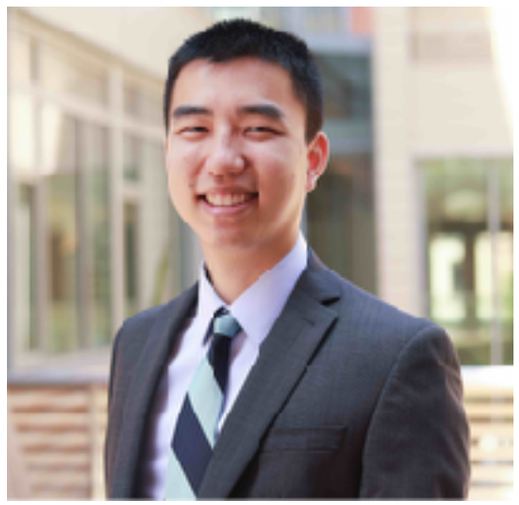
Usable Information and Evolution of Optimal Representations During Training

Michael Kleinman
UCLA

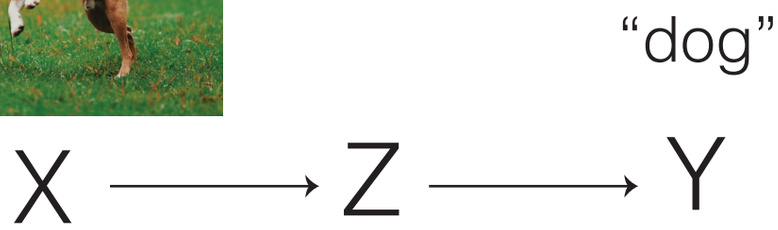
Alessandro Achille
Caltech

Daksh Idnani
UCLA



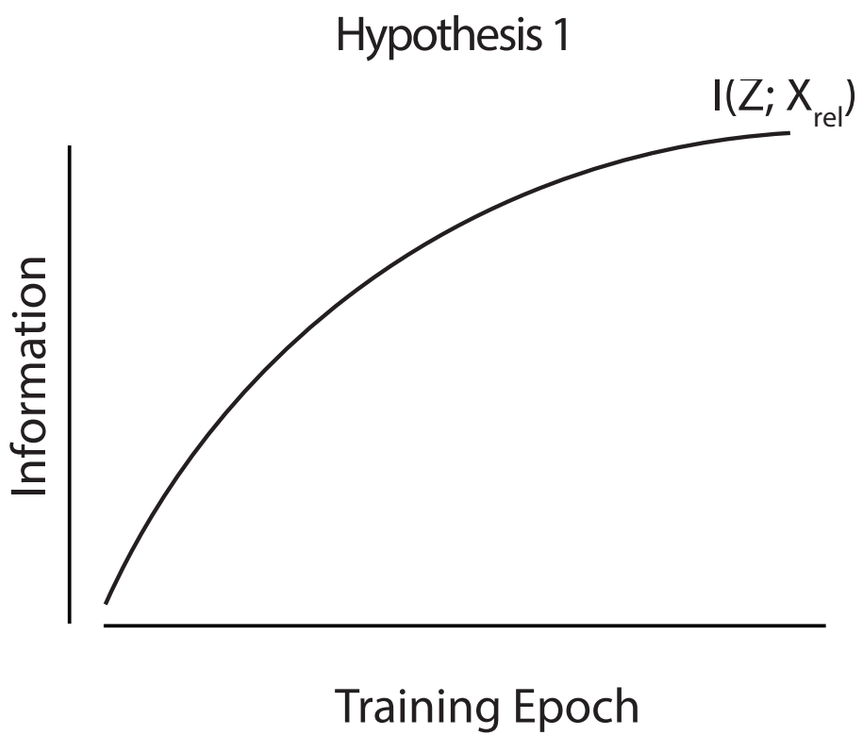
Jonathan Kao
UCLA

Overview

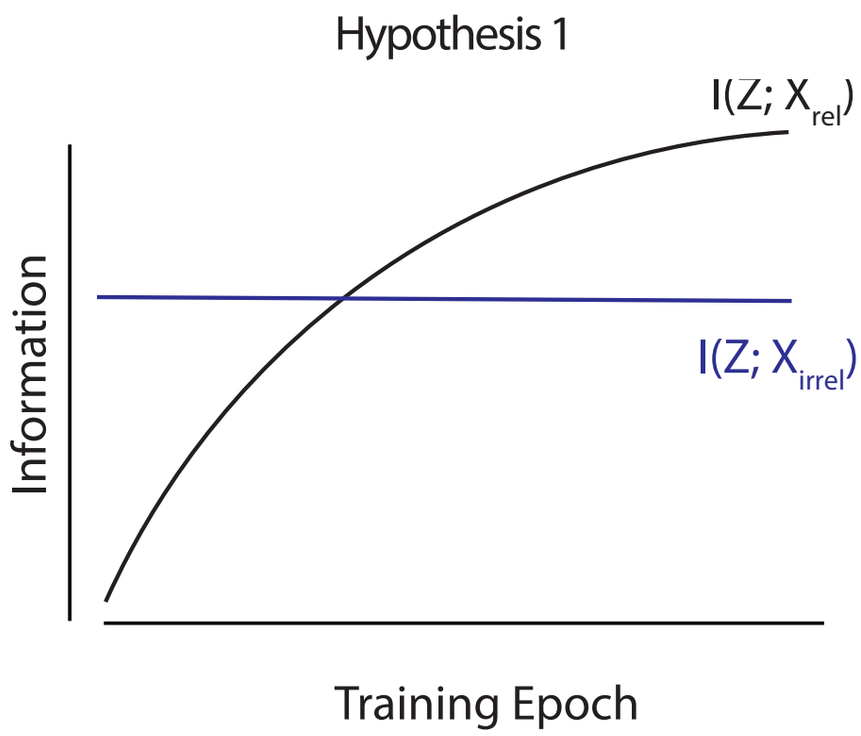


- How is relevant and irrelevant information about the input X represented during training?
- How can we quantify the information contained in a representation Z in a deep network?
- How are the learning dynamics affected by the implicit regularization coming from SGD?

