Discussion Quiz #2

Ashutosh Tiwari (ashutiwa@iu.edu)

February 7, 2022

1 Papers

1.1 Paper 1

Deeply supervised nets

1.2 Paper 2

Network in Network

1.3 Paper 3

Distilling the Knowledge in a Neural Network

2 Questions

2.1 How deep supervision in a network help it learn highly discriminative feature maps?

Authors' version of deep supervision means that at each layer of their network they introduced SVM or softmax classifier along with companion loss. If we look at **equation 3** in paper, we can notice that final objective function is a sum of final loss function and sum of all these companion losses present with all layers. Since network tries to optimize this loss function, it forces it to create discriminative feature maps.

Authors point out that by making use of this feature quality feedback they are able to influence the updation of latent variables to favor discriminative feature map learning at each hidden layer.

2.2 What are the three aspects in convolutional neural network style architectures are being looked at in paper 1?

Paper explores these three aspects of convolutional NNs

1. transparency and contribution of hidden layers in overall classification scheme
2. discriminativeness and stability of learned features
3. effectiveness of training in presence of exploding and vanishing gradients

2.3 How deep supervision is different from normal supervised training? How deep supervision affects (a) Training, (b) Backpropagation

Deep supervision is different from normal supervised training in the sense that it includes a term in objective function which has a term which takes into consideration the discriminative and sensibility in features at each hidden individual level. In other words, it is insured at every logical level of network that it learns according to a strict criteria instead of abstract concepts in case of normal neural network. It does same by somehow including terms in objective function which measure the difference between representations learnt vs expected ones.

1. **Training**
   It affects training by including an extra term in objective function. This forces NN to learn feature maps that are more discriminative towards the target classification task. They do point out that this might result in gradient vanishing and exploding depending on supervision from true labels.

2. **Backpropagation**
   There is no fundamental change in the way backpropagation proceeds. However there are few qualitative changes. For example authors point out that using companion objective acts as a kind of feature regularization. Secondly they point out that in their experiments it results in a faster convergence.

2.4 What are differences you observe in the feature visualization shown in Figure 3 of paper 1. explain briefly

In figure 3 authors show top 30% activations from each feature map. Feature maps from deep supervision network for sure look for intuitive. We can see that features are of more prominent components from image and they make more sense to humans.

In case of image on right, these feature maps are not intuitive and show that network is trying to learn abstract concepts not perceivable to humans.

2.5 Label smoothing is a technique to train a model to be less confident about its prediction by smoothing out the one-hot vector, eg. from [0, 1, 0] to [0.1, 0.8, 0.1]. How the soft-labels generated using a teacher model is different from it?

Label smoothing and knowledge distillation are both types of soft labeling. The general training objective for both of them is

$$\mathcal{L}(\widetilde{q}, p) = (1 - \alpha)\mathcal{L}(q, p) + \alpha\mathcal{L}(\hat{q}, p)$$

, where $\mathcal{L}(\widetilde{q}, p)$ is the final output. However purpose of both of them is very different.

In label smoothing, second term comes from a uniform distribution over all classes. In other words its purpose is to regularize so that it doesn't overfit.

In knowledge distillation, second term comes from the teacher network and first from the student network. Along with regularizing the student network, paper claims that it also increases the detection horizon of the student network.

2.6 How the softmax activation is modified to control the soft-labels generated by the teacher model? What is the relation between the temperature T and soft-labels produced?

In equation 1 of paper 3, authors have shown the formulation of q_i. As you can see, from equation itself, as temperature T increases the softer is the probability distribution over classes. In normal cases this value is generally 1.

While distillation, the student model is generally trained on transfer set with a high value of T but while making predictions these values are set to 1 and then the student model is used to make predictions.

2.7 What is the implicit assumption made by a conventional CNN model? How paper 2 address this problem? Explain.

Conventional CNNs make an assumption that latent features are linearly separable. Authors attempt to solve this problem by choosing multilayer perceptrons. These MLP layers are attached to the convolution layer and then this combination is slided over whole image. Because these MLPs are fully connected networks, which are universal function approximators, these tend to work better.

2.8 "The MLP is shared among all local receptive fields" Explain this sentence from paper 2.

The fully connected neural network or the network authors call MLP is connected to the convolution layer. Feature maps from these convolution maps are then passed on to the FC network. Now because this FC network is same for a convolution layer authors make that statement. In other words it means that weights(latent variables) learnt for an MLPConv layer are shared by all feature maps from a layer.

2.9 Define Global Average Pooling (GAP) and state its advantage(s). Also, explain how the application of GAP is more robust to spatial translations?

There are different types of pooling operations. Their purpose is to downscaling the feature maps without learning any new parameters, as opposed to convolutions themselves. These use fixed tensor operations (instead of learning transformations) to downscale the feature maps. Average Pooling is one of them. It is just the average of all values of the feature map, where size of feature map is called pool size. In case of GAP, pool size is equal to size of output features of last layer.

GAP was first introduced in this paper. Paper talks about three advantages of GAPs over traditional FC layers which it tries to replace. Since their outputs are directly fed to softmax, its output can directly be interpreted as category classification confidence. Secondly as explained in last paragraph, it has no parameter to optimize. And the last is that since it averages out the feature map, it is more robust to spatial translations.

The last statement means that, as it does not pass any spatial information through them they are ignorant of spatial invariabilities. In other words, their output is same as long as feature map has same set of values, doesn't matter arrangement of those.

2.10 Describe each term briefly: (a) Companion Objective, (b) deep-supervision, (c) Teacher and student networks, (d) Knowledge Distillation

1. **Companion Objective**
   Companion objectives are class of objective functions used in deep learning architecture. They are called companion objectives because they normally are summed up with overall output layer loss and similar to that it also depends on the output class. In case of semi-supervised and unsupervised settings these companion objectives are normally dependent on variables which overall function is dependent on.

2. **Deep Supervision**
   Deep supervision is just a way to include the class based discrimination at every layer. There can be many ways to do so. One can be including a loss which depends on target label (like this paper), other can be forcing decoder at every layer to output semantic output. Ultimately benefits from deep supervision is two fold. First is, it results in strong regularization in form of strictness in forcing each layer to learn a specific format. Secondly it results in faster convergence.

3. **Teacher and Student Networks**
   Teacher and Student network represent a format of knowledge distillation or a knowledge transfer method, where a larger teacher network is used to train a small student network which can then be deployed in production.

   Intuition is that the student model which is generally much smaller than the teacher model, instead of learning all cumbersome features and representations, is only required to learn mapping between input and output tensors.

4. **Knowledge Distillation**
   Papers refers knowledge distillation as transferring knowledge from a massive, expensive and non-efficient model to a lean model, which can be deployed in production. This knowledge transfer is generally paramount in success of these kinds of strategies because student network being small in size is generally not able to learn on its own from the training dataset and thus requires mappings (feature or labels) from teacher network.

   There are many proposed methods to do that. In some cases smaller models tends directly learning mapping from input to output vectors, while in some other cases they try learning the feature maps. There is a third kind as well where student networks try to learn relation between successive feature maps and then try to scale that to further layers.

References

1. Affects of soft labeling in DNNs
2. Unsupervised companion objective
3. Deep Supervision using forcing decoding consistency at each layer
4. Knowledge Distillation Survey