Paper: *How transferable are features in deep neural networks?*

Nirajan Koirala

CSC 9010_001

Summary

Deep neural networks are found to exhibit a common phenomenon: the first few layers learned by the network are features similar to Gabor filters and color blobs, where as later layers learn features that are specific to the target tasks. So in the way networks learn, we can call their first layers features general and last-layers feature specific. Authors raise and try to tackle various questions regarding the learning mechanism of different layers within a neural network in this paper. In particular, is it possible quantify the degree to which particular layer is general or specific, when does exactly the transition from general layers to specific takes place and what is the mechanism behind this transition. They found that splitting the network makes the optimization hard since neurons in neighboring layers work adaptively for detecting features. Also, a degradation in performance is observed when features are transferred without fine-tuning and few benefits are occurred when features are transferred between more dissimilar base and target tasks. Additionally, by initializing a network with transferred features from any layer we can boost generalization performance after fine-tuning with a new dataset.

Authors construct pairs of non-overlapping features from ImageNet and randomly split the 1000 classes into two groups with each containing 500 classes. Two networks A and B are trained with 8-layers CNN and they are called baseA and baseB. For the B3B and B3B+, the first 3 layers are copied from baseB and remaining 5 layers are trained on dataset B. However, unlike B3B which freezes the weights from previous 3 layers, B3B+ is made to fine-tune the weights. A3B and A3B+ are made to work same as the B3B and B3B+ except they copy the first 3 layers from baseA and train on dataset B and this network is used to understand the generality of the layers. They further provide the description of their approach on ImageNet and in section 4.2 show that features transfer poorly when the datasets used are less similar.

As expected performance drop is observed in BnB network which is a evidence that the original network contained fragile co-adapted features on successive layers and this co-adaptation could not be relearned by the upper layers alone. Graph of BnB+ shows that when copied, lower layer features also learn on the target dataset and performance is similar to the base case. Fine-tuning in BnB+ prevents the performance drop observed in BnB networks. In AnB networks, the performance starts to decrease and do not rise again which indicates the final layers are feature specific. For AnB+ networks a surprising effect is observed in which they are able to generalize better than those networks trained directly on the target dataset. This suggests that transferring features boosts generalization performance even if the target dataset is large.

Hence, despite the similarity between target and base dataset, freezing the previous few layers produce competitive results to pre-trained models. Transferability is negatively affected by optimization difficulties related to splitting the networks in the middle of fragilely co-adapted layers and the specialization of higher layer features to the original task at the expense of performance on the target task. Either of these two issues may dominate depending upon where the features are transferred from. Authors are also able to quantify how the transferability gap grows as the distance between tasks increases. And finally, they found that initializing with transferred features can improve generalization performance even after substantial fine-tuning on a new task which could be a useful technique to improve the network's performance.

Strengths

- The methods described in the paper of transferring features are intuitive and easy to follow.
- Paper is highly organized.
- Figures and tables provided in the experimental setup are really helpful.

Weaknesses

- Captioning on the figures could have been made a little shorter.

Points of Confusion

- I couldn't understand why transferring features from even distant tasks is better than using random features.

Discussion Questions

- Would the network overfit the training set while transferring features if the target dataset is really small?
- Would features transferred also help us in localization tasks?
- Is it also possible to transfer features in RNNs?