Data augmentation as a biologically plausible alternative to explicit regularization in CNNs

IIT - NCSR “Demokritos”

Aθήνα, 8th of October 2018
Alex Hernández-García
Main conclusions

SPOILER ALERT
Main conclusions

In convolutional neural networks trained with sufficient level of data augmentation, explicit regularizers (weight decay and dropout) might not provide any additional generalization improvement.

Data augmentation presents much higher adaptability to changes in the architecture and the amount of training data than weight decay and dropout, which require specific fine-tuning of their hyper-parameters.

Models trained with heavier data augmentation exhibit higher representational similarity to the human inferior temporal (IT) cortex.
Motivation

Statistical learning theory and DL

According to statistical learning theory, too high a model complexity yields overfitting and poor generalization.

However:

<table>
<thead>
<tr>
<th></th>
<th>CIFAR-10</th>
<th>ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td># train</td>
<td>50,000</td>
<td>1,200,000</td>
</tr>
<tr>
<td>Inception</td>
<td>1,649,402</td>
<td></td>
</tr>
<tr>
<td>Inception V4</td>
<td>42,681,353</td>
<td></td>
</tr>
<tr>
<td>Resnet-{18,152}</td>
<td>11,689,512,60,192,808</td>
<td></td>
</tr>
<tr>
<td>Alexnet</td>
<td>1,387,786</td>
<td></td>
</tr>
<tr>
<td>Alexnet</td>
<td>61,100,840</td>
<td></td>
</tr>
<tr>
<td>MLP 1x512</td>
<td>1,209,866</td>
<td></td>
</tr>
<tr>
<td>VGG-{11:19}</td>
<td>132,863,336,143,667,240</td>
<td></td>
</tr>
</tbody>
</table>

Zhang et al. (2016). Understanding deep learning requires rethinking generalization. ICLR.

Convolutional networks do generalize, despite its comparably huge amount of parameters.
Motivation

Statistical learning theory and DL

According to statistical learning theory, too high a model complexity yields overfitting and poor generalization.

Zhang et al. (2016). Understanding deep learning requires rethinking generalization. ICLR.

Zhang et al. claimed that DL operates in an overparameterization regime where theory might not hold
Motivation

“Rethinking generalization”?

Some observations from Zhang et al., 2016:

- “Explicit regularization may improve generalization performance, but is neither necessary nor by itself sufficient for controlling generalization error.”
- “While regularization is important, bigger gains can be achieved by simply changing the model architecture.”
- “It seems that data augmentation is more powerful than just tuning weight decay or preventing low training error.”

A global answer is however still open.
We propose to rethink regularization and data augmentation instead.
Rethinking regularization

The common practice

Most CNNs trained for image object recognition have a much larger number of parameters than training images. How is this fought in practice?

- **REGULARIZATION**
  - The large capacity of CNNs is usually constrained by regularization techniques. Isn’t this a waste?
- **EARLY STOPPING**
  - Srivastava et al. (2014)
- **DROPOUT**
  - Srivastava et al. (2014)
- **WEIGHT DECAY**
  - Hanson & Pratt (1989)
- **L1 PENALTIES**
- **TIKHONOV**
  - Tikhonov (1943)
- **STOCHASTIC DEPTH**
  - Huang et al. (2016)

[Graph showing training and testing error over model complexity]

Hanson & Pratt (1989)
Srivastava et al. (2014)
Huang et al. (2016)
Tikhonov (1943)
An important distinction

There exist many regularization techniques but the literature often overlooks an important distinction: explicit vs. implicit regularization.

The terms have been used before (Neyshabur et al., 2014; Zhang et al., 2017) without a formal definition. Therefore, we propose (Hernández-García & König, 2018a):

- **Explicit regularization** techniques are those specifically and solely designed to constrain the effective capacity of a given model in order to reduce overfitting. Furthermore, explicit regularizers are not a structural or essential part of the network architecture, the data or the learning algorithm and can typically be added or removed easily. Ex: weight decay and dropout

- **Implicit regularization** is the reduction of the generalization error provided by characteristics of the architecture, the training data or the learning algorithm, which are not specifically designed to constrain the effective capacity of the model. Ex: SGD and data augmentation

Neyshabur et al. (2014). In search of the real inductive bias: On the role of implicit regularization in deep learning. ICLR
Zhang et al. (2016). Understanding deep learning requires rethinking generalization. ICLR.
There exist many regularization techniques but the literature often overlooks an important distinction: explicit vs. implicit regularization.

