A Kernel Theory of Modern Data Augmentation

Tri Dao, Albert Gu, Alex Ratner, Virginia Smith, Chris De Sa, Chris Ré
Data augmentation example
Data augmentation is important to accuracy...

3.7 pt. average gain across top ten CIFAR-10 models
13.9 pt. average gain for CIFAR-100

A form of weak supervision: expresses domain knowledge (invariance)
… but is not well understood

**How does data augmentation affect the model?**

- Learning process
- Parameters and decision surface
Goal: Understand effects of data augmentation

Invariance

Regularization

Practical utility

speeding up training

as a diagnostic
Kernel with feature map

\[ k(x, y) = \langle \phi(x), \phi(y) \rangle \]
Model of data augmentation

Non-augmented: \[
\min_w \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top \phi(x_i))
\]

Augmented: \[
\min_w \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim T(x_i)} \ell(w^\top \phi(z_i))
\]

- Loss function
- Feature map
- Transformed versions of data point
Data augmentation: 1\textsuperscript{st} order approximation

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim T(x_i)} \ell(w^\top \phi(z_i)) \approx \frac{1}{n} \sum_{i=1}^{n} \ell(w^\top \mathbb{E}_{z_i \sim T(x_i)} \phi(z_i))
\]

Data augmentation induces invariance by \textbf{feature averaging}
Data augmentation: 2\textsuperscript{nd} order approximation

\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{z_i \sim T(x_i)} \ell(w^\top \phi(z_i)) \approx \]

\[
\frac{1}{n} \sum_{i=1}^{n} \ell(w^\top \phi(x_i)) + w^\top \left( \frac{1}{2n} \sum_{i=1}^{n} \text{Cov}_{z_i \sim T(x_i)} [\phi(z_i)] \ell''(w^\top \phi(x_i)) \right) w
\]

Average of augmented features

Data augmentation reduces model complexity via a \textbf{data-dependent} regularization term
How accurate are the approximations?

Relative objective difference (MNIST)

Relative prediction disagreement (MNIST)
A diagnostic: kernel alignment metric

Averaged features: \[ \psi(x) = \mathbb{E}_{z \sim T(x)} \phi(z) \]

Kernel target alignment (Cristianini et al., NIPS 2002):

how well separated are features from different classes
A diagnostic: kernel alignment metric

Kernel alignment correlates with accuracy.
Efficient augmentation: random features

Random Fourier features:

\[
\tilde{\phi}(x) = \frac{1}{\sqrt{m}} \left[ \exp(i\omega_1^\top x) \ldots \exp(i\omega_m^\top x) \right]
\]

For linear transforms: \(\exp(i\omega^\top Ax) = \exp(i(A^\top \omega)^\top x)\)

Augmented random Fourier features (Raj et. al., AISTATS 2017):

\[
\tilde{\psi}(x) = \frac{1}{s\sqrt{m}} \left[ \sum_{j=1}^{s} \exp(i(A_{\alpha_j} \omega_1)^\top x) \ldots \sum_{j=1}^{s} \exp(i(A_{\alpha_j} \omega_m)^\top x) \right]
\]
Efficient augmentation: random features

<table>
<thead>
<tr>
<th>Model</th>
<th>MNIST</th>
<th>CIFAR-10</th>
<th>DDSM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Acc. (%)</td>
<td>Time</td>
<td>Acc. (%)</td>
</tr>
<tr>
<td>No augmentation</td>
<td>96.1 ± 0.1</td>
<td>34s</td>
<td>39.4 ± 0.5</td>
</tr>
<tr>
<td>Traditional augmentation</td>
<td>97.6 ± 0.2</td>
<td>220s</td>
<td>45.3 ± 0.5</td>
</tr>
<tr>
<td>Augmented RFFs</td>
<td>97.6 ± 0.1</td>
<td>54s</td>
<td>45.2 ± 0.4</td>
</tr>
</tbody>
</table>

Augmented random features retains 70-100% accuracy boost, with 2x-4x faster training.
Accuracy vs computation

Apply cheap 1\textsuperscript{st} order approximation at earlier layers of network.
Augmentation as sequence modeling

- TANDA, Ratner et al., NIPS 2017
- AutoAugment, Cubuk et al., 2018

Model augmentation as a Markov chain
Augmentation as kernels

Base classifier: k-nearest neighbors
+ Data augmentation
= Asymptotic kernel classifier
Summary

• Data augmentation + k-NN = asymptotic kernel classifier.

• Data augmentation induces invariance and regularizes

• Application in speeding up training and diagnostics