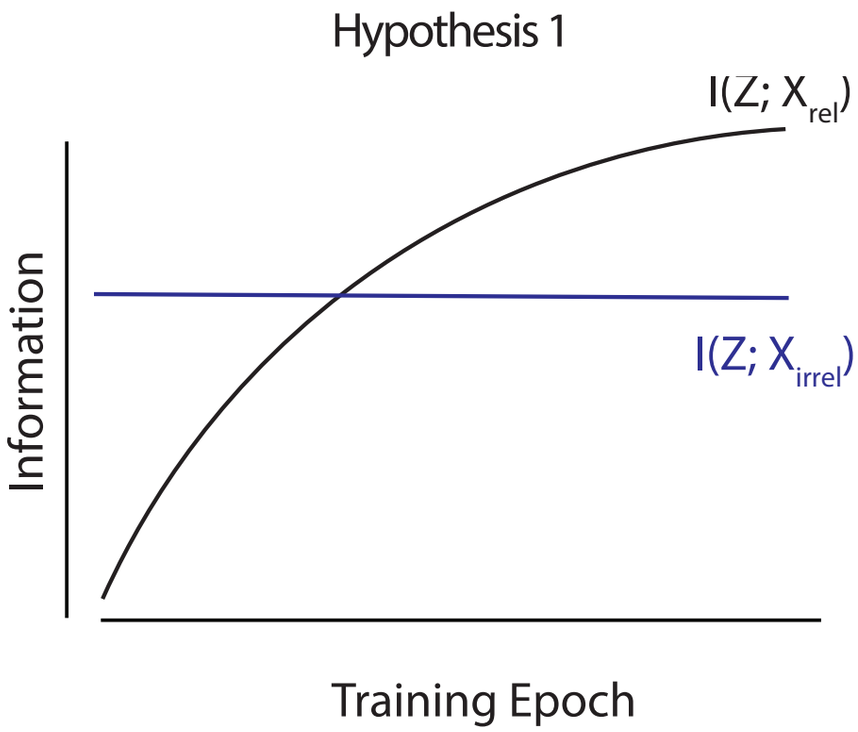
Possible learning dynamics

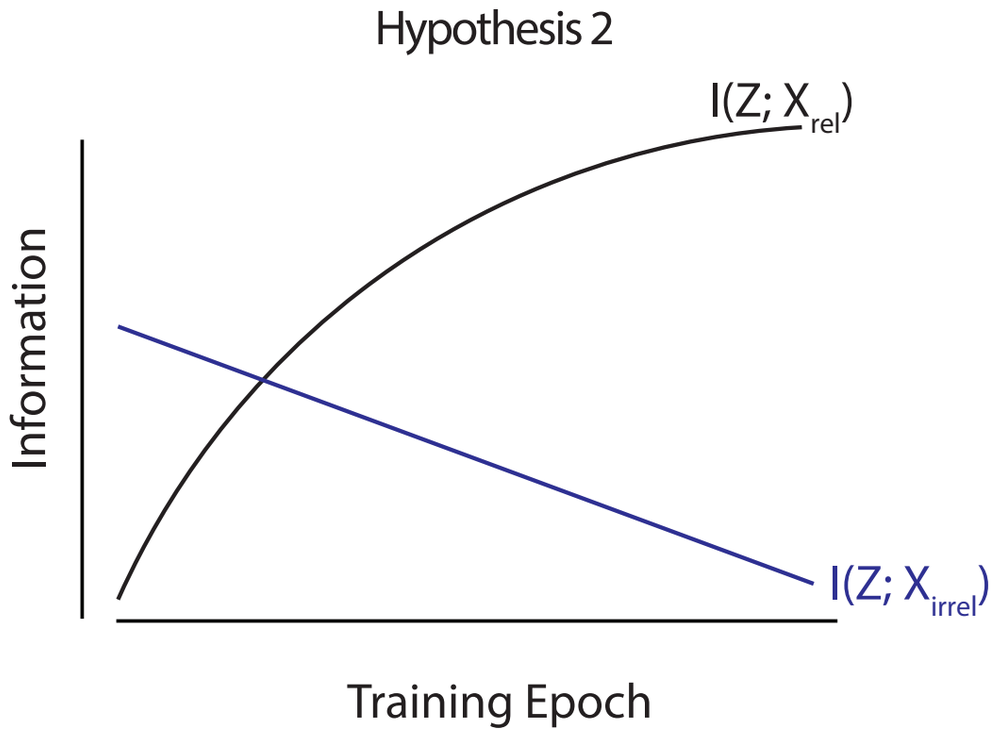


Possible learning dynamics

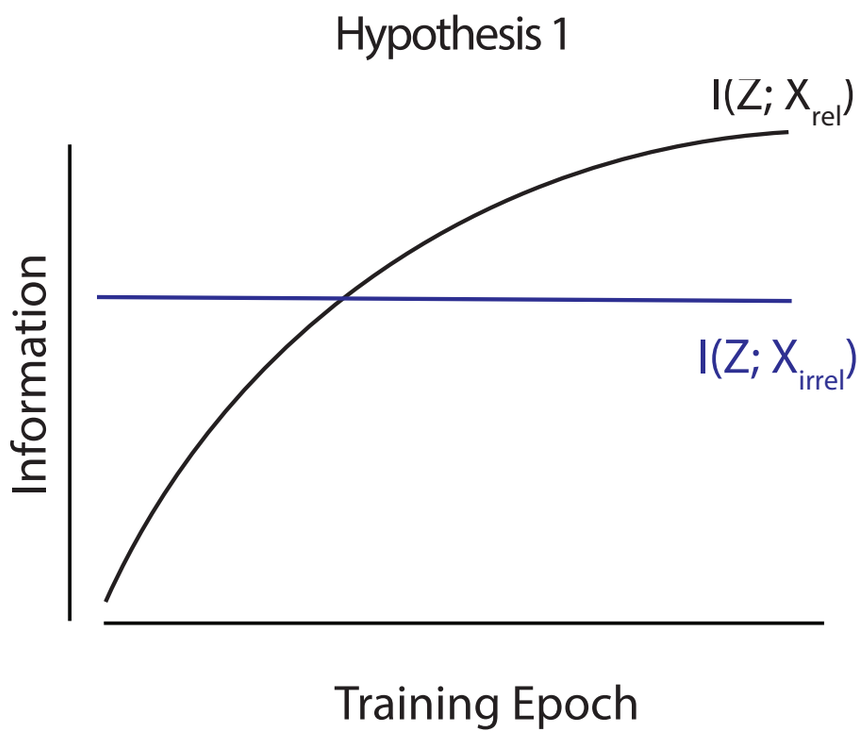


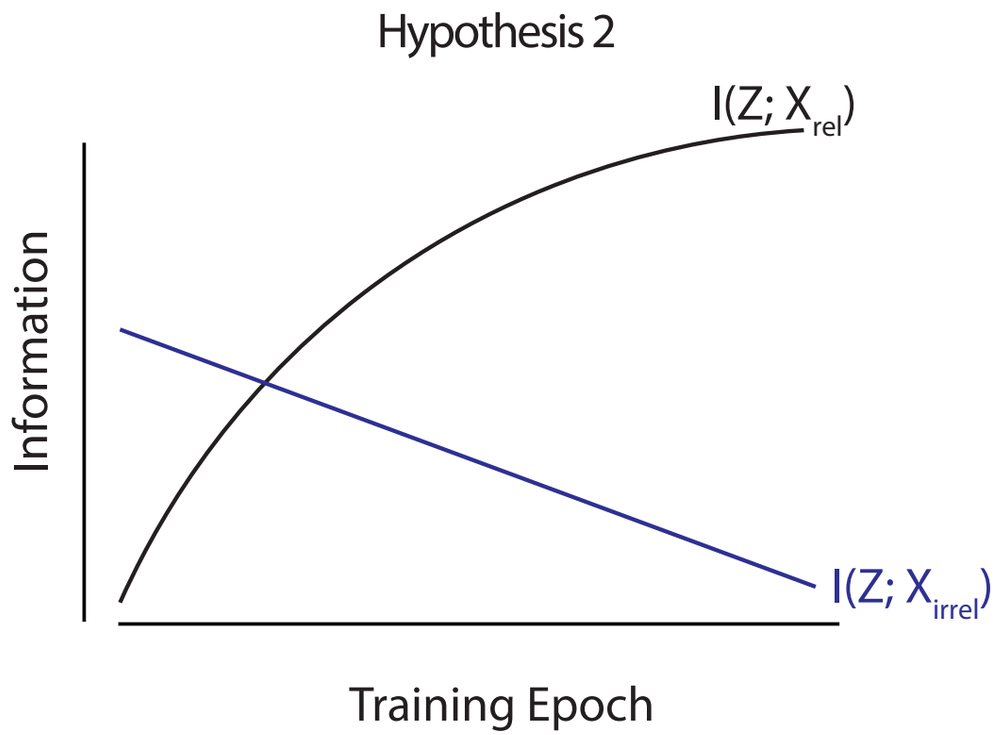
Possible learning dynamics

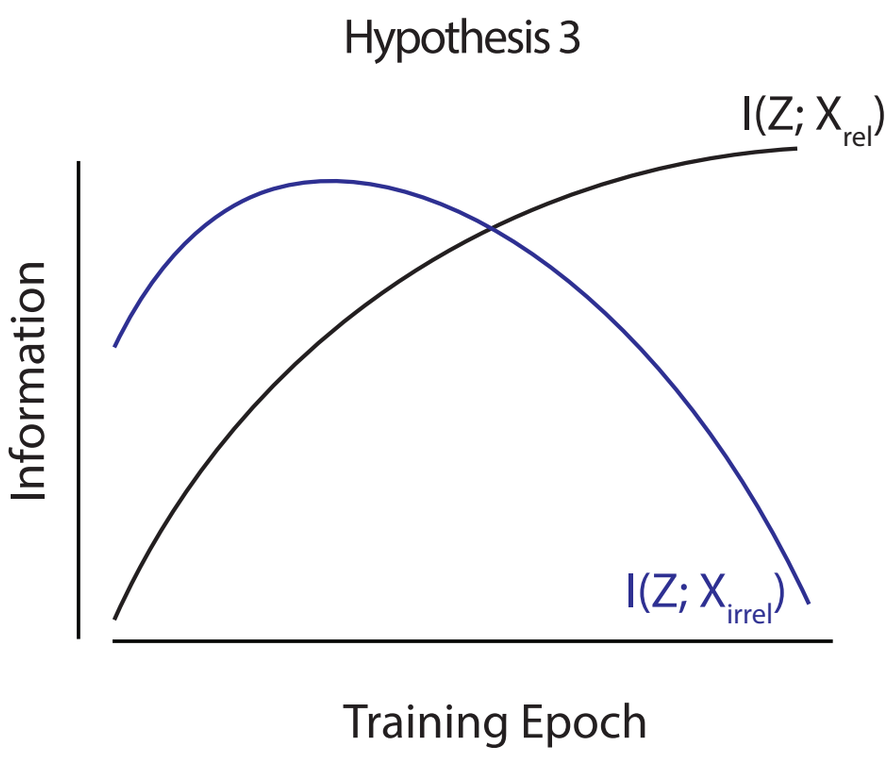




Possible learning dynamics

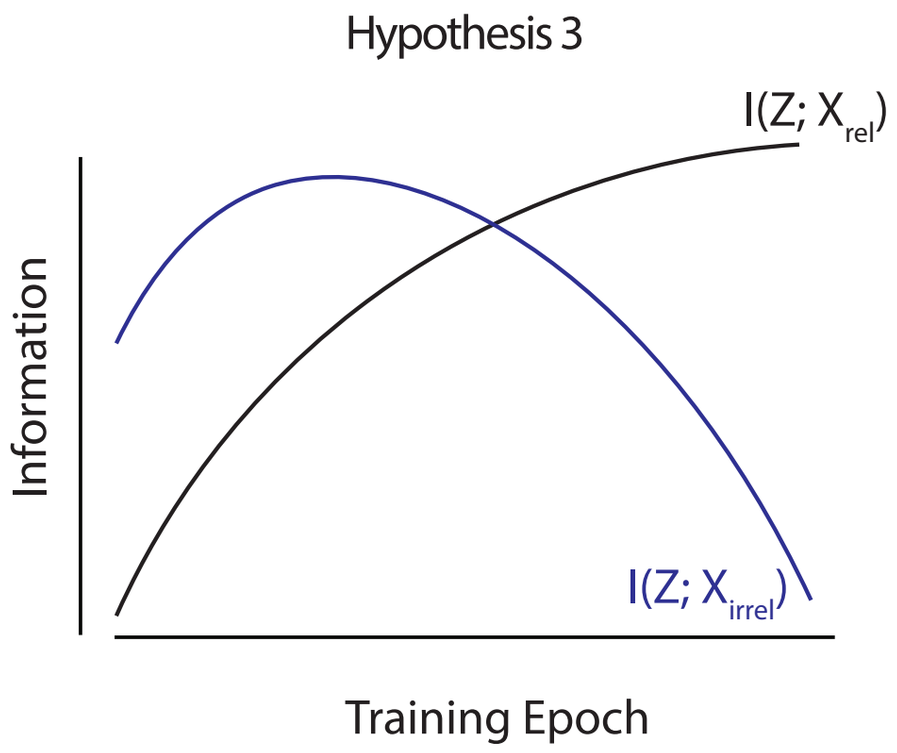






Possible learning dynamics

