Ablation Studies in Artificial Neural Networks

Richard Meyes
Institute of Information Management in Mechanical Engineering, RWTH Aachen University
Dennewartstr. 27, 52064 Aachen, Germany
richard.meyes@ima-ifu.rwth-aachen.de

Melanie Lu
Institute of Information Management in Mechanical Engineering, RWTH Aachen University
Dennewartstr. 27, 52064 Aachen, Germany
melanie.lu@ima-ifu.rwth-aachen.de

Tobias Meisen
Institute of Information Management in Mechanical Engineering, RWTH Aachen University
Dennewartstr. 27, 52064 Aachen, Germany
tobias.meisen@ima-ifu.rwth-aachen.de

Abstract

1 Introduction

Recent research on deep learning (DL) has brought fourth a number of remarkable applications in a variety of different domains, such as computer vision (CV) (ref: object rec, object loc), natural language processing (NLP) (ref: speech rec, language separation) or continuous control (ref: DRL, Lillicrap, Silver, DeepMind, etc...).

- Recent research, many applications for ANNs
- different domains, CV, NLP, DRL+self learning agents,
- Networks become more complex, grow larger, algorithms become more sophisticated
- methods to understand the networks did not develop accordingly, or at least were limited to a specific viewpoint developed within the domain of AI research.
- recent growth in complexity of the networks raises the question if a viewpoint from neuroscience is may be helpful to gain new insights about mechanisms within the networks that make them work as they do
- prominent method is ablation study to answer questions about localization of knowledge representation within networks.
- e.g. topographical mapping of motor cortex
- can something like a mapping be found within neural networks
- neuroscience past, ablation studies were useful to learn something about functionality
- however, ethical problem with destroying permanently parts of the brain
- those methods can be used for mice, maybe cats in smaller quantities, but not for primates let alone humans
- even though ANNs are vastly different from brains, the methods inspired from neuroscience may help to understand some aspects of these ANNs that have been not previously looked at

- in this paper: MNIST trained MLP
- different locations knockout and different proportions
- look at how knockouts influence the accuracy of the network
- different knockouts show class selective influence, other knockouts show strong global influence
- some classes seem to be split onto several units (different knockouts, different parts of same class influenced)
- mostly damage to the network, but in some cases, accuracy for some classes improves
- gives rise to the notion, that specific influence on weights can improve the networks accuracy
- this specific weight adjustment must be carefully done, rather than just continuing backprop training

2 Related Work

- pruning! (purposefully removing parts of a neural network)
- https://jacobgil.github.io/deeplearning/pruning-deep-learning
- https://www.inference.vc/pruning-neural-networks-two-recent-papers/
- Method is similar, ablate some structural components
- goal however is to speed up the model and reduce complexity
- then DeepMind single directions
- https://deepmind.com/research/publications/importance-single-directions-generalization/
- bridge between neuroscience and AI to understand each other a little better
- In this paper, we complement this approach
- specifically: damage and recovery?
- some related work from neuroscience about ablation studies or lesion studies?

3 Background and Methods

- ablation study and recovery training in simple MLPs on MNIST Dataset
- simple MLP trained on MNIST
- selectively ablating single units, test accuracy of the network
- characteristics of the unit which determines its importance WITHOUT a functional test?
- similar to estimate whether a banana is gonna taste well without actually eating it
- characteristics such a color, firmness, etc...
- correlates of unit importance to some characteristics
-
- ablation study and recovery training in VGG19 on ImageNet Dataset

4 Results

- MNIST - MLP
- Ablation in first layer, different kinds of effects on functionality
- Class selective representation
- question: what makes a neuron important?
- answer: weights distribution difference between untrained and trained state? Correlation evidence
- usual effect: negative on performance, however ALSO POSITIVE EFFECTS but smaller magnitude

- question: can the positive effect be achieved without the negative?
- ImageNet - VGG19
- ablation, depth dependent effects on performance
- effects are class selective
- for different classes, different depth dependence, but overall trend
- damage recovery successful, potential of the recovery depending on damage to the network

5 Conclusions and Future Work

...

Acknowledgments

...

References