The terms have been used before (Neyshabur et al., 2014; Zhang et al., 2017) without a formal definition. Therefore, we propose (Hernández-García & König, 2018a):

- **Explicit regularization** techniques are those specifically and solely designed to constrain the effective capacity of a given model in order to reduce overfitting. Furthermore, explicit regularizers are not a structural or essential part of the network architecture, the data or the learning algorithm and can typically be added or removed easily. Ex: weight decay and dropout

- **Implicit regularization** is the reduction of the generalization error provided by characteristics of the architecture, the training data or the learning algorithm, which are not specifically designed to constrain the effective capacity of the model. Ex: SGD and data augmentation

Many features of CNNs implicitly regularize the model and play a crucial role in generalization
Data augmentation
A new view of data augmentation

Data augmentation has been often considered a form of explicit regularization (Zhang et al., 2016) and sort of a hack to train CNNs.

Definition: data augmentation in machine learning refers to the techniques that synthetically expand a data set by applying (random, but constrained) transformations on the existing examples, thus augmenting the amount of available training data.

We claim that data augmentation...

● does not reduce the effective capacity of the model
● increases the number of training examples
● improves the robustness against input variability
● and as a desirable side effect, it implicitly regularizes the model and improves generalization

Data augmentation is not explicit regularization!
Methods

Analysis of data augmentation on CNNs

In spite of being widely used, the literature lacks a systematic analysis of the impact of data augmentation on CNNs, compared to explicit regularization.

An ablation study of data augmentation and weight decay and dropout:

<table>
<thead>
<tr>
<th>REGULARIZATION</th>
<th>DATA AUGMENTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>no expl. regularization</td>
<td>no data aug.</td>
</tr>
<tr>
<td>weight decay &amp; dropout</td>
<td>no data aug. &amp; no expl. regularization</td>
</tr>
<tr>
<td></td>
<td>no data aug. &amp; expl. regularization</td>
</tr>
</tbody>
</table>

Our study aims at revealing the interaction between explicit regularization and data augmentation.
We are interested in studying the effect of data augmentation on training CNNs.

Data augmentation schemes

Data sets CIFAR10, CIFAR100 and ImageNet

3 schemes: no augmentation, light transformations as in the literature and heavier transformations
Methods

Data augmentation schemes

\[
\begin{bmatrix}
x' \\
y' \\
1
\end{bmatrix} = \begin{bmatrix}
f_h z_x \cos(\theta) & -z_y \sin(\theta + \phi) & t_x \\
z_x \sin(\theta) & z_y \cos(\theta + \phi) & t_y \\
0 & 0 & 1
\end{bmatrix} \begin{bmatrix}
x \\
y \\
1
\end{bmatrix}
\]

\[x' = x + \delta\]

\[x' = \gamma(x - \bar{x}) + \bar{x}\]

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>(f_h)</td>
<td>Horizontal flip</td>
<td>(1 - 2B(0.5))</td>
</tr>
<tr>
<td>(t_x)</td>
<td>Horizontal translation</td>
<td>(\mathcal{U}(-0.1, 0.1))</td>
</tr>
<tr>
<td>(t_y)</td>
<td>Vertical translation</td>
<td>(\mathcal{U}(-0.1, 0.1))</td>
</tr>
<tr>
<td>(z_x)</td>
<td>Horizontal scale</td>
<td>(\mathcal{U}(0.85, 1.15))</td>
</tr>
<tr>
<td>(z_y)</td>
<td>Vertical scale</td>
<td>(\mathcal{U}(0.85, 1.15))</td>
</tr>
<tr>
<td>(\theta)</td>
<td>Rotation angle</td>
<td>(\mathcal{U}(\frac{-\pi}{180}, 22.5), \frac{\pi}{180}, 22.5)</td>
</tr>
<tr>
<td>(\phi)</td>
<td>Shear angle</td>
<td>(\mathcal{U}(\frac{-\pi}{180}, 0.15))</td>
</tr>
<tr>
<td>(\gamma)</td>
<td>Contrast</td>
<td>(\mathcal{U}(0.5, 1.5))</td>
</tr>
<tr>
<td>(\delta)</td>
<td>Brightness</td>
<td>(\mathcal{U}(\frac{-\pi}{180}, 0.25))</td>
</tr>
</tbody>
</table>
Methods

Network architectures

- **All-CNN** (Springerberg et al., 2014):
  - Only convolutional layers.
  - Shallow, only 12 layers with a total of 1.3 million parameters.