- 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)

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

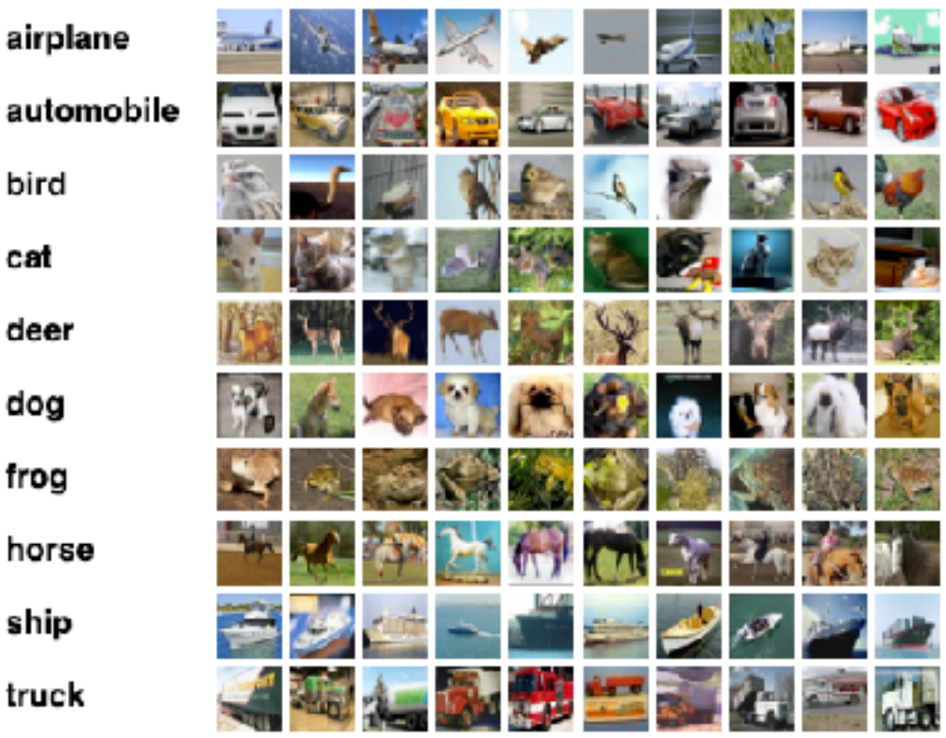
- *H(Y)* is the entropy, or uncertainty, of *Y*
- L_{ce} is the cross-entropy loss on the test set of
- *q(y|z)* is a discriminator network trained to approximate the true distribution *p(y|z)*
- Related to V-Information (Xu et al., 2020)

