CS7150.37338.202130 Day 3 Paper Questions

Xiaolin Zhao, Aishwarya Vantipuli, Morgan Kohler

TOTAL POINTS

4.5 / 5

QUESTION 1

1 Question 1 **1 / 1**

✓ **- 0 pts** Correct

- 0.5 pts Minor inaccuracies

- 1 pts Major inaccuracies or not clear answer

- 0 pts Click here to replace this description.

Very good! Especially the contrast with the past "engineered" methods!

QUESTION 2

2 Question 2 **1 / 1**

✓ **- 0 pts** Correct

- 1 pts Not clear description and missing key point in the analysis.

- 0.5 pts Minor inaccuracy

Good answer!

QUESTION 3

3 Question 6 **1 / 1**

✓ **- 0 pts** Correct

- 0.5 pts Some inaccuracies

Ok answer to regularization, but careful with the conclusions from the paper, those are arguments of an heated debate!

QUESTION 4

4 Question 7 **1 / 1**

✓ **- 0 pts** Correct

- 0.5 pts Minor inaccuracies or not clear answer

- 1 pts Click here to replace this description.

Great job!

QUESTION 5

5 Question 9 **0.5 / 1**

- 0 pts correct

- 0.2 pts Some inaccuracies in "Memorize"

- 0.2 pts Some inaccuracy in "Shatter"

✓ **- 0.5 pts** Some inaccuracies

- 0.8 pts Major inaccuracies

CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 3 — Preparation Questions For Class
Due: Wednesday 1/27/2021 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 'Deep Learning' (Three Giants Paper)

- Read the whole paper

Question 1. *What is representation learning? In what sense are deep learning methods considered to be representation learning?*

Response:

Representation learning refers to a set of techniques used to learn features and other representations from raw data to predict classification/regression task. Deep learning methods are similar to representation learning because of their architecture with multiple levels/layers and each layer is considered as a method to perform a sub task .These layers are composed of non linear functions to extract relevant features. Starting with raw input data, when these layers are placed into a sequential pipeline, the representations learned at each layer are much more complex than the previous layers.As layers go deeper, so does the transformations such that very complex features can be learned. One important thing to notice about deep learning methods is that these layer transformations are not designed by humans but learned by the model itself from raw data using a general purpose learning procedures.

Question 2. *Explain Figure 1a. In your explanation, explain how it connects to the claim that 'The hidden layers can be seen as distorting the input in a non-linear way so that categories become linearly separable by the last layer'. See the preamble above for comments on how to explain a figure.*

Response:

Context:

Author tries to illustrate how a simple Feed Forward Neural Network works in a binary classification setting with one hidden layer.

1 Question 1 **1 / 1**

- ✓ **- 0 pts** Correct
- **- 0.5 pts** Minor inaccuracies
- **- 1 pts** Major inaccuracies or not clear answer
- **- 0 pts** Click here to replace this description.

💬 Very good! Especially the contrast with the past "engineered" methods!

CS 7150: Deep Learning — Spring 2021 — Paul Hand

Day 3 — Preparation Questions For Class
Due: Wednesday 1/27/2021 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 'Deep Learning' (Three Giants Paper)

- Read the whole paper

Question 1. *What is representation learning? In what sense are deep learning methods considered to be representation learning?*

Response:

Representation learning refers to a set of techniques used to learn features and other representations from raw data to predict classification/regression task. Deep learning methods are similar to representation learning because of their architecture with multiple levels/layers and each layer is considered as a method to perform a sub task .These layers are composed of non linear functions to extract relevant features. Starting with raw input data, when these layers are placed into a sequential pipeline, the representations learned at each layer are much more complex than the previous layers.As layers go deeper, so does the transformations such that very complex features can be learned. One important thing to notice about deep learning methods is that these layer transformations are not designed by humans but learned by the model itself from raw data using a general purpose learning procedures.

Question 2. *Explain Figure 1a. In your explanation, explain how it connects to the claim that 'The hidden layers can be seen as distorting the input in a non-linear way so that categories become linearly separable by the last layer'. See the preamble above for comments on how to explain a figure.*

Response:
Context:

Author tries to illustrate how a simple Feed Forward Neural Network works in a binary classification setting with one hidden layer.

What is plotted:

The neural network is used to classify two non linear classes (blue and red). This representation is then transformed to a binary output using the same sigmoid function. The figure represents three phases, one is the raw data, the other after passing it to the hidden layer with sigmoid function and finally predicting an output using sigmoid.

What we observe:

the figure illustrates that the raw data is clearly non-linearly separable considering the two classes having a convex boundaries. The hidden representation depicts a distorted version of the input where the two classes have been transformed to a coordinate space with a much clearer line of separability.

Interpretation:

The statement ”The hidden layers can be seen as distorting the input in a non-linear way so that categories become linearly separable by the last layer” relates to DARPA video where speaker mentions neural networks stretches and squashes the data space to make the non linear classes linearly separable by introducing a non-linear activation function (sigmoid). At the last layer, the sigmoid function uses a threshold/probability concept to determine what class each data point belongs to.

