Computer Vision Discussion Quiz – 3

1. Answer:

   The discovery of a discriminative classifier that has been trained on more discriminative features functioning better than one that has been trained on the less discriminative data, inspired them to make the network acquire feature maps that are extremely discriminative. This finding suggests that the effectiveness of a discriminative classifier trained with hidden layer feature maps could act as a substitute for the selectivity of such hidden layer feature maps, as well as affect the performance of higher layer feature maps and, ultimately, impact the hidden state filter upgradation to promote extremely discriminative feature maps with optimal use of quality feedback of the features for every hidden layer of the network.

2. Answer:

   The three elements of CNN type architectures that are being analyzed are as follows: First one is about the openness of the layers, the overall classification's transparency of the intermediate layers, second one is the performance of features in starting layers, learnt features' selectiveness and durability, particularly in the initial layers and the third one that's being addressed is about gradients, training productivity with the existence of vanishing and exploding gradients.

3. Answer:

   Deep supervised learning addresses few drawbacks of the normal supervised learning. The following are some of the major pitfalls of supervised learning frameworks: lowered transparency and selectivity of hidden layer features, training problems due to exploding and vanishing gradients, and, even with some theoretical attempts, there is absence of a comprehensive mathematical interpretation of algorithmic ways, reliance on huge amounts of training data, and ambiguity of manual tuning during while training. Deep supervised learning attempts to reduce these disadvantages by presenting a companion objective to each of the hidden layers. Deep supervision impacts

   A) Training: As in deep supervised learning, a companion objective is being introduced to the individual hidden layers and is utilized as an extra constraint to the learning process. Existing supervised training functions will perform better as a result of this. Considering the optimization function's local strong convexity, this illustrates that deep supervised learning has a higher convergence rate than standard supervised training. This suggested strategy in deep supervised learning avoids pre-training and focuses on decreasing the output classification error whilst lowering the prediction error of every successive layer.

   B) Back Propagation: In deep supervised learning, we know that a companion objective is enforced at the convolutional layers. This enforcement of companion objective at layers helps in backpropagation. With this companion objective we backpropagate the classification error information to the deeper convolutional layers. In normal supervised learning, we backpropagate form the final layer, but with deeply supervised nets we can backpropagate not just from the ultimate layer, but also from our localized companion output at the same time.

4. Answer:

   From the figure of visualization of feature maps of Deep Supervised Networks and Conventional Neural Networks, we could see that the DSN-learned feature maps are much clearer compared to CNN-learned feature maps. We could see that the clarity of DSN-learned feature maps is more than CNN-learned feature maps. The DSN-learned feature maps are more intuitive, clear, automatic. From

the figure, the DSN-learned feature maps are more transparent compared to CNN-learned feature maps. These are the differences that I could spot from both figures.

5. Answer:

   Overfitting can be reduced by using a method called label smoothening. We don't declare any image as belonging 100 percent to a given label using this method. Because declaring every picture with a 100% probability would result in overfitting, we utilize label smoothening to proclaim that image as belonging to a certain label 70% or 80% of the time and distribute the remaining likelihood to other labels. We allocate in this manner and assign a label to train. This process is known as label smoothening, and it helps to prevent overfitting.
   Soft labels, on the other hand, are not the same as label smoothing; they are the predictions that are created after delivering an image to an already trained model or the teacher model. Soft labels are the probabilities that are created in this manner. For good results, we employ the created soft labels and the picture to train the smaller network or student network.

6. Answer:

   We incorporate a temperature variable T in the softmax function, which helps to regulate the softness of the labels. In general, there will be times when only a few of the soft labels are significant. The training would be unsuccessful if we used these negligible soft labels. So, we can use the Temperature variable T, of the softmax function and increase the T value to shorten the space between soft labels and make the inconsequential soft labels as effective as possible. The final categorized label would not change as a result of the rise in T value, but the soft labels proportionality would alter. So, now that we have well-proportioned soft labels for training, the training will be more effective since we have better data. T value is directly proportional to the softness of the variables. As T increases, variable softness increases, as T decreases, variable softness decreases.

