Discussion Quiz #5

Ashutosh Tiwari (ashutiwa@iu.edu)

February 28, 2022

1 Papers

1.1 Paper 1

Wide RN

1.2 Paper 2

ResNeXt

2 Questions

2.1 Explain how the architecture of a Wide ResNet is different from original ResNet architecture? (3pts)

Wide Resnet attempts to increase the width of network. This increases the size of feature maps for each convolutional layer. They suggest two the biggest advantages of same are decrease in number of layers and significant decrease in training time of the network. They do it without losing quality of the network. They are at least as good as ResNet.

2.2 Explain how identity connection in ResNets might be a potential problem for training a network? (3pts)

Authors point out that having identity skip connection is a weakness. It can provide a escape path for the network to avoid to learn anything. This might result in very few residual blocks actually contributing to any learning. They call it **diminished feature reuse**.

2.3 List all the problems in the original ResNet that a wide ResNet tries to solve? (3pts)

These are the problem Wide RN tries to solve:

1. ResNet is designed on basis of Circuit Complexity Theory which states that wider networks need exponentially more number of parameters. Therefore ResNet was designed to be too thin.

2. Diminished feature reuse, where a lot of residual blocks contribute very less towards the final objective.

2.4 Explain the sentence, "The main power of deep residual networks is in residual blocks, and that the effect of depth is supplementary". (3pts)

Authors acknowledge that ResNet is good because of presence of residual blocks and not because they are deep. They point out that authors of Resnet were forced by circumstances to make their networks deep because they were very thin. They seems to suggest that somebody can train resdual networks which are wide and will still be able to get comparable performance.

2.5 State the three simple ways to increase the representational power of residual blocks. (3pts)

Authors list down these three ways to increase representational power of residual blocks

1. adding more convolutional layers per block
2. widening convolutional layers by adding more feature planes
3. increasing filter sizes in covolutional layers

2.6 B(M) denotes residual block structure, where M is a list with the kernel sizes of the convolutional layers in a block. What does B(3, 1, 1) denote? Explain how it is similar to network-in-network architecture. (3pts)

B(3, 1, 1) is equivalent to having a sequence of 3×3, 1×1 and 1×1 and 1×1. It is similar to Network in Network architecture since the later 1×1 and 1×1 kernels act as a network attached to 3×3 convolution kernel. Each 1×1 kernel acts as a small layer and the combination is an output which has an increased receptive field and is like an output generated from a neural network.

2.7 State the relations of a number of parameters of a model with deepening factor (l), the number of ResNet block (d) (2pts)

l and d are block deepening factor and number of blocks respectively. Authors point out that total number of parameters in a network are proportional to both of them and therefore authors try to keep the multiplication of both of both of them constant, i.e. one decreases whenever one increases.

2.8 Describe in detail the architectural difference between paper 1 and paper 2? (2pts)

Main architectural difference between paper 1 and paper 2 is that in ResNext, authors do not try to go wide, instead they are in favor of separate paths capturing different types of features. In paper one or Wide Resnet, authors try to increase width or features in every residual block.

Apart from this big difference, this decision leads to other small changes which are a repercussions of this previous decision. For example in Wide resnet, they use Dropout, while in second they don't.

2.9 Explain the operation of a simple neuron in artificial neural networks with non-linear activation as a combination of splitting, transforming, and aggregating techniques as described in paper 2. (3pts)

Authors point out that a simple neuron does weighted sum of all its inputs and passes on the aggregated result. They say that this operation can be thought of as a combination of splitting, transforming and aggregating.

Splitting because vector $\boldsymbol{x}$ is sliced into lower dimensional space. Transformation because it scales i.e. applies linear transformation to each of vector $\boldsymbol{x}$ dimensions and Aggregation because it sums up all the embeddings of vector $\boldsymbol{x}$.

2.10 Describe briefly. (a) Basic block (paper 1), (b) Bottleneck block, (c) deepening factor and Widening factor, (d) Cross-validation, (e) cardinality (5pts)

1. **Basic Block**

   Basic block is a block of sequence of similar convolutions. In paper authors give an example of 3×3 followed by 3×3. Basic block has deepening factor, l =2 and $k = 1$.

2. **Bottleneck block**

   Bottleneck block is the block with a larger convolutional kernel is in between kernels with lower dimensions. Bottleneck Architectures are when an input is passed though a set of layers which are of lower dimensionality that input and output layers. In most of cases bottle neck architecture are used to lower the computational considerations.

3. **Deepening factor and Widening Factor**

   Deepening Factor or as authors represent it l is number of convolutions in a block. Widening factor k is the width or in other words the multiplication factor which multiplies with number of features to expand the dimensions.

4. **Cross-validation**

   Cross validation is the practice of dividing training dataset in K different sets and then using $K - 1$ sets to train and 1 set to evaluate the learning algorithm. this creates the opportunity to evaluate and train model on K different sets of training and validation data.

5. **Cardinality**

   Cardinality is the number of independent paths available in a residual block. For Resnext authors have this split-transform-merge residual block which splits tensor $\boldsymbol{x}$ into n paths (which is cardinality) before concatenating it back as the residual block ends.

References