Usable Information (definition)

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

Property: $I_u(Z;Y) \leq I(Z;Y)$

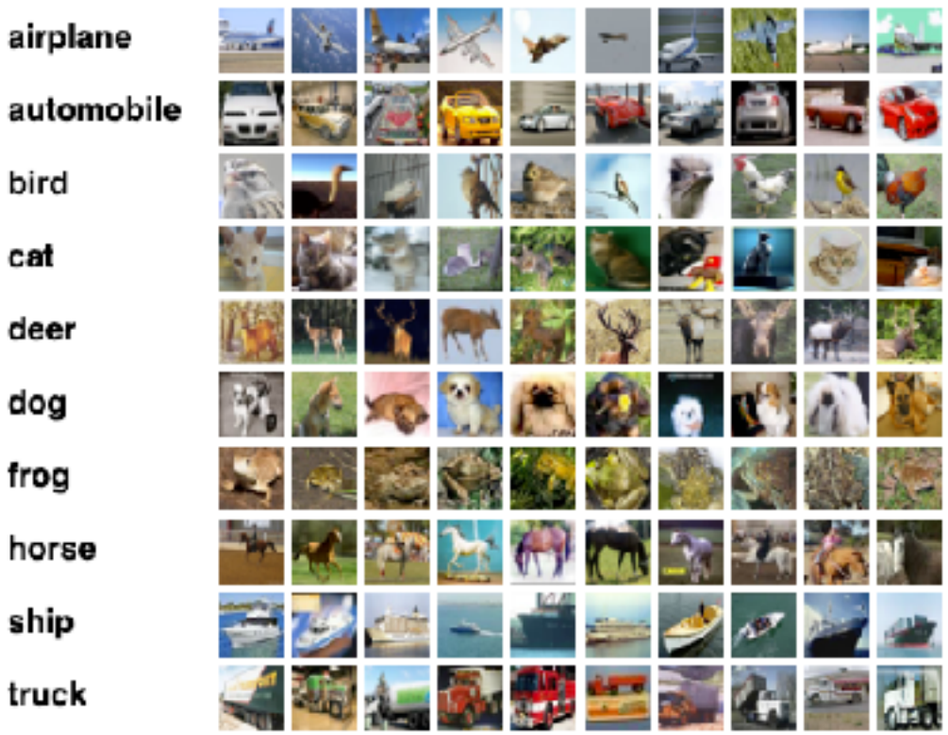
Results: CIFAR-10



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

Fine Labels

Results: CIFAR-10



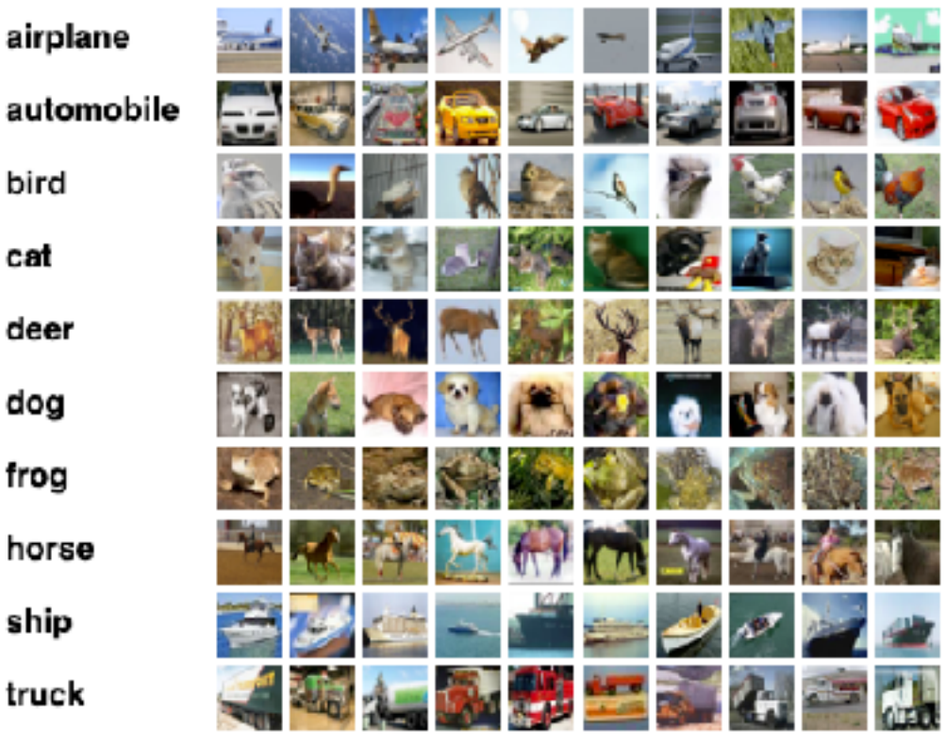
Fine Labels

Coarse Label 1

Coarse Label 2

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

