal{N}(0, \Sigma_{aERM}),$$

with

$$\Sigma_{aERM} = \sigma^2 \cdot \left(I_\theta^{-1} - I_\theta^{-1}\mathbb{E}\left[\mathrm{Cov}_G \nabla f(\theta, gX)\right] I_\theta^{-1}\right).$$

2L NN

- Two-layer neural network

$$f_W(x) = 1^\mathsf{T}\sigma(Wx).$$

Theorem

1. *The Fisher information matrix* $I_W = \mathbb{E}\nabla f_W(X)\nabla f_W(X)^\mathsf{T}$ *can be viewed as a tensor*

$$I_W = \mathbb{E}(\sigma'(WX) \otimes \sigma'(WX)) \cdot (X \otimes X)^\mathsf{T}.$$

2. *If the activation function is quadratic,* $\sigma(x) = x^2/2$*, then the information is a product of the* $p^2 \times d^2$ *tensor* $W \otimes W$ *and the* $d^2 \times d^2$ *4th order moment tensor of* X*:*

$$I_W = (W \otimes W) \cdot \mathbb{E}(XX^\mathsf{T} \otimes XX^\mathsf{T}).$$

2L NN

Theorem

1. *Consider now augmentations acting by circular shift (translation-invariance). Let C_v be the circulant matrix with entries $C_v(i,j) = v_{i-j+1}$. Then the augmented Fisher information, $\bar{I}_W = \mathbb{E}[\mathrm{Cov}_G \nabla f(\theta, gX)]$ is*

$$\bar{I}_W = (W \otimes W) \cdot d^{-2}\mathbb{E}(C_X \otimes C_X) \cdot (C_X \otimes C_X)^{\mathsf{T}}.$$

2. *If the distribution of X is normal, $X \sim \mathcal{N}(0, I_d)$, then*

$$\bar{I}_W = (W \otimes W) \cdot F_2^* \cdot (F_2^2 \odot M) \cdot F_2^*.$$

Here $F_2 = F \otimes F$, F is the $d \times d$ DFT matrix, and M is the $d^2 \times d^2$ tensor with entries

$$M_{iji'j'} = F_i^{\mathsf{T}} F_j \cdot F_{i'}^{\mathsf{T}} F_{j'} + F_i^{\mathsf{T}} F_{j'} \cdot F_{i'}^{\mathsf{T}} F_j + F_i^{\mathsf{T}} F_{i'} \cdot F_i^{\mathsf{T}} F_{j'}.$$

2L NN

- Note: $\mathbb{E}\,\mathrm{tr}\, I_W = p\,\mathrm{tr}(XX^{\mathsf{T}})^2$, $\mathbb{E}\,\mathrm{tr}\, \bar{I}_W = p\,\mathrm{tr}(C_X C_X^{\mathsf{T}})^2/d^2$. Plot ratio, relative efficiency:

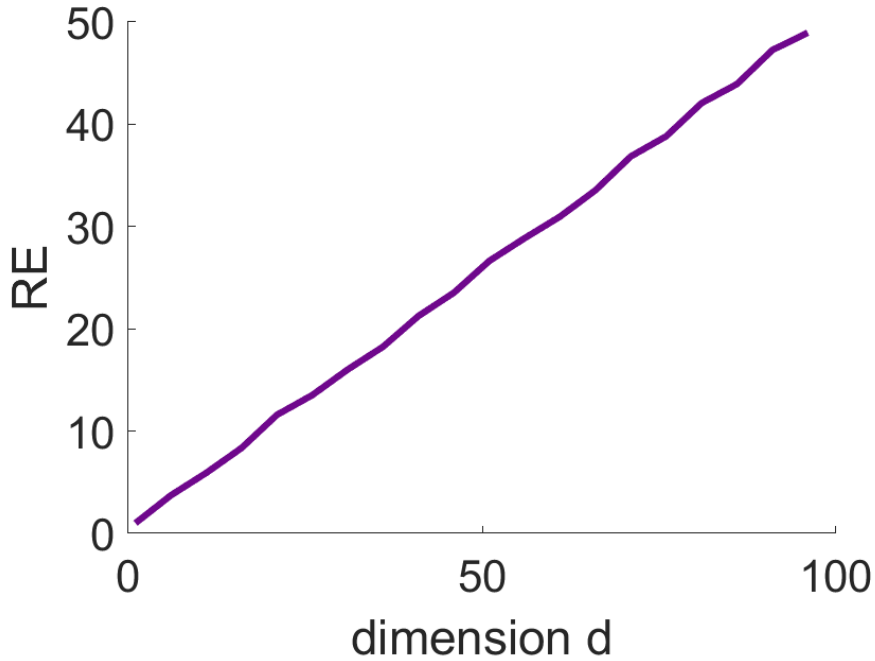


Figure: Plot of the increase in efficiency achieved by data augmentation in a *circular symmetry* model.