- **WRN** (Zagoruyko & Komodakis, 2016):
  - A version of ResNet with fewer layers, but more units per layer.
  - Deeper, 28 layers with a total of 36.5 million parameters.

Both networks are trained with the original training hyperparameters (learning rate, iterations, etc.)

Springerberg et al. (2014). Striving for simplicity: The all convolutional net. ICLR
Zagoruyko & Komodakis (2016). Wide residual networks. BMVC.
Methods

Network architectures

- **All-CNN** (Springerberg et al., 2014):
  - Only convolutional layers.
  - Shallow, only 12 layers with a total of 1.3 million parameters.

- **WRN** (Zagoruyko & Komodakis, 2016):
  - A version of ResNet with fewer layers, but more units per layer.
  - Deeper, 28 layers with a total of 36.5 million parameters.

Both networks are trained with the original training hyperparameters (learning rate, iterations, etc.)

All-CNN and WRN are two distinct, well known CNNs, architecturally simple, but successful neural networks, suitable as a testing ground.
Results

Original architectures

The original architectures trained with/without weight decay and dropout and different levels of data augmentation.

Data augmentation alone achieves very close or even better accuracy than explicitly regularized models.
With the original setup, data augmentation alone already obtains similar results to the explicitly regularized models.

How do weight decay, dropout and data augmentation adapt to changes in the original setup?

Shallower and deeper architectures

- **All-CNN**
  - (original: 12 layers, 1.3 M param.)
  - **Shallower**: 9 layers, 374 K param.
  - **Deeper**: 15 layers, 2.4 M param.

Reduced training sets

- 50 %
- 10 %

of the original number of examples

With the original setup, data augmentation alone already obtains similar results to the explicitly regularized models.

How do weight decay, dropout and data augmentation adapt to changes in the original setup?

Shallower and deeper architectures

All-CNN
(original: 12 layers, 1.3 M param.)

Shallower
9 layers, 374 K param.

Deeper
15 layers, 2.4 M param.

Reduced training sets

- 50 %
- 10 %

of the original number of examples

We vary the depth of the architectures and the amount of available training examples.
Further advantages of data augmentation

Shallower and deeper architectures

When trained with data augmentation alone, the deeper models perform slightly better and the shallower models slightly worse. With explicit regularization, both architectures present a dramatic drop.
Further advantages of data augmentation

Reduced training sets
Further advantages of data augmentation

Reduced training sets

CIFAR100
Further advantages of data augmentation

When fewer training examples are available, the benefits of training with data augmentation and without explicit regularization become even larger.
Why does data augmentation adapt much better to changes in the architecture and the amount of training data, compared to weight decay and dropout?

- Weight decay and dropout are explicit regularization methods, that is
- they directly influence (reduce) the effective capacity of the models, whereas
- data augmentation does not reduce the effective capacity of the model and
- only implicitly regularizes it by increasing the amount of training examples.
- Weight decay and dropout require specific fine-tuning of their hyperparameters. On the contrary,
- data augmentation only depends on the type of data (e.g. natural images)

Data augmentation is a much more adaptable technique for training CNNs than explicit regularization
Discussion (I)

Rethinking data augmentation

Data augmentation is often regarded as a *hack* to train convolutional networks, while weight decay are almost ubiquitously used.

We propose to switch the roles:

- Stop using explicit regularization because
  - it has been shown unnecessary
  - it hampers the development process

- Incorporate data augmentation into the training procedures
  - encourage better augmentation techniques
  - better models should better exploit the data

Data augmentation seems to be a more advantageous alternative to explicit regularization
Limitations

Although the results of our experiments present a reasonable level of consistency, we cannot guarantee that the same conclusions would be observed

- on every other CNN architecture
- with different types of image transformations
- on other data domains (e.g. speech, text, music...)

We encourage further research to confirm the validity of these hypotheses at a broader scale
Main conclusions

In convolutional neural networks trained with sufficient level of data augmentation, explicit regularizers (weight decay and dropout) might not provide any additional generalization improvement.

Data augmentation presents much higher adaptability to changes in the architecture and the amount of training data than weight decay and dropout, which require specific fine-tuning of their hyper-parameters.

Models trained with heavier data augmentation exhibit higher representational similarity to the human inferior temporal (IT) cortex.
Deep neural networks trained with heavier data augmentation learn features closer to representations in hIT

Alex Hernández-García, Johannes Mehrer, Nikolaus Kriegeskorte, Peter König and Tim C. Kietzmann
(Cognitive Computational Neuroscience 2018)
The features learnt by some convolutional neural networks (CNN) explain inferior temporal (IT) cortex representations (Yamins et al., 2013; Khaligh-Razavi & Kriegeskorte, 2014).