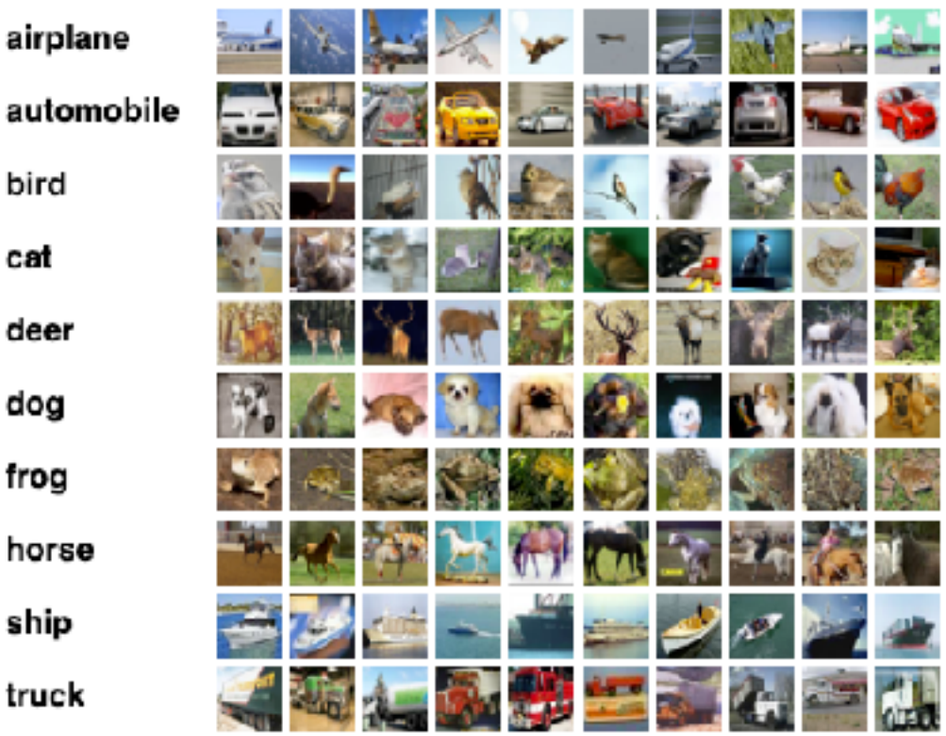
Results: CIFAR-10



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

Task: Output coarse label

Results: CIFAR-10

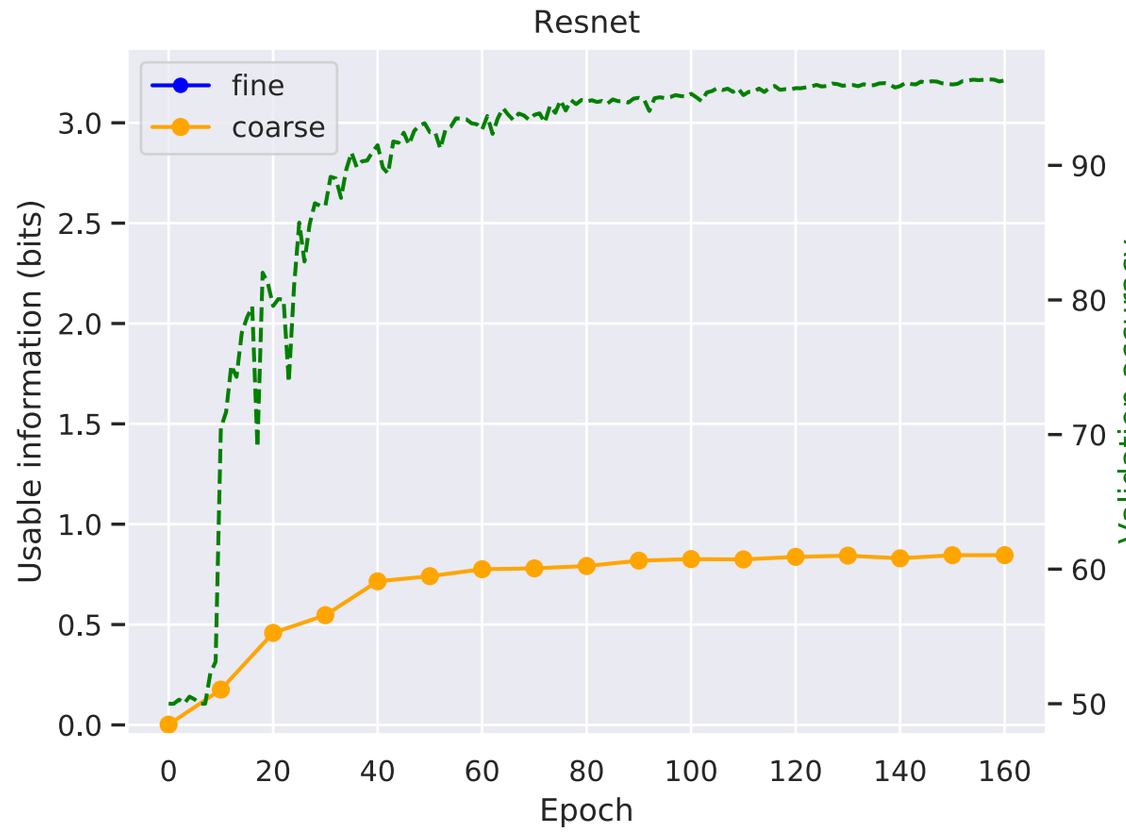


Fine Labels

Coarse Label 1

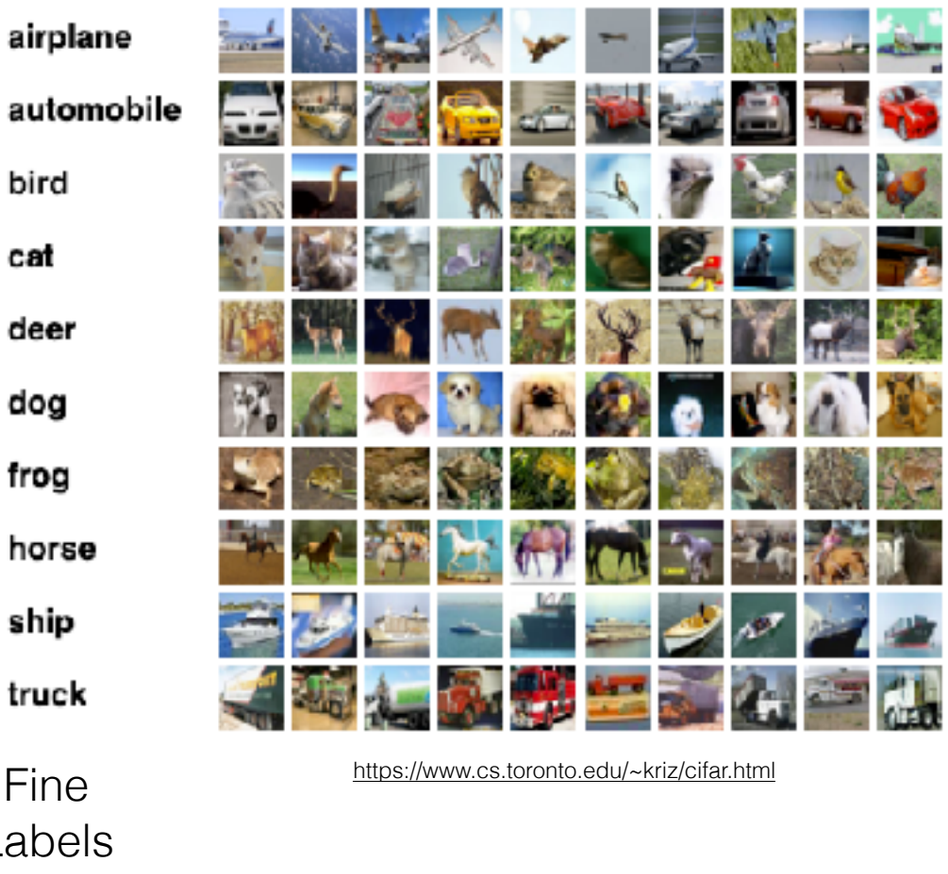
Coarse Label 2

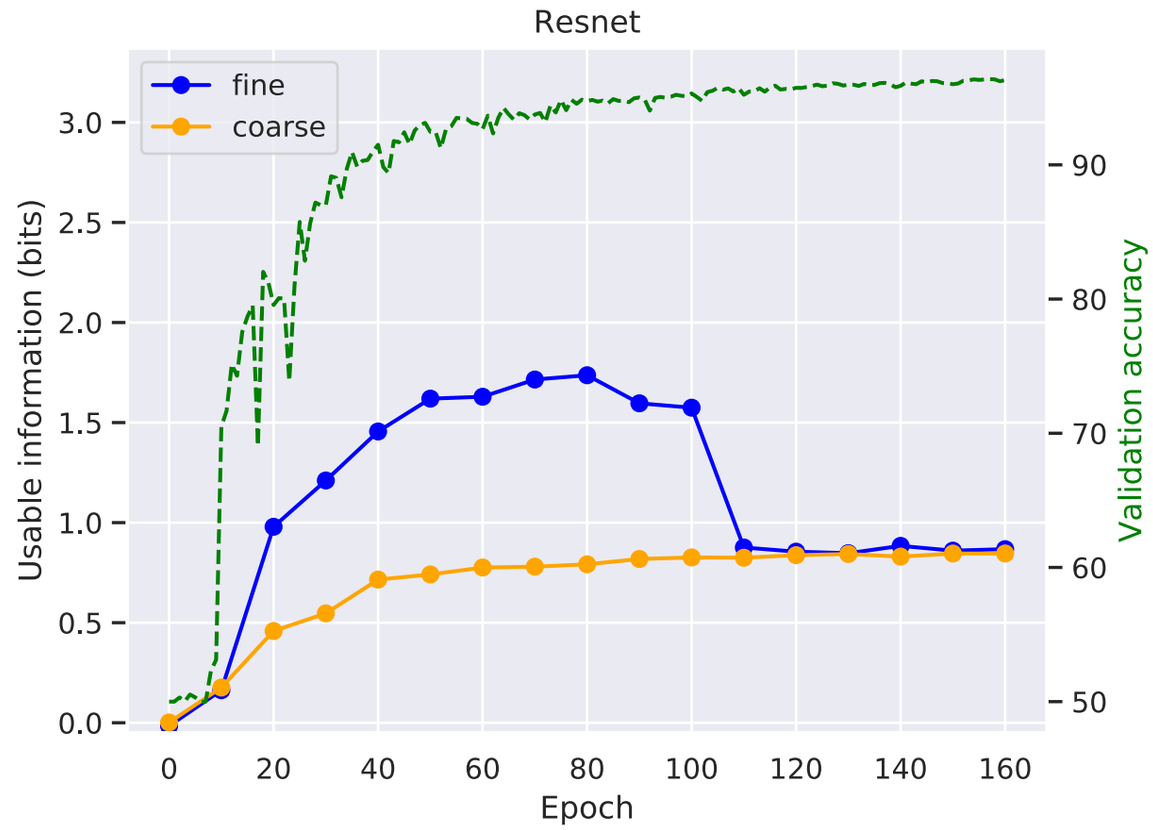
https://www.cs.toronto.edu/~kriz/cifar.html



