# Discussion Quiz #3

Ashutosh Tiwari (ashutiwa@iu.edu)

February 14, 2022

## 1 Papers

### 1.1 Paper 1

GoogLeNet

### 1.2 Paper 2

Inception v3

## 2 Questions

### 2.1 What are the drawbacks presented in paper 1 for uniformly increased network size?

Authors present two drawbacks of uniformly increasing network size

1. Bigger networks are prone to over-fitting
2. Bigger networks demand more computational resources for training and predictions

### 2.2 Describe how an inception module handles multi-scale processing?

In case of Inception's architecture, every stage has kernels of different sizes stacked parallel to each other and then the last layer of the module just does a concatenation, so that next layer can abstract important features from different scales simultaneously.

### 2.3 Paper 1 uses auxiliary classifiers (similar to deep supervision) connected to intermediate layers, What is the author's explanation for this?

Authors point out that one of the biggest successes of shallower, small networks is that sufficient gradients can flow all the way through to initial layers. To do this they intend to make layers in the middle of network very discriminative and for that reason they attach auxiliary classifiers to these middle layers.

Auxiliary classifiers are nothing but simple classifiers whose loss is added to final loss of the whole network with some discounted weights. There only purpose is to regulate training and thus they are ignored at prediction time.

### 2.4 1x1 convolutions have a dual purpose. Explain.

1×1 convolutions are used for:

1. reducing dimensions before increasing them using 3×3 and 5×5 convolutions. Authors wanted to have a check on computational complexity of the network. This helps them achieve that in one of the ways.

2. Their second purpose is to be used as rectified linear activation units.

### 2.5 Why it is preferred to use 1x1 convolutions inside inception modules? How does it help?

Authors point out the necessity of creating a local optimal block and then repeating same block over and over again spatially. This intuition is based on a previous work which suggests a multi-layer construction with highly correlated clusters into same groups. But problem with this type of construction is that dimensions can blow up pretty quickly. Because of this reason authors use $1 \times 1$ convolutions mostly to reduce dimensions before blowing them up in next layer with $3 \times 3$ and $5 \times 5$.

### 2.6 Explain in detail :) , why did the author cite reference [1] in paper 1?

Authors point out that their work is inspired by Network-In-Network paper along with this internet theme in the sense that they are including a layer of organization in Network-In-Network work and they are going deeper (as the architecture of model is) as this meme suggests.

### 2.7 What changes were made inside an inception module in paper 2 as compared with the inception modules in paper 1?

There are series of changes inside inception module:

1. First they factorize $5 \times 5$ block into two $3 \times 3$ blocks. Then they go one step ahead and factorize $n \times n$ into $n \times 1$ and $1 \times n$. While doing so they point the reason to be computational efficiency.

2. Secondly they expanded the filter banks with asymmetric convolutions to avoid representational bottleneck, with the same logic as the last part.

3. Authors found out that auxiliary classifiers didn't contribute to convergence early in the training. Therefore these are replaced with BatchNorms.

   They included label smoothing.

## 2.8 What are the two problems described in paper 2, which happen are a result of the model being too confident about its classification prediction?

The two problems which are result of model too confident in its predictions are:

1. It may result in overfitting or in other words it doesn't generalize and assigns full probability to what it thinks is the correct label.
2. Secondly it increases the difference between most confident logit and other logits, and thus the final update has a limit because of this difference combined with backpropagated gradient.

## 2.9 What are the general design principles presented in paper 2. Give at least 4 points.

Authors present four general design principles as part of this paper, which are result of various experiments on different architectures and are supposed to work as guiding principles.

1. First point they mention is to avoid bottlenecks in early layers of the network. They suggest various ways of addressing this issue. They point out that probably it should be avoided with compression.
2. They say that local high dimensional information is easier to process and these networks train faster when combined with high number of activations per tile.
3. Dimensionality reduction of spatial representations helps in faster learning without loss of any representation power.
4. This point presents a general guiding principle to manage the width and depth of network. They say that we need to balance both and that can only be done when both are increased at the same time.

## 2.10 What are the advantages of factorizing a convolution are presented in paper 2? Does this factorization leads to any loss of expressiveness for the neural network?

There are two types of factorizations in the network. One is the factorization of bigger kernels in small ones. And second is the spatial factorizations of the filter banks into asymmetric convolutions.

Authors point out two advantages of these changes. One and the bigger one is decrease in number of trainable parameters. Second is the increase in representational dimension because of different orientation of features, mostly in case of asymmetric arrangement of filters.

Authors point out that there is they ran several experiments and deduced that there is no loss of expressiveness because of these changes, particularly when they use rectified linear units in all stages of the factorization.

# References