Usable Information and Evolution of Optimal Representations During Training

Michael Kleinman

22/02/2022
Overview

- How is relevant and irrelevant information about the input $X$ represented during training?
- How can we quantify the information contained in a representation $Z$ in a deep network?
- How are the learning dynamics affected by the implicit regularization coming from SGD?
Possible learning dynamics

Hypothesis 1

$I(Z; X_{rel})$
Possible learning dynamics

Hypothesis 1

\[ I(Z; X_{\text{rel}}) \]

\[ I(Z; X_{\text{irrel}}) \]

Training Epoch
Possible learning dynamics

Hypothesis 1

Hypothesis 2
Possible learning dynamics

Hypothesis 1

Hypothesis 2

Hypothesis 3
Possible learning dynamics

• Prior work using **Shannon’s mutual information** suggested these learning dynamics (Shwartz-Ziv and Tishby, 2017) but has been disputed in part over the **approximation** of mutual information (Saxe et al., 2018).
“Usable Information” in a representation

• A representation Z may store information in a variety of ways.

• It may be that a complex transformation is required to read out the information, or it may be that a simple linear decoder could read out the information.

• In both cases, from an information-theoretic perspective, the same information is contained in the representation, however, there is an important distinction regarding how “usable” this information is.
Usable Information (definition)

\[ I_u(Z; Y) := H(Y) - L_{CE}(p(y|z), q(y|z)) \]

- \( H(Y) \) is the entropy, or uncertainty, of \( Y \)
- \( L_{ce} \) is the cross-entropy loss on the test set of
- \( q(y|z) \) is a discriminator network trained to approximate the true distribution \( p(y|z) \)
- Related to V-Information (Xu et al., 2020)

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
Usable Information (definition)

\[ I_u(Z; Y) := H(Y) - L_{CE}(p(y|z), q(y|z)) \]

Property: \( I_u(Z; Y) \leq I(Z; Y) \)
Results: CIFAR-10

airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

Fine Labels

https://www.cs.toronto.edu/~kriz/cifar.html

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
Results: CIFAR-10

https://www.cs.toronto.edu/~kriz/cifar.html

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
Results: CIFAR-10

<table>
<thead>
<tr>
<th>Coarse Label 1</th>
<th>Coarse Label 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>airplane</td>
<td></td>
</tr>
<tr>
<td>automobile</td>
<td></td>
</tr>
<tr>
<td>bird</td>
<td></td>
</tr>
<tr>
<td>cat</td>
<td></td>
</tr>
<tr>
<td>deer</td>
<td></td>
</tr>
<tr>
<td>dog</td>
<td></td>
</tr>
<tr>
<td>frog</td>
<td></td>
</tr>
<tr>
<td>horse</td>
<td></td>
</tr>
<tr>
<td>ship</td>
<td></td>
</tr>
<tr>
<td>truck</td>
<td></td>
</tr>
</tbody>
</table>

Task: Output coarse label

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
Task: Output coarse label
Results: CIFAR-10

Task: Output coarse label

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
Effect of learning rate and batch size

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
Effect of learning rate and batch size

Increasing batch size

Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021
**Effect of learning rate and batch size**

**Decreasing learning rate**

**Increasing batch size**

![Graphs showing the effect of learning rate and batch size on usable information and validation accuracy during training.](image)

*Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021*
Conclusion

• We introduce a notion of *usable information* contained in the representation learned by a deep network, and use it to study how optimal representations for the task emerge during training.

• We show that the implicit regularization coming from training with Stochastic Gradient Descent with a high learning rate and small batch size plays an important role in learning minimal sufficient representations for the task.

*Kleinman, Achille, Idnani, Kao. Usable Information and Evolution of Optimal Representations During Training. ICLR 2021*