Task: Output coarse label

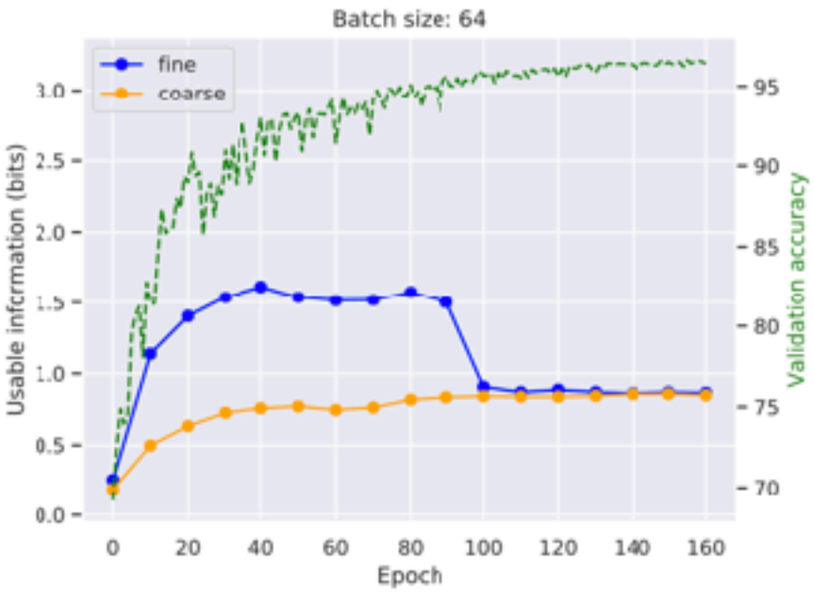
Results: CIFAR-10





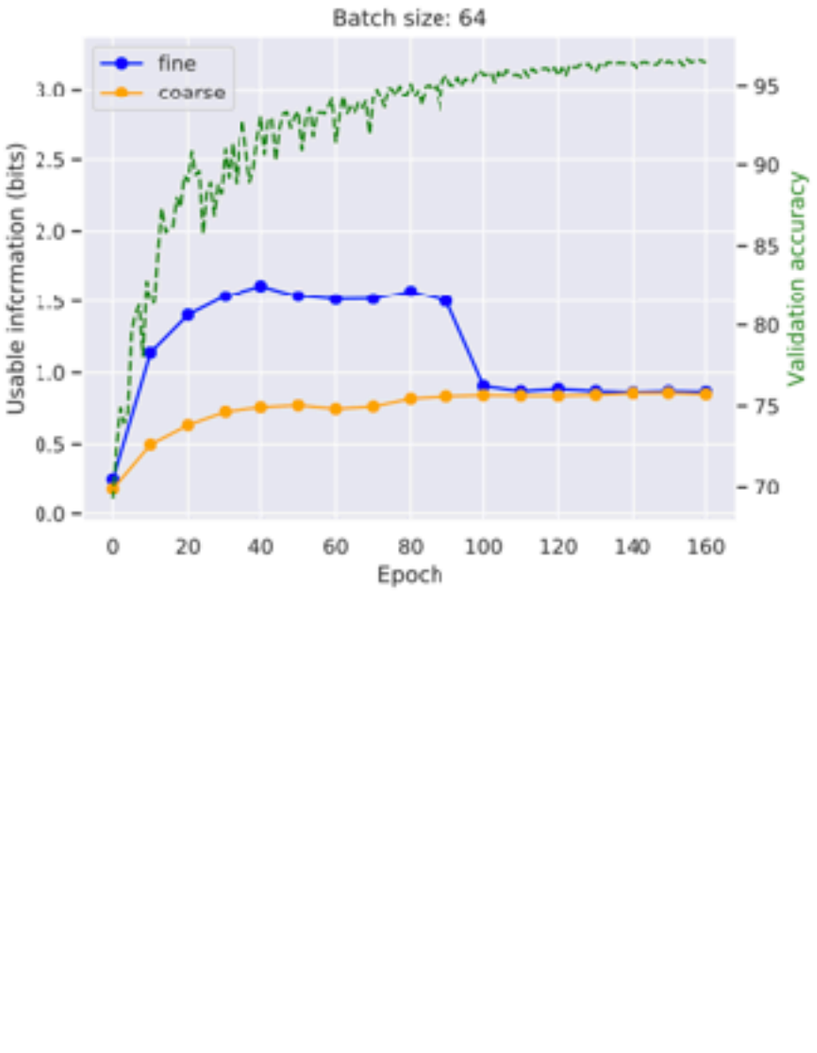
Task: Output coarse label

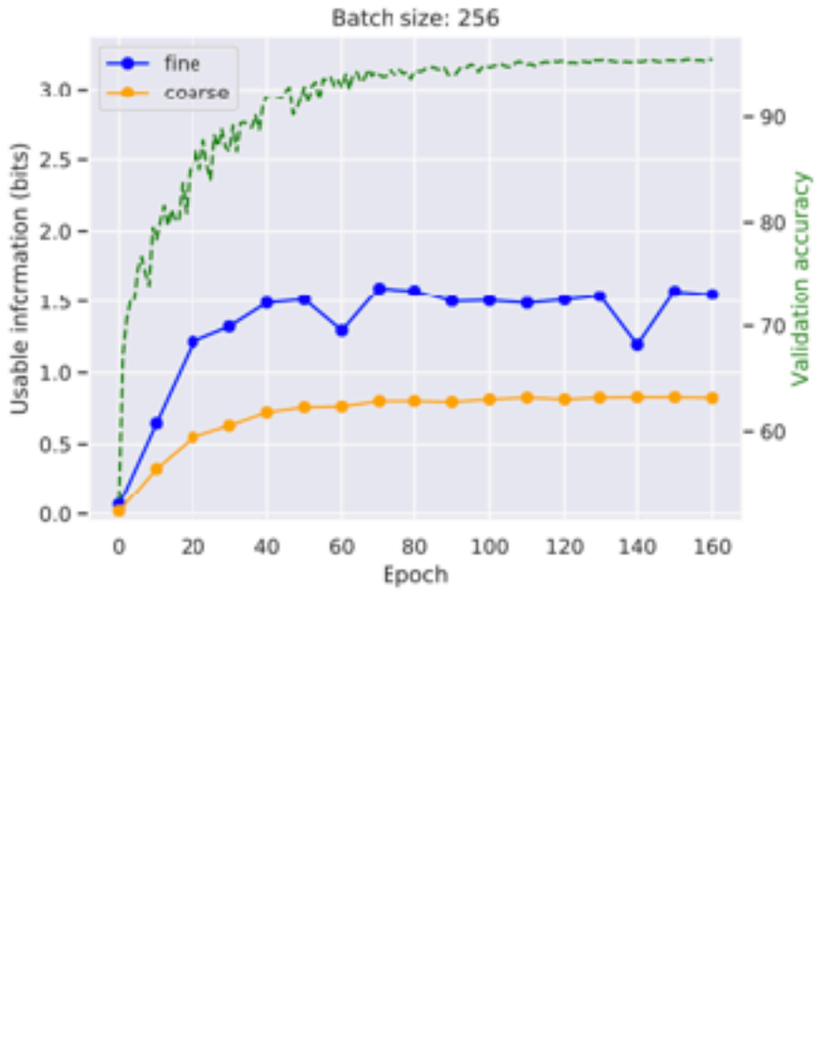
