Jenish Maharjan **Week 8**

Paper: *How transferable are features in Deep Neural Networks?*

Summary:

Yonsinski et al in this paper, as the title suggests, discuss how transferable are features in deep neural networks. The motivation behind their study is a commonly observed phenomenon in deep neural networks – the first few layers of the network learn features similar to Gabor filters and color blobs while the later layers learn features that are specific to the target tasks. In this paper they raise three issues – quantifiability of the degree to which a particular layer is general or specific and the point of transition (at which layer?) and nature of transition (sudden or gradual?) from general layers to specific layers in the network. Answers to these questions would help, according to the authors, to understand the extent to which features could be transferred – transfer general features and retrain on target set for specific features. They describe and adapt the usual transfer learning approach where first a base network is trained on a base dataset and task and then transfer the learned features to a second target network to be trained a target dataset and task. They also implement both freezing and fine-tuning in various degrees in their approach.

In their work, they create pairs of classifications tasks A and B by constructing pairs of non-overlapping subsets of the ImageNet dataset. They randomly split the 1000 classes of the ImageNet dataset into two groups making sure that the classes in task A and B are semantically different. They train two networks A and B with 8-layers CNN – baseA and baseB, one for each of the two tasks. They extend these networks to four variations –BnB, BnB+ (Selfer networks) and AnB, AnB+ (transfer networks). In BnB and BnB+, the first 3 layers are copied from baseB and the remaining 5 layers are trained on the dataset B. BnB freezes the weights for the copied 3 layers while BnB+ does not. AnB and AnB+ work the same as BnB and BnB+ except that they copy the first 3 layers from baseA and train the remaining layers on dataset B. AnB and AnB+ help to understand the generality of layers – if performance of AnB and BnB are comparable, it can be understood that the first 3 layers of A learn general features.

The authors, in the "Experimental Setup", note that their focus in this study is to learn about transfer results on a well-known architecture rather than to maximize absolute performance. For this purpose, they performed three experiments. Their main experiment had a random A/B split for which they were able to obtain top-1 error of 37.5% error which is lower than the top-1 error attained on the 1000 class network. The other experiment consisted of a man-made/natural split and the third experiment had random weights. The accuracy stays the same for removing less than 3 layers for all the networks which indicates that the network is learning features that can be generalized despite the complexity of the dataset. For the selfer BnB and transfer AnB, both of which freeze the weights from pretrained models, the performance drastically decreases after chopping three more layers. Unlike selfer AnB, the performance of selfer BnB improves when only the last layer is removed. Both BnB+ and AnB+ have increased performance regardless at which layers networks are split and interestingly AnB+ performs better than BnB+.

The authors, in this work, demonstrated a method to quantify transferability of features from each layer of a neural network. They showed the negative relation between transferability and optimization related to splitting 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. They also showed that initializing with transferred features can improve generalization performance even after substantial fine-tuning on a new task.

Strengths:

- The paper is well written and the language used is easy to read and understand.
- The paper uses bullet points when talking about their observations which increased the readability of the paper.
- The graphs and the textual content for the results/observations complement very well.

Weaknesses:

- Some sections in the paper could have been shortened and broken down into smaller sections to increase the readability.

Confusions:

- Wouldn’t retraining the network on the same dataset (like B3B and B3B+) lead to overfitting?
- Would these findings and observations be consistent with other domains and tasks?

Discussion Questions:

- For what kind of dataset or tasks would transfer learning not work?
- A brief discussion on Gabor filters and color blobs would be really helpful.
- Many computer vision tasks use ImageNet to pretrain CNNs. Why is this so successful and what could be the limitations?