Question 3. *What is the selectivity-invariance dilemma and why is it challenging to overcome?*

Response:

The selectivity-invariance dilemma in machine learning refers to the process of selecting for characteristics necessary for classifying an object while ignoring irrelevant features. In the context of object recognition, the eyes, nose, and mouth of an animal are likely necessary for correct classification while the pose of the animal or objects in the background should not be considered. If the pose or background is considered than it could lead to false positives or false negatives if for example a different animal is in the same pose or the same animal is present in a different background. It is difficult problem to solve because it requires that an algorithm be sufficiently advanced to extract robust features from the input and be invariant to changes in pose, lighting, or noise. This level of advanced feature extraction is difficult to solve as it requires an algorithm sufficiently advanced to extract robust features which is difficult without hand selecting features. It also requires that the algorithm does not select for irrelevant common features such as the background which can game the loss function to decrease loss while not learning the features relevant to the true object in consideration. This problem can be mitigated by having a large and diverse data set but the problem will still remain in some capacity.

Question 4. *Explain Figure 4.*

Response:

Context

2 Question 2 1 / 1

✓ - **0 pts** Correct

- **1 pts** Not clear description and missing key point in the analysis.

- **0.5 pts** Minor inaccuracy

💬 Good answer!

networks and recurrent neural networks (RNNs) trained with reinforcement learning to attend to specific areas of the input image. In this way, the algorithm could "look" around the image to find the region of highest interest instead of treating the entire image equally. The third direction is that of NLP algorithms which can attend to different parts of the input instead of processing the entire sequence at once. This prediction has largely come true with the advent of attention models such as transformers. Each of these predictions are common in that they combine the representative learning of current deep learning models with higher level consideration of the input. The authors present the final direction of AI in general as the development of algorithms which can reason about the input in a more complex way than current rule-based methods and do so through autonomous representation learning. This represents a higher-level intelligence than current methods which can learn task-based representations but are unable to reason about such representations or which can reason using contextual rules but are unable to derive these rules themselves.

Directions:
Read 'Understanding deep learning requires rethinking generalization.'

- Read only: Abstract, Sections 1.0, 1.1, 2.0, 2.1, 6

Question 6. *What is regularization? What is the explanation of why it is 'needed' to obtain good generalization performance?*

Response:

Regularization is making the solution of a machine learning algorithm generalize well to unseen data, often by reducing the complexity of the function that model fits, or by de-coupling the parameters in the model. This can come in the form of implicit regularization where the algorithms architecture implicitly generalizes. It can also be explicit regularization where techniques are applied to tweak the algorithm towards a solution which generalizes well. If a machine learning algorithm is sufficiently complex to memorize an entire training dataset (assuming training and testing set are independent and identically distributed, we have 100% training accuracy but much lower testing accuracy) then the algorithm has over-fit to the solution and memorized irrelevant details of the input. To force the model to focus on the features which are relevant to a generalized solution, some form of regularization must be present in the solution. The authors prove that deep neural networks can memorize randomized input data and can thus over-fit to any presented solution. The striking nature of neural networks is how well they can provide generalizable solutions with no intervention to apply regularization to the model. This observation implies that there is an inherent form of implicit regularization present in the architecture of neural networks trained with SGD.

Question 7. *Explain Figure 1a.*

Response:

Context:

The paper seeks to understand the model capacity of a feed forward neural network. To do so, the authors trained deep learning models on true labels and on randomized labels. They

3 Question 6 **1 / 1**

✓ **- 0 pts** Correct

- 0.5 pts Some inaccuracies

Ok answer to regularization, but careful with the conclusions from the paper, those are arguments of an heated debate!

networks and recurrent neural networks (RNNs) trained with reinforcement learning to attend to specific areas of the input image. In this way, the algorithm could "look" around the image to find the region of highest interest instead of treating the entire image equally. The third direction is that of NLP algorithms which can attend to different parts of the input instead of processing the entire sequence at once. This prediction has largely come true with the advent of attention models such as transformers. Each of these predictions are common in that they combine the representative learning of current deep learning models with higher level consideration of the input. The authors present the final direction of AI in general as the development of algorithms which can reason about the input in a more complex way than current rule-based methods and do so through autonomous representation learning. This represents a higher-level intelligence than current methods which can learn task-based representations but are unable to reason about such representations or which can reason using contextual rules but are unable to derive these rules themselves.

Directions:
Read 'Understanding deep learning requires rethinking generalization.'

- Read only: Abstract, Sections 1.0, 1.1, 2.0, 2.1, 6

Question 6. *What is regularization? What is the explanation of why it is 'needed' to obtain good generalization performance?*

Response:

Regularization is making the solution of a machine learning algorithm generalize well to unseen data, often by reducing the complexity of the function that model fits, or by de-coupling the parameters in the model. This can come in the form of implicit regularization where the algorithms architecture implicitly generalizes. It can also be explicit regularization where techniques are applied to tweak the algorithm towards a solution which generalizes well. If a machine learning algorithm is sufficiently complex to memorize an entire training dataset (assuming training and testing set are independent and identically distributed, we have 100% training accuracy but much lower testing accuracy) then the algorithm has over-fit to the solution and memorized irrelevant details of the input. To force the model to focus on the features which are relevant to a generalized solution, some form of regularization must be present in the solution. The authors prove that deep neural networks can memorize randomized input data and can thus over-fit to any presented solution. The striking nature of neural networks is how well they can provide generalizable solutions with no intervention to apply regularization to the model. This observation implies that there is an inherent form of implicit regularization present in the architecture of neural networks trained with SGD.

Question 7. *Explain Figure 1a.*

Response:

Context:

The paper seeks to understand the model capacity of a feed forward neural