Effect of learning rate and batch size

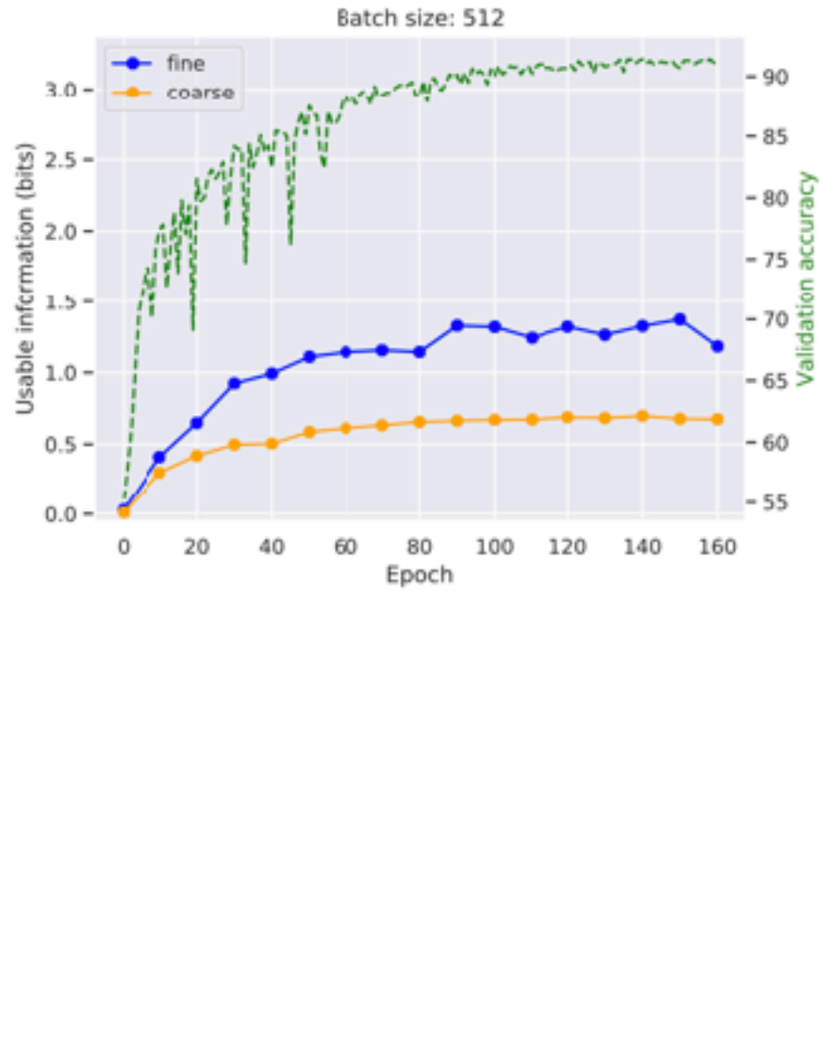


Effect of learning rate and batch size

Increasing batch size →

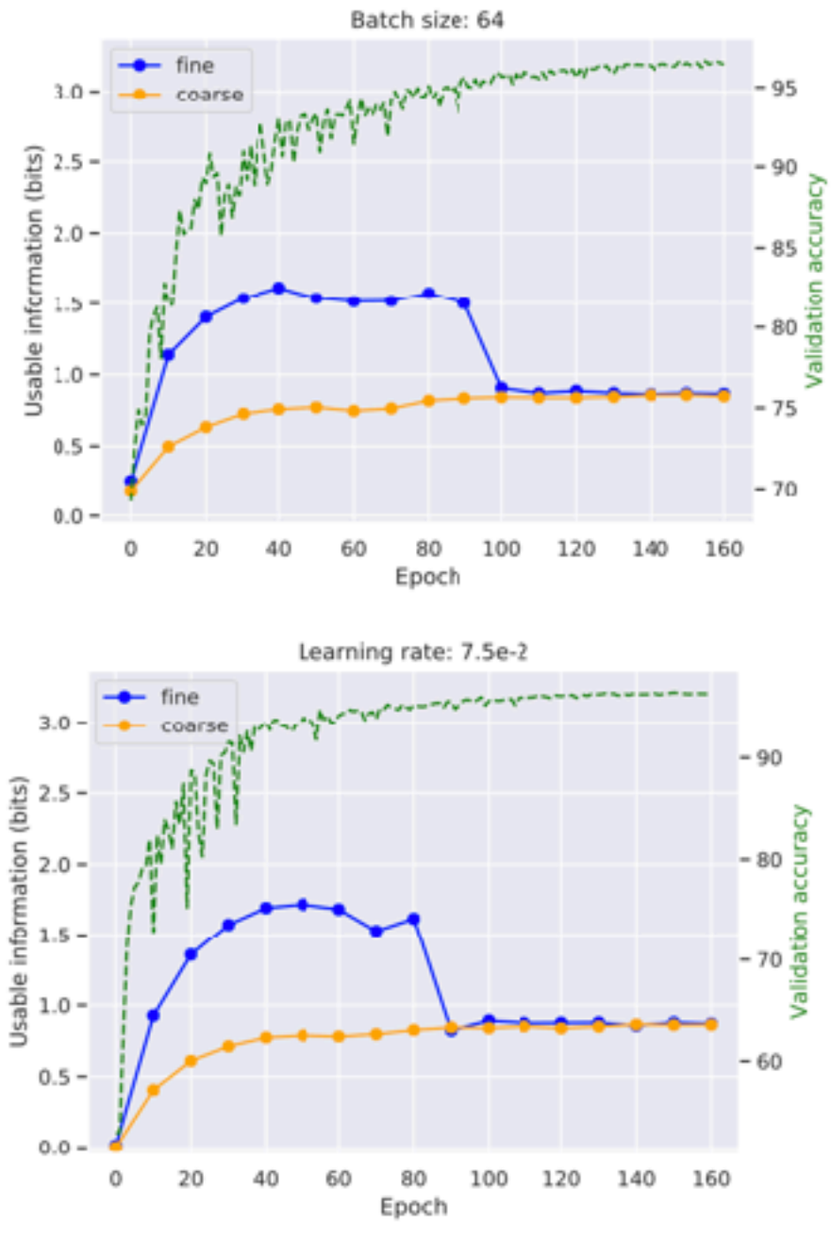


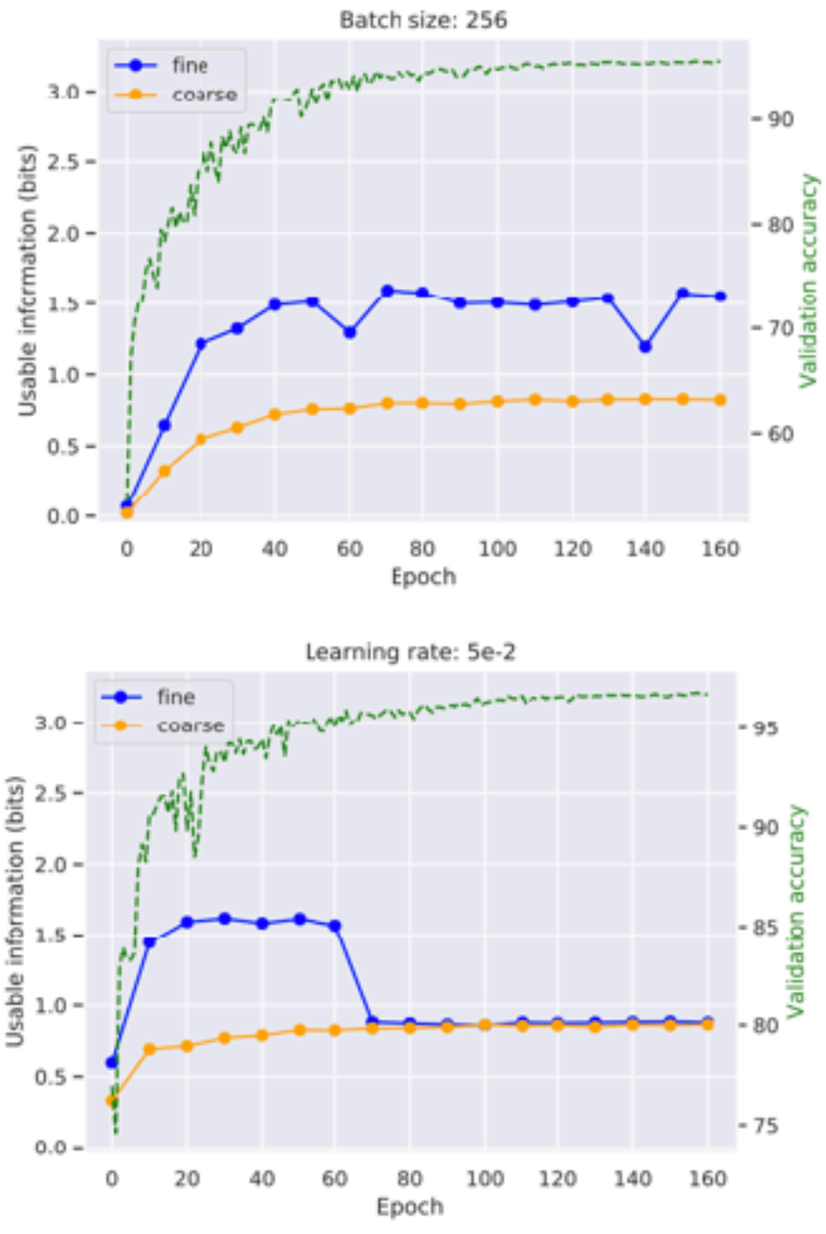


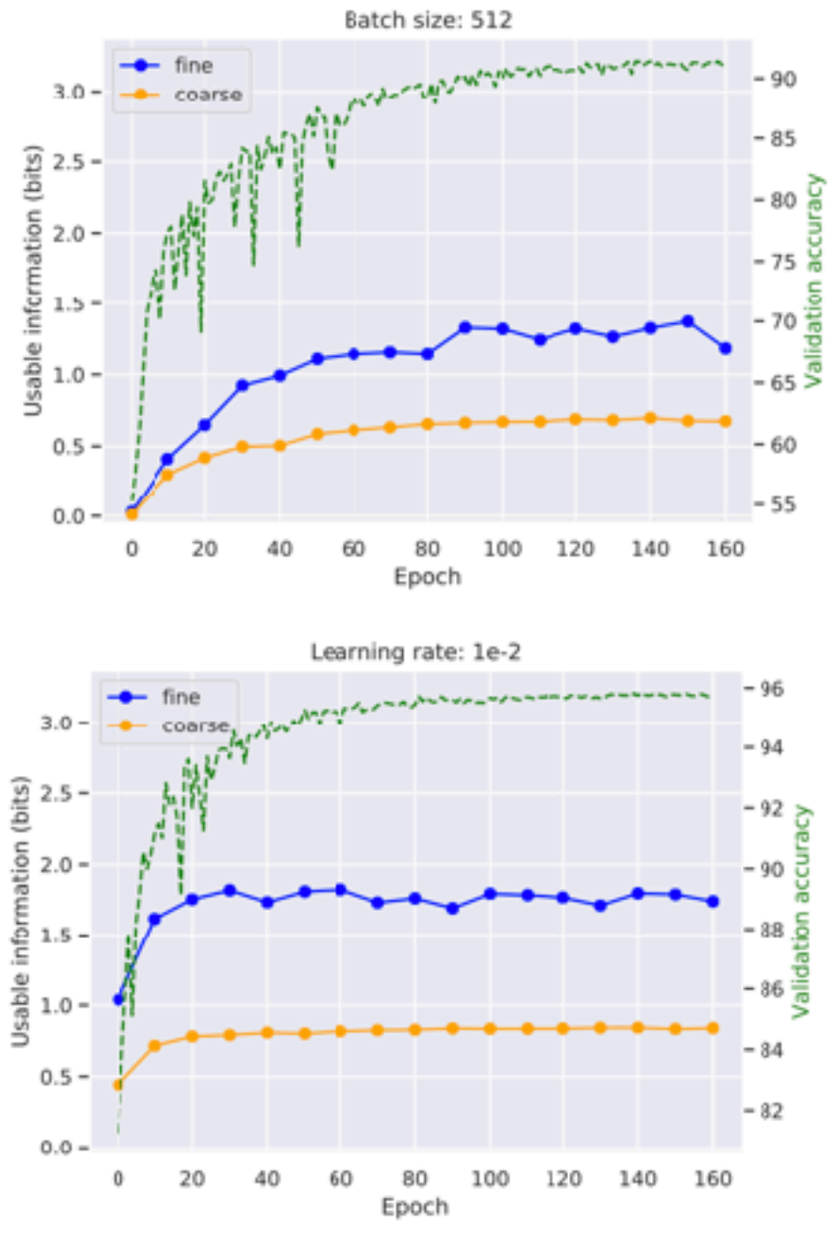


Effect of learning rate and batch size

Increasing batch size →







Decreasing learning rate →

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.