CS7150.37338.202130 Day 5 Preparation Questions

Aishwarya Vantipuli, Xiaolin Zhao, Morgan Kohler

TOTAL POINTS

6.5 / 7

QUESTION 1

1 Question 1 1 / 1

✓ **- 0 pts** Correct

- 0.1 pts Some inaccuracies

- 0.3 pts Drawbacks

- 0.5 pts Click here to replace this description.

QUESTION 2

2 Question 2 1 / 1

✓ **- 0 pts** Correct

QUESTION 3

3 Question 3 0.5 / 1

- 0 pts Correct

✓ **- 0.5 pts** Some inaccuracies

- 0.2 pts Explanation difference in plot structure

Be more specific in your observation and interpretation. This is a very important figure and has been extensively discussed in class. See also the justification for why the differences in the structure of the plot.

QUESTION 4

4 Question 4 1 / 1

✓ **- 0 pts** Correct

- 0.5 pts Click here to replace this description.

QUESTION 5

5 Question 6 1 / 1

✓ **- 0 pts** Correct

- 0.5 pts Some inaccuracies

- 1 pts Click here to replace this description.

QUESTION 6

6 Question 7 1 / 1

✓ **- 0 pts** Correct

- 0.2 pts Some inaccuracies

- 0.8 pts Click here to replace this description.

QUESTION 7

7 Question 10 1 / 1

✓ **- 0 pts** Correct

gradescope

CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 5 — Preparation Questions For Class
Due: Wednesday 2/3/2021 at 2:30pm via Gradescope

Names: [Morgan Kohler, Aishwarya Vantipuli, Xiaolin Zhao]

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. Your answers should be as concise as possible. When asked to explain a figure, your response should have the following structure: provide context (state what experiment was being run / state what problem is being solved), state what has been plotted, remark on what we observe from the plots, and interpret the results.

Submit one document for your group and tag all group members. We recommend you use Overleaf for joint editing of this TeX document.

Directions: Read 'Visualizing Higher-Layer Features of a Deep Network'.

- Read only Section 3.

Question 1. *Explain the visualization method the authors propose. What are its drawbacks?*

Response:

The visualization methods authors proposed are: Activation maximization and Sampling.

Activation Maximization:

In this method, we look at input patterns which maximize the activation function used in hidden layer. This could be good because a pattern to which the unit is responding could be a good representation of what a unit is doing internally.

Disadvantage of this technique is choosing number of samples to keep for each unit and combining them. Also, not all input samples contribute to high activation which makes it harder to find what these samples have in common.

One can take a more general view of this method by not constricting oneself to the training and test samples and just set it up as an optimization problem instead. Which is to say, we optimize the input to achieve the highest activation possible. This way, a global or more likely local minima can be found which contains the direct patterns that the function is looking for. The same problem presents itself in that different starting points will result in different local minima however combining samples from different starting points can alleviate this.

Sampling and Clamp:

In a Deep Belief Network, using top down approach we successively sample from last layer and clamp one unit to 1. This will produce a distribution which summarizes the information of units

1 Question 1 1 / 1

✓ **- 0 pts** **Correct**

- 0.1 pts Some inaccuracies

- 0.3 pts Drawbacks

- 0.5 pts Click here to replace this description.

gradescope

by computing expectation. Disadvantage of this is the Markov chain at last layer doesn't always mixes well.

Along with these two new methods, authors also used previously published Linear Combination of previous layer filters technique which involves taking weighted linear combination of previous layer filters to determine characteristics of a unit. However it is hard to choose the right number of filters to keep at each layer for understanding and visualizing units. Also it ignores the non-linear relationship between layers.

Directions: Read 'Visualizing and Understanding Convolutional Networks' (ZFNet).

- Read the whole paper. You can skip Sections 4.3 and 5.3.

Question 2. *Summarize the contributions of ZFNet paper. Be as succinct yet comprehensive as possible.*

Response:

ZFNet is the winner of ILSVRC 2013, an image classification competition; which has significant improvement over other existing cutting edge state of art models at that time. While only shallow layer features can be observed previously, this paper introduces a technique to effectively visualize the CNN kernels in deep layers using Multi-layered Deconvolutional Network to project the feature activations back to the input pixel space.

