CS 7150: Deep Learning — Summer-Full 2020 — Paul Hand

Week 2 — Preparation Questions For Class
Due: Monday April 11, 2020 at 12:00 PM Eastern time via Gradescope

Name: [Sumeet Gajjar]
Collaborators: [Saurabh Vaidya]

You may consult any and all resources in answering the questions. Your goal is to have answers that are ready to be shared with the class (or on a hypothetical job interview) as written.

Directions: Read the articles 'Deep Learning' (Three Giants Paper) and 'Visualizing and Understanding Convolutional Networks'. (ZFNet)

Questions on Three Giants:

Question 1. *Briefly explain the selectivity-invariance dilemma.*

Response:

Answer 1. *Selectivity and invariance are considered to be important ingredients in biological visual systems. The features or representations should be selective to the aspects of the image that are important for discrimination, and at the same time invariant to irrelevant aspects such as the pose of the animal, scale or rotation.*

Question 2. *Provide a sentence (or two) from the Three Giants Paper that you find interesting and would like to discuss in class.*

Response:

Answer 2.

- *How is backpropagation handled for convolutional and pooling layers in CNN?*
- *What Distributed representations are and how does it help in generalization during learning?*

Questions on ZFNet:

Question 3. *Summarize the contributions of the paper. Be as succinct yet comprehensive as possible.*

Response:

Answer 3. *The ZFNet paper proposes a novel visualization technique to get insight into the function of intermediate feature layers of CNN and understand the functions of the classifier. The authors demonstrate how this can be used as a diagnostic tool to improve already existing CNN architecture and they achieve this by finding a model architecture that outperforms Krizhevsky et al. on the ImageNet classification benchmark. The authors also show their ImageNet model generalizes to other datasets. By using the pre-trained model and retraining the softmax classifier, they beat the current state-of-the-art results on Caltech-101 and Caltech-256 datasets. While doing so, the authors question the utility of benchmarks with small datasets.*

Question 4. *Precisely explain what is being plotted in Figure 2. Make sure to explain why the organization/presentation of layer 1 is different than layers 2 and higher. You do not need to summarize the visualization algorithm in your explanation.*

Response: Your answer should be begin with something similar to: A ... net was trained to solve [some task].

Answer 4. *A convnet was trained to solve the problem of image classification. Figure 2 represents visualization of features for validation data in the fully trained convnet. Each 1x1 figure is a reconstructed pattern which indicates high activation in a given feature map. Each visualization has a corresponding image patch.*

The layer 1 detects simple gradients, patterns and lines. Similar images causes almost same section of a feature map to be activated and hence only 1 visualization is necessary for 9 similar images.

From layer 2 onwards the patterns are complex and composite. Similar images are causing activation in different sections of the same feature map and hence each image has a different visualization.

Thus for layer 1, 1 visualization is used for 9 similar images and for layer 2 onwards each visualization has a corresponding image patch.

Question 5. *In Figure 2, why even bother showing the visualizations instead of the patches? Provide two examples where the visualization is needed (and the image patches are insufficient) to understand the qualitative interpretation of a particular feature map.*

Response:

Answer 5. *The visualizations are necessary since the image patches have greater variation and very little in common than visualizations as the visualizations solely focus on the discriminant structure within each patch.*

- *For (layer 5, row 1, col 2), Glancing at the image patches, one can say the feature map is focusing on 4 legged animal, however that is not true. The visualization reveals that feature map focuses on grass in background and not on foreground objects.*
- *For (layer 5, row 3, col 2), The images patches consists of humans, dogs, and cars which have nothing in common. The visualization reveals that feature map focuses on almost circular structure which in humans and dogs it's face and in case of car it's type.*

Question 6. *What do the authors mean by "invariances" of a feature map? Why does the method presented in this paper better show the ÒinvariancesÓ of the feature maps in comparison to gradient- or Hessian-based analysis of the provided neuron?*

Response:

Answer 6. *According to the authors the invariances of a feature map means the ability to be agnostic to translation, scaling or rotation of the images.*

The gradient analysis method requires careful or any initialization and does not provide any information on neuron's invariance.

The Hessian method overcomes the shortcoming of neuron's invariance, however the invariance in the higher layers is complex and is poorly captured by a simple quadratic equation.

The method presented in this paper provides a non-parametric view of invariance in form of visualizations, showing which patterns from the training set activate the feature map. This visualizations are top-down projections that reveal structures within each patch that stimulate a particular feature map.

Hence the method presented in this paper is better in comparison to gradient- or Hessian-based analysis of the provided neuron

Question 7. *The authors saw that their classifier's performance was not invariant to image rotation except for the category 'entertainment center.' Explain what is going on with this category. What could you do to improve the rotational invariance of the classifier?*

Response:

Answer 7. *The classifier is not invariant to rotation except for the "entertainment center" category due to presence of the objects with rotational symmetry in the images. Just by looking at image of "entertainment center" category from figure 5, we can it is almost symmetrical on horizontal and vertical axis. Hence the features of this image rotated by 90,180 and 270 degrees will also be similar and we can conclude this from Figure 5, c3. Since the feature vectors of the rotated images are similar to the original image, they are classified correctly.*

Following are the methods to improve the rotational invariance of the classifier

- *Introduce rotated copies of the image in the training dataset.*
- *We duplicate the filters thrice and rotate each of the duplicated copy by 90 degrees, this produces equivariant feature maps and this approach also requires the classifier to be trained with rotated copies.*
- *We can use feature map back-rotation as proposed by Follmann et. al. in A Rotationally-Invariant Convolution Module by Feature Map Back-Rotation. In this approach, in addition to duplicating and rotating the filters, we back-rotate the feature maps with respect to the angle that the filter is rotated before. This method does not require the training dataset to include the rotated copies.*

Question 8. *How useful/general do you find the process the authors used for improving the architecture of a network by visualizing the activations of its inner layers? What are the challenges of this as a general process? What alternative procedures can you envision for improving neural network architectures?*

Response:

Answer 8. *The process that authors used for improving the architecture of a network can be useful in general given the fact that we how to visualize the activations of inner layers. However expertise or experience is required to identify the problem/issue and apply the necessary improvements. The first and foremost challenge would to develop a visualization technique which can be applicable to not just CNN but all kind of networks. The visualization technique need not be same of other networks., each category of networks can have their respective way to visualize activations. CNN has an inherit advantage to be able to represent the activation in 2D form but this may or may not be true for RNN, LSTM and other category of networks which makes the task of visualization difficult.*

In order to improve the Neural network architecture, we can automate the process that authors suggested and include the same in the learning process. Here once the training finishes, for each layer we

can look for coverage for high and low frequencies, aliasing effects and adjust the filter size and stride respectively. Once modified we retrain the network and this loop continues until a desired performance.

In a deep CNN network we can create visualizations for each layer and then compare the visualizations on consecutive layers or maybe pair of layers within a certain window. We compare the patterns and if there is a similarity within a certain threshold, then it can be a indication that both layers may be doing similar feature extraction. Hence we can modify the parameters(no of filters, filter size or stride) or drop the layer entirely. If the similarity of the patterns is under a certain threshold, it can be a indication that a new inner layer can be introduced. Once the modifications are made we retrain the network to see improvements or deterioration in performance.

Question 9. *Identify which parameters in Figure 3 are user specified, and which are derived from those user-specified parameters. Verify that the derived parameters are correct.*

Response:

Answer 9. ***The following are the user specified parameters:***

- *For input layer: Image size, Filter size and stride*
- *For layer 1-5: No of Filters, Filter size, stride*
- *For layer 6,7: The number of neurons i.e. 4096*
- *For output layer: The number of softmax classes*
- *For layer 1,3 and 4: a padding of 1 in conv and maxpooling layers.*
- *For layers 2 and 5: a padding of 1 in conv layer.*

Note: *In order to verify the derived parameters are correct, we use the following formula to calculate the output of a convolutional or pooling layer.*

$$\frac{w - f + 2 * p}{s} + 1$$

where w is the input size, f is the Kernel size, p is the padding, s is the stride

The following are the derived parameters:

- *Input size Layer 1 conv : 110,* $\frac{224-7+2}{2} + 1 = 110$
- *Input size Layer 1 max-pooling : 55,* $\frac{110-3+2}{2} + 1 = 55$
- *Input size Layer 2 conv : 26,* $\frac{55-5}{2} + 1 = 26$
- *Input size Layer 2 max-pooling : 13,* $\frac{26-3+2}{2} + 1 = 13$
- *Input size Layer 3 conv : 13,* $\frac{13-3+2}{1} + 1 = 13$
- *Input size Layer 4 conv : 13,* $\frac{13-3+2}{1} + 1 = 13$
- *Input size Layer 5 conv : 13,* $\frac{13-3+2}{1} + 1 = 13$

- *Input size Layer 5 max-pooling : 6,* $\frac{13-3}{2} + 1 = 6$

Question 10. *The authors make the claim that their result Òbrings into question the utility of benchmarks with small (i.e.* $< 10^4$*) training sets.Ó Why do they say this? Do you agree? Are small datasets of no value anymore?*

Response:

Answer 10. *The authors trained and tested their convnet for Caltech-101 and Caltech-256 datasets independently. Training the network from scratch on these datasets did not yield high accuracy's as compared to the ImageNet-pretrained convnet. Hence the authors claim that their result Òbrings into question the utility of benchmarks with small (i.e.* $< 10^4$*) training sets.Ó*

However this just shows how the ImageNet trained model can generalize well to other datasets but that does not mean small datasets are of no value.

In my opinion, the small datasets are still of value, since they can used for benchmarking generalization capability of pretrained networks. We can use the pretrained network to generate better representations and retrain the classifier on the small datasets.