7. Answer:

   When the instances of the latent ideas are linearly separable, that is, when the variations of the ideas all exist across one side of separation plane, GLM can attain a high level of abstraction. As a result, traditional CNN implicitly assumes that the latent ideas are linearly separable. Nevertheless, because the information for the similar notion is frequently found on a nonlinear manifold, the approximations that describe such concepts are frequently extremely nonlinear functions of the input. To address this problem, we need a universal function approximator, as it can estimate more depictions of latent ideas. In paper-2 they picked Multilayer Perceptron as universal function approximator to address the latent ideas issue because it is consistent with the design of back-propagation-trained convolutional neural networks. For using same features again and to reduce computational cost, multilayer perceptron could also be a deep model in and of itself. They named this unique form of layer as mlpconv.

8. Answer:

   Multilayer Perceptron is applied same as a convolutional filter. Multilayer Perceptron works same like a filter. As how a filter convolves on an image, in the same way MLP is also convolved on an image. So, MLP is shared across all local receptive fields in the same way as convolution filters are shared among all local receptive fields.

9. Answer:

   Global Average Pooling is intended to take the role of the conventional fully connected layers in conventional neural networks. In this pooling, we calculate the mean of every feature map and this

mean output is used in the soft-max layer, despite of using fully connected layers over the feature maps.

Benefits of Global Average Pooling are: Main benefit is, by mandating correlations among feature maps and categories, Global Average Pooling is organic to the convolution framework. As a result, the feature maps may be simply understood as confidence maps for categories. Another benefit of global average pooling is that there exist zero parameters to tune, therefore overfitting is prevented. This makes it a structural regularizer.

So, now we know that in global average pooling, we take the spatial mean of feature maps through final layer and send to soft-max layer, this process of taking the mean of the spatial data makes the global average pooling more resistant to spatial transitions.

10. Answer:

    a. Companion Objective: Companion objective operates as a form of feature regularization or proxy. The testing error would affect with the Companion objective, it helps in large decrease of this testing error. But companion objective may or may not affect the test error. This can be applied at the convolutional layers to backpropagate the classification error advice to the bottom convolutional layers. And, when the training data is limited, it also leads to speedier convergence.

    b. Deep-supervision: In deep-supervision we employ network in network. So, instead of linear filters in CNN, using a multilayer perceptron which constructs a micro neural network is called as Deep supervision. In deep supervision, we could combine global average pooling on feature maps to make it far less subjected to overfitting than fully connected layers. Here in deep supervision, feature maps are made by moving micro networks over the input and sending to the proceeding layers.

    c. Teacher and Student networks: While performing knowledge distillation, that means, while shrinking the big ensemble model's knowledge into one model, the big ensemble model is the teacher network, and the solo post compressed model is the student network. So, the bigger network is the teacher network and smaller one is the student network. For the student network, rather than training on raw information, the student model is taught to replicate the outcome of the bigger ensemble network that means the teacher network's outcome. Rather than Boolean or probabilities, the teacher network results in float labels or soft labels. And as the student acquires from the teacher, and the teacher network is aware of the subtleties, this helps the student in learning more effectively.

    d. Knowledge Distillation: As with a large cast of models for predictions is time consuming and requires a big computational cost. So, for big neural networks this process becomes impractical as it demands more time and computational cost. So, to reduce this problem, we do a process called Knowledge Distillation. Knowledge Distillation is nothing but, condensing the ensemble's knowledge into a solo model, this helps in efficient deployment, as it's considerably simpler to deploy from solo model. In the process of Knowledge Distillation, forwarding knowledge to condensed model is performed through training on a transfer set. Additionally for distribution, on every instance in this transfer set, we utilize a soft target distribution.

Citation: https://arxiv.org/pdf/1409.5185.pdf

https://arxiv.org/pdf/1312.4400.pdf

https://arxiv.org/pdf/1503.02531.pdf

https://www.quora.com/What-is-a-teacher-student-model-in-a-Convolutional-neural-network