1. Deconvnet Techniques for Visualization

For the visualization of features, authors used deconvoulutional networks (deconvnet) which uses same components as a convnet model but in reverse. Instead of mapping pixels to features it does the opposite.

Unpooling: Max pooling operation is non-invertible, however we can obtain an approximate inverse by recording the locations of the maxima within each pooling region.

Rectification: ReLU is used as activation which tunes all values to positive. By performing ReLU again, we achieve feature reconstruction.

Filtering: Deconvnet uses transposed versions of convnet filters which means flipping each filter vertically and horizontally.

2. Visualization for Each Layer

Using above technique on ImageNet Dataset, the top 9 activated patterns in randomly selected feature maps are shown for each layer. Layer 2 responds to corners and other edge/color conjunctions. Layer 3 has more complex invariances, capturing similar textures.Layer 4 shows significant variation, but is more class-specific. Layer 5 shows entire objects with significant pose variation.

3. Modifications of AlexNet Based on Visualization Results

To improve performance, ZFNet makes two changes to AlexNet: Reducing filter size and stride in first layer which lead to increase in sampling rate.

By visualizing the convolutional network layer by layer, ZFNet adjusts the layer hyper parameters such as filter size or stride of the AlexNet and successfully reduces the error rates.

2 Question 2 1 / 1

✓ - 0 pts Correct

gradescope

Question 3. *Explain Figure 2. In your explanation, you do not need to summarize the visualization algorithm. Make sure that you response explains why the structure of the plot for Layer 1 is different than the structure of the plot for Layers 2-5.*

Response:

Context:

A convolutional Neural Network was trained to perform ImageNet classification. Then a deconvnet is used to visualize the feature activations on validation set.

What is plotted:

Figure shows feature visualizations when the model is completely trained. By using deconvnet techniques, the top 9 activated patterns in randomly selected feature maps are shown for each layer. For each feature map corresponding image patches are shown.

What we observe:

Filters at layer 1 are a mix of extremely high and low frequency information, with little coverage of the mid frequencies. Layer 2 shows aliasing artifacts. Layer 3 starts to learn some general patterns such as mesh patterns, and text pattern. Layer 4 shows significant variation, and is more class-specific, such as dogs' faces and birds' legs. Layer 5 shows entire objects with significant pose variation, such as keyboards and dogs.

Interpretation:

There are problems in first two layers. Without the mid frequencies in layer 1, there is a chain effect that deep features can only learn from extremely high and low frequency information. Aliasing in layer 2 occurs when sampling frequency is too low. Hence authors tried to reduce filter size and stride to tackle these problem thus increased AlexNet performance.

Question 4. *In Figure 2, why even bother showing the visualizations instead of the raw image 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:

Because the image patches could have great variations, but by projecting each of the top 9 activations separately, different structures that excite a given feature map can be revealed. This is because these visualizations only focused on the discriminant structure within each patch.

e.g.1 In layer 2, row 3, col 1, the patches also seem to be different in terms of color and shape, but they all have certain lines that separate the image as suggested by the visualization, this means that the neuron is able to detect the lines across the image.

e.g.2 In layer 5, row 1, col 2, the patches seem to be very different in terms of object, color and theme, but from the feature map visualization in the left, we can see this particular feature map

3 Question 3 0.5 / 1

- **0 pts** Correct

✓ - **0.5 pts** **Some inaccuracies**

- **0.2 pts** Explanation difference in plot structure

Be more specific in your observation and interpretation. This is a very important figure and has been extensively discussed in class.

See also the justification for why the differences in the structure of the plot.

gradescope

Question 3. *Explain Figure 2. In your explanation, you do not need to summarize the visualization algorithm. Make sure that you response explains why the structure of the plot for Layer 1 is different than the structure of the plot for Layers 2-5.*

Response:

Context:

A convolutional Neural Network was trained to perform ImageNet classification. Then a deconvnet is used to visualize the feature activations on validation set.