Data augmentation helps convolutional neural networks achieve better generalization and can be a substitute of explicit regularization (Hernández-García & König, 2018a).

Hypothesis: does training with data augmentation facilitate learning representations better aligned with the human IT cortex?

A bigger, open question: what makes CNNs learn representations that more closely mirror the ones in IT?

The features learnt by some convolutional neural networks (CNN) explain inferior temporal (IT) cortex representations (Yamins et al., 2013; Khaligh-Razavi & Kriegeskorte, 2014).

Data augmentation helps convolutional neural networks achieve better generalization and can be a substitute of explicit regularization (Hernández-García & König, 2018a).

Hypothesis: *does training with data augmentation facilitate learning representations better aligned with the human IT cortex?*

A bigger, open question: *what makes CNNs learn representations that more closely mirror the ones in IT?*

We aim at exploring the biological plausibility of training CNNs with data augmentation.
Object recognition in the brain

The inferior temporal (IT) cortex is the highest-level area for object recognition in the brain.

Methods

**CNNs**

- **Networks:** All-CNN and WRN
- **Dataset:** ImageNet ILSVRC 2012
- **(1.3 M images, 1,000 classes)**
- **Data augmentation:** light vs. heavier

We want to compare CNNs trained with data augmentation with the human IT cortex.

We compare the representational similarity of CNNs trained with light and heavier data augmentation.
We want to compare CNNs trained with data augmentation with the human IT cortex.

**Methods**

**CNNs**

- **Networks:** All-CNN and WRN
- **Dataset:** ImageNet ILSVRC 2012 (1.3 M images, 1,000 classes)
- **Data augmentation:** light vs. heavier

**Diagram:**

- **All-CNN**
  - Top-5 accuracy
  - Accuracy

- **WRN**
  - Top-5 accuracy
  - Accuracy

The diagram shows that All-CNN and WRN perform reasonably well on ImageNet. Heavier augmentation yields worse accuracy.

All-CNN and WRN perform reasonably well on ImageNet. Heavier augmentation yields worse accuracy.
Methods

Brain

- Measurements: BOLD responses from 3T fMRI by 15 participants
- Stimuli: 92 images of isolated objects, known to elicit different representations in IT.

We want to compare CNNs trained with data augmentation with the human IT cortex.
Methods

Representational similarity analysis (RSA)

Definition: representational similarity analysis (RSA) characterizes a representation in a brain or a computational model by the dissimilarities in the response patterns elicited by a set of stimuli.

- It allows direct comparisons between model systems without having to explicitly align the different measurement types.
- In practice, a representational dissimilarity matrix (RDM) is built for each model, storing the pairwise distances between stimuli responses.
- Then, the correlation between the RDMs of different systems characterize the similarity between their representations.

We want to compare CNNs trained with data augmentation with the human IT cortex.

RSA provides a framework to compare the representations in the brain and CNNs.
Methods

Representational similarity analysis (RSA)

We want to compare CNNs trained with data augmentation with the human IT cortex.

We compute an RDM for every system we want to compare (CNN, IT cortex).
Results
Kendall correlation between IT and CNN RDMs

The models trained with heavier data augmentation correlate significantly higher with IT fMRI representations, than their counterparts trained with light transformations.

This is true in spite of the worse classification performance achieved by the models trained with heavier transformations.

Why do models train with heavier data augmentation better resemble the representations in the brain?

- CNNs trained with heavier data augmentation exhibit higher representational similarity with the human inferior temporal (IT) cortex.
- Humans develop robust object recognition by constantly being exposed to visual stimuli.
- Data augmentation might better mimic the way the visual cortex develops object recognition.
- However, the present study is only preliminary and future work should confirm this hypothesis and further develop better models of the primate visual system.
Summary

1. Motivation: propose to *rethink regularization* and *data augmentation* instead of *rethinking generalization* (Zhang et al., 2016)

2. New definitions of *explicit* and *implicit* regularization

3. New view of data augmentation as implicit (and not explicit) regularization

4. Experiments: data augmentation alone outperforms weight decay and dropout on the original setup and presents much higher adaptability to changes in the architecture and the training data.

5. Biological plausibility: comparison of CNNs with the IT cortex

6. Models trained with heavier data augmentation might be more similar to the human brain
Thank you very much!

Alex Hernández-García
ahernandez@uos.de

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 641805.