How to do data augmentation beyond ERM?

- General estimators $\hat{\theta}(x)$ that are not ERM:
  1. Regularized ERM, e.g., Ridge regression, Lasso
  2. Shrinkage estimators, e.g., Stein Shrinkage
  3. Nearest neighbors, e.g., k-NN regression (at least not in a natural way)
  4. Heuristic algorithms like Forward stepwise variable selection
- Augment: $\hat{\theta}_g(x) = \mathbb{E}_G \hat{\theta}(gx)$
- Augmentation decreases MSE of general estimators:

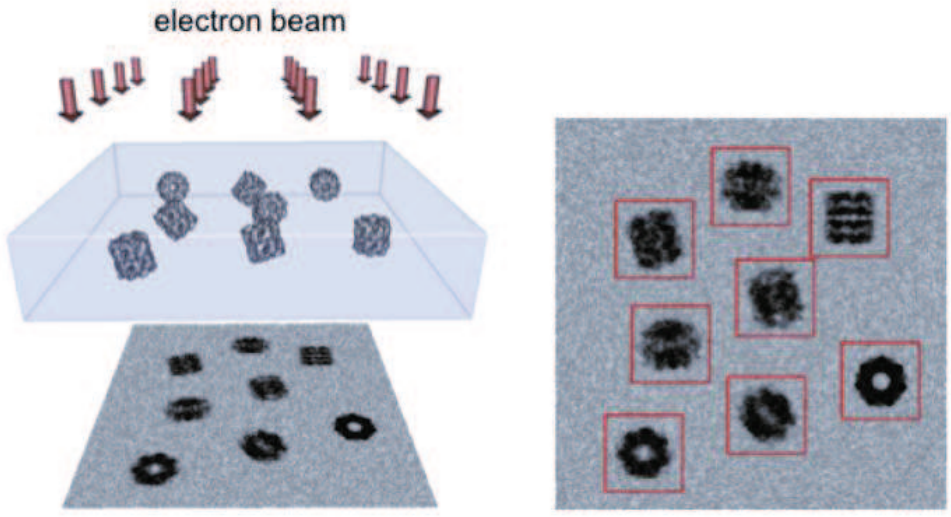
Proposition

Consider an estimator $\hat{\theta}(x)$ of θ, and its augmented version $\hat{\theta}_g(x) = \mathbb{E}_G \hat{\theta}(gx)$. The bias of the augmented estimator is the same as the bias of the original estimator, and the covariance matrix of the augmented estimator becomes smaller in the Loewner order. Hence, the MSE decreases by augmentation.

Potential new applications of data augmentation: cryo-EM

- Cryo-electron microscopy (cryo-EM) is a revolutionary technique in structural biology.
- Determine the structure of molecules to an unprecedented resolution.
- Nobel Prize in Chemistry in 2017.
- Data has invariance properties. Hard to exploit: massive, noisy.

Cryo-EM



Other new applications?

- X-ray Free Electron Lasers (XFEL)
- Spherically invariant data: marginal density estimation
- Unbalanced random effects models

Overview

Summary

- Big message: Using structure is key.
- Framework for data augmentation: averaging over group action
- Theory: variance reduction in finite sample and asymptotic regime.
- Examples: 2L NN
- Potential new applications: cryo-EM
- Many open problems:
  - Is this the right framework?
  - Optimization
  - New augmentation methods

Notes and Acknowledgements

- thanks to Zongming Ma for proposing the topic, Tony Cai for organizing DL reading group, many people for helpful suggestions
- preprint: `arxiv.org/abs/1907.10905`
- talk slides: `github.com/dobriban`
- funding: NSF BIGDATA IIS 1837992, NSF TRIPODS 1934960.
- deep learning course at Wharton Stats: `github.com/dobriban/Topics-in-deep-learning`

A. Antoniou, A. Storkey, and H. Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

H. S. Baird. Document image defect models. In *Structured Document Image Analysis*, pages 546–556. Springer, 1992.

Y. Bengio, F. Bastien, A. Bergeron, N. Boulanger-Lewandowski, T. Breuel, Y. Chherawala, M. Cisse, M. Côté, D. Erhan, J. Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172, 2011.

B. Bloem-Reddy and Y. W. Teh. Probabilistic symmetry and invariant neural networks. *arXiv preprint arXiv:1901.06082*, 2019.

T. Cohen and M. Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999, 2016a.

T. Cohen, M. Geiger, and M. Weiler. A general theory of equivariant cnns on homogeneous spaces. *arXiv preprint arXiv:1811.02017*, 2018a.

T. S. Cohen and M. Welling. Steerable cnns. *arXiv preprint arXiv:1612.08498*, 2016b.

T. S. Cohen, M. Geiger, J. Köhler, and M. Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018b.

E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

T. Dao, A. Gu, A. Ratner, V. Smith, C. De Sa, and C. Re. A kernel theory of

modern data augmentation. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

T. DeVries and G. W. Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017.

S. Dieleman, J. De Fauw, and K. Kavukcuoglu. Exploiting cyclic symmetry in convolutional neural networks. *arXiv preprint arXiv:1602.02660*, 2016.

L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

C. Esteves, C. Allen-Blanchette, A. Makadia, and K. Daniilidis. Learning so(3) equivariant representations with spherical cnns. In *The European Conference on Computer Vision (ECCV)*, September 2018a.

C. Esteves, A. Sud, Z. Luo, K. Daniilidis, and A. Makadia. Cross-domain 3d equivariant image embeddings. *arXiv preprint arXiv:1812.02716*, 2018b.

C. Esteves, Y. Xu, C. Allen-Blanchette, and K. Daniilidis. Equivariant multi-view networks. *arXiv preprint arXiv:1904.00993*, 2019.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4):193–202, 1980.

S. Hauberg, O. Freifeld, A. B. L. Larsen, J. Fisher, and L. Hansen. Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Artificial Intelligence and Statistics*, pages 342–350, 2016.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

A. Hernández-García, J. Mehrer, N. Kriegeskorte, P. König, and T. C. Kietzmann. Deep neural networks trained with heavier data augmentation learn features closer to representations in hit. In *Conference on Cognitive Computational Neuroscience*, 2018.

H. Javadi, R. Balestriero, and R. Baraniuk. A hessian based complexity measure for deep networks. *arXiv preprint arXiv:1905.11639*, 2019.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

S. Liu, D. Papailiopoulos, and D. Achlioptas. Bad global minima exist and sgd can reach them. *arXiv preprint arXiv:1906.02613*, 2019.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

S. Rajput, Z. Feng, Z. Charles, P.-L. Loh, and D. Papailiopoulos. Does data augmentation lead to positive margin? *arXiv preprint arXiv:1905.03177*, 2019.

A. J. Ratner, H. Ehrenberg, Z. Hussain, J. Dunnmon, and C. Ré. Learning to compose domain-specific transformations for data augmentation. In *Advances in neural information processing systems*, pages 3236–3246, 2017.

L. Sixt, B. Wild, and T. Landgraf. Rendergan: Generating realistic labeled data. *Frontiers in Robotics and AI*, 5:66, 2018.

T. Tran, T. Pham, G. Carneiro, L. Palmer, and I. Reid. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems*, pages 2797–2806, 2017.

D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. Harmonic networks: Deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5028–5037, 2017.

Overview

Why has deep learning been successful?

- Why? *Data* and *Compute*
- From AlexNet paper (won ImageNet 2012, dubbed "start of deep learning revolution")

  "It has only recently become possible to collect labeled datasets with millions of images"

  "current GPUs, paired with a highly-optimized implementation of 2D convolution, are powerful enough to facilitate the training of interestingly-large CNNs"

Data Augmentation in AlexNet

4 Reducing Overfitting

4.1 Data Augmentation

The easiest and most common method to reduce overfitting on image data is to artificially enlarge the dataset using label-preserving transformations (e.g., [25, 4, 5]). We employ two distinct forms of data augmentation, both of which allow transformed images to be produced from the original images with very little computation, so the transformed images do not need to be stored on disk.

The first form of data augmentation consists of generating image translations and horizontal reflections. We do this by extracting random 224×224 patches (and their horizontal reflections) from the 256×256 images and training our network on these extracted patches[4]. This increases the size of our training set by a factor of 2048, though the resulting training examples are, of course, highly interdependent.

Data Augmentation in AlexNet

The second form of data augmentation consists of altering the intensities of the RGB channels in training images. Specifically, we perform PCA on the set of RGB pixel values throughout the ImageNet training set. To each training image, we add multiples of the found principal components, with magnitudes proportional to the corresponding eigenvalues times a random variable drawn from a Gaussian with mean zero and standard deviation 0.1.

Variance reduction for MLE

- Let $X_1, \ldots, X_n \in \mathbb{R}^d \sim_{iid} P_\theta$ from a statistical model $\{P_\theta, \theta \in \Theta\}$
- Let $\ell(\theta, X) = \log p(X; \theta)$, where p is the density. Consider the MLE.

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \ell(\theta, X_i).$$

- Under regularity conditions, asymptotic normality holds

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta) \Rightarrow \mathcal{N}(0, I_\theta^{-1}),$$

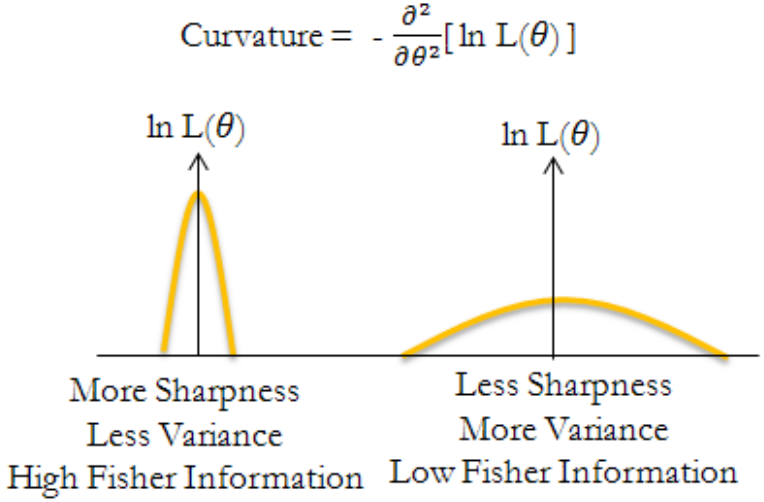
- The Fisher information I_θ is

$$I_\theta := \mathbb{E}\nabla\ell(\theta, X)\nabla\ell(\theta, X)^\mathsf{T} = -\mathbb{E}\nabla^2\ell(\theta, X)$$

the covariance of the score (gradient of the log-likelihood)

Some theory: background

- The Fisher information captures the local curvature of the model
- If the gradient varies a lot, then the information is large



- Fundamental lower bound: Among all "regular estimators", no estimator can have asymptotic variance less than the inverse Fisher information on a set of positive measure. (Fisher, Cramer-Rao, Le Cam, Hajek, ...)

Some theory

- Augmented MLE:

$$\hat{\theta}_{aMLE} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \mathbb{E}_G \ell(\theta, gX_i).$$

- Beats the fundamental Fisher information lower bound over non-invariant estimators.
- **Theorem**: We have

$$\sqrt{n}(\hat{\theta}_{aMLE} - \theta) \Rightarrow N(0, J_\theta^{-1})$$

where

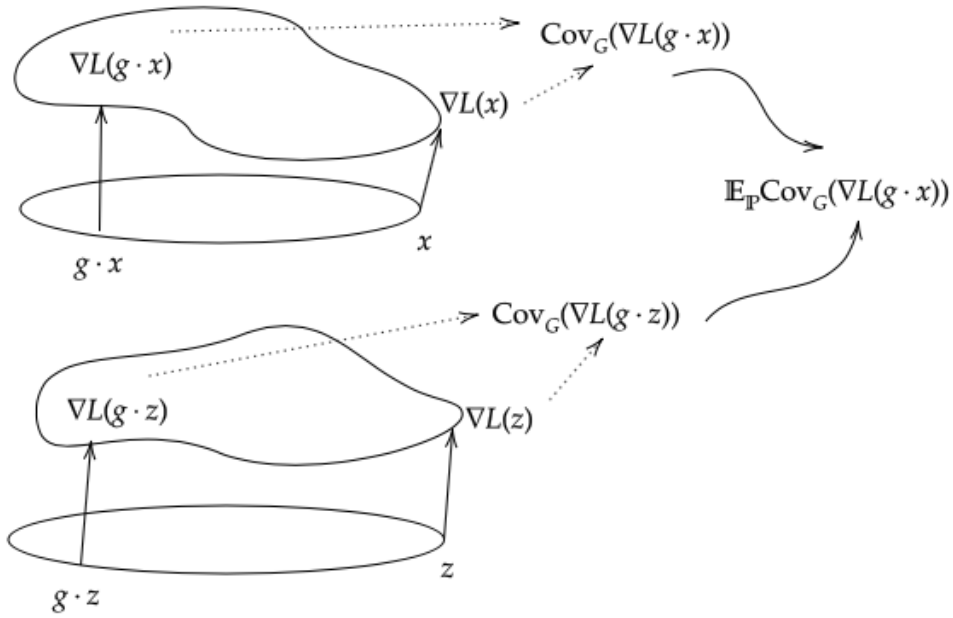
$$J_\theta^{-1} = I_\theta^{-1}(I_\theta - \mathbb{E}_\theta \mathrm{Cov}_G \nabla \ell(\theta, gX)) I_\theta^{-1} \leq I_\theta^{-1}.$$

- The efficiency gain is governed by the covariance of the gradient

$$\mathrm{Cov}_G \nabla \ell(\theta, gX)$$

over G. "How much variance reduction we get by averaging over the group."

Covariance of the gradient



- The average covariance of the gradient $\mathbb{E}_\theta \mathrm{Cov}_G \nabla \ell(\theta, gX)$ over G is large, if...?
- ... Averaging the score over the orbit reduces the variance

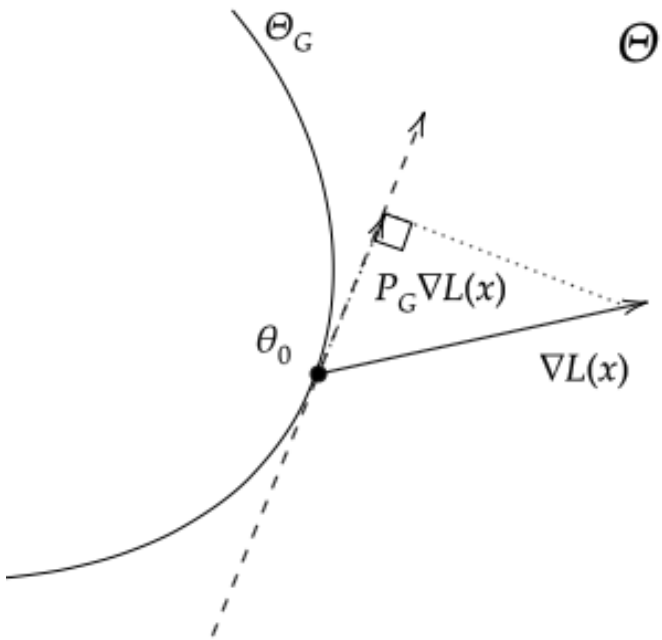
$$\mathbb{E}_\theta \mathrm{Cov}_G \nabla \ell(\theta, gX) = \mathrm{Cov}_\theta \nabla \ell(\theta, X) - \mathrm{Cov}_\theta \mathbb{E}_G \nabla \ell(\theta, gX)$$

How much do we gain?

- *Invariant subspace*

$$\Theta_G = \{\theta \in \Theta : gX =_d X \ \forall g \in G, \text{where } X \sim P_\theta\}.$$

- In general cERM may be hard to compute.
- Projection into the tangent space is invariant; so only the orthogonal part reduces variance $\mathbb{E}_\theta \mathrm{Cov}_G(\nabla \ell(\theta, gX)) = \mathbb{E}_\theta \mathrm{Cov}_G(P_G^\perp \nabla \ell(\theta, gX))$.



- Big subspace $\rightarrow$ large gains.