What is plotted:

Figure shows feature visualizations when the model is completely trained. By using deconvnet techniques, the top 9 activated patterns in randomly selected feature maps are shown for each layer. For each feature map corresponding image patches are shown.

What we observe:

Filters at layer 1 are a mix of extremely high and low frequency information, with little coverage of the mid frequencies. Layer 2 shows aliasing artifacts. Layer 3 starts to learn some general patterns such as mesh patterns, and text pattern. Layer 4 shows significant variation, and is more class-specific, such as dogs' faces and birds' legs. Layer 5 shows entire objects with significant pose variation, such as keyboards and dogs.

Interpretation:

There are problems in first two layers. Without the mid frequencies in layer 1, there is a chain effect that deep features can only learn from extremely high and low frequency information. Aliasing in layer 2 occurs when sampling frequency is too low. Hence authors tried to reduce filter size and stride to tackle these problem thus increased AlexNet performance.

Question 4. *In Figure 2, why even bother showing the visualizations instead of the raw image 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:

Because the image patches could have great variations, but by projecting each of the top 9 activations separately, different structures that excite a given feature map can be revealed. This is because these visualizations only focused on the discriminant structure within each patch.

e.g.1 In layer 2, row 3, col 1, the patches also seem to be different in terms of color and shape, but they all have certain lines that separate the image as suggested by the visualization, this means that the neuron is able to detect the lines across the image.

e.g.2 In layer 5, row 1, col 2, the patches seem to be very different in terms of object, color and theme, but from the feature map visualization in the left, we can see this particular feature map

4 Question 4 1 / 1

✓ **- 0 pts** Correct

- 0.5 pts Click here to replace this description.

gradescope

focuses on the grass in the background, not the foreground objects, which means the neuron found the pattern of grass and can detect whether the grass region appears in foreground or background.

Question 5. *What do the authors mean by "invariances" of a feature map? Why does the method presented in this paper better show the invariances of the feature maps in comparison to the gradient-based analysis gradient- or Hessian-based analysis of the provided neuron?*

Response:

The invariance of feature map means, when the images are transformed, for example translated, rotated or scaled, how much the feature vectors change before and after transformation.

The problem of gradient-based or Hessian-based analysis of the provided neuron is that, for higher layers, the invariances are very complex, so they are poorly captured by a simple quadratic approximation. But visualization with deconvnet can provide a non-parametric view of invariance, showing which pattern from the training set activate the feature map.

Question 6. *Explain Figure 4.*

Response:

Context:

A convnet was trained to perform image classification, and deconvnet was used to visualize the features.

What is plotted:

The figure plotted the evolution of the strongest activation of a randomly chosen subset of 6 feature maps in 5 layers at epochs 1, 2, 5, 10, 20, 30, 40, 64 through training.

What we observe:

We observed that in lower layers it's roughly the lines/edges in input images or slightly complex combination of the lines; in mid layers it's some patterns/sketches of object that human might understand; in higher layers it's a description of which of these patterns/sketches occur and where are they, or basically how is the input image consisted of these patterns/sketches. Features in higher layers tend to be more abstract and human may not be able to read. We also notice that in each layer the feature maps converge when epochs increase, but with different speed: converge happens within a few epochs in lower layers, but generally the higher the layers, the more epochs it takes to converge.

Interpretation:

The feature maps become more complex/abstract with layers go deeper is because that, in lower layers the receptive field is very small and can only capture patterns within a small

5 Question 6 1 / 1

✓ **- 0 pts** **Correct**

- 0.5 pts Some inaccuracies

- 1 pts Click here to replace this description.

gradescope

neighborhood of pixels, thus results in simple features such as lines, shape and corners; when layers go deeper, the receptive field of each neuron becomes larger and can capture more complicated and abstract features. And this is also why the upper layers would take more epochs to converge compared to lower layers.

Question 7. *Explain Figure 5c. In the interpretation, make sure to explain what is going on with the category 'entertainment center.' What could you do to improve the rotational invariance of the classifier?*

Response:

Context:

A convolutional neural network was trained to perform classification of images from ImageNet. The authors seek to understand how modifying the images changes the activation of the feature maps and the probability of choosing the correct class.

What is plotted:

There are several images used as input to the model shown in various degrees of rotation around the image center. Then there are two graphs representing the canonical distances from the activation of the unrotated input images vs. the images at various degrees of rotations. The two graphs represent the activations at layers 1 and 7. The last graph is a rotation degree vs. the probability of correct classification by the model.

What we observe:

The distances from the unchanged images' feature maps all increase as the absolute value of the degrees of rotation increases. This distance peaks at 180 degrees returning to 0 as the absolute value of the rotation decreases. The first few degrees offset is where the majority of the change occurs and the distance changes more slowly after this. The distances for all classes are much less for layer 7 than layer 1. The distances in layer 7 are not as uniform as the distances in layer 1 with each class falling into different patterns of how the distance changes. The probability of correct classification decreases as the rotation increases however different classes decrease differently.

Interpretation:

Since the feature map difference for layer 7 is significantly less than layer 1 we can conclude that the feature maps of layer 7 are more invariant of rotational changes than those of layer 1. Different classes are more invariant to rotation than other classes. Lawnmower may be the class least invariant to rotation which makes intuitive sense as most of the input lawnmower images are not likely to be rotated by many degrees. The African crocodile feature maps change the least likely based on more variability in the orientation of natural images in the training set. For the entertainment center, the greatest change in layer 7 feature map vectors occur in between intervals of 90 degrees. This is expected as the entertainment center is made up of square sections and is most often presented with the square edges parallel to the image edges.

6 Question 7 1 / 1

✓ **- 0 pts** **Correct**

- 0.2 pts Some inaccuracies

- 0.8 pts Click here to replace this description.

gradescope

$$\frac{W-(F-1)+2P}{S}+1$$
$$\frac{224-(7-1)}{2}+1$$
$$110$$

This is because we must subtract the unused pixels on the edges if the padding doesn't make up for it. Then you divide by the stride to account for the unused pixels in the stride pattern (by unused I mean not directly reflected in the next layer as they are still reflected in the other filters).
After the relu there is a max pooling layer which has a windows size of 3 and stride 2. These parameters are also user defined for each pooling layer. These two operations result in an output size of 55. This operation also follows the same formula.

$$\frac{W-(F-1)+2P}{S}+1$$
$$\frac{110-(3-1)}{2}+1$$
$$55$$

The next 4 layers follow this same scheme with convolutional layers and pooling layers. The user specified parameters are the same as for layer 1 with the output size being derived from the parameters of the convolutional and pooling components. These derived sizes use the same formula so will be omitted here. The output of the final layer is derived at size 6 and the number of channels is user specified at 256. The following 2 layers are fully connected with user specified 4096 channels. The size of the output layer is also user specified as the number of classes in the problem.

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

Response:

The authors' experiments show that small variations in rotation, occlusion, pose, etc. can result in worse feature map expressivity and worse classification (figures 5, 7, and eight). This result demonstrates that the neural architecture used doesn't generalize well to test data which is much different than their training data. One solution to this is to include more training data which has high degree of within-class variability. Thus if a model trained on the entire imagenet, which has many examples of high variability relative to other common datasets, cannot generalize well, than a model trained on less data, with likely less variability, would logically be worse.

If a smaller dataset is used and has overall low variability then the model may score well on it's test data but would not generalize well to unseen data of the same classes but of different

orientations and backgrounds.

This does not necessarily mean that small datasets have no value as they can be useful for classification of classes which are not represented within larger datasets. Small datasets can often be used to finetune models trained on larger datasets to be specific to new classes while retaining the representational power of the larger dataset. Besides, when solving real-world problems, we can use small dataset to do sanity check. If we have a candidate architecture that does not perform well on overfitting, then this architecture may not be suitable for solving the problem, and we should consider other architectures; otherwise, we can proceed to train with larger dataset.

7 Question 10 1 / 1

✓ - 0 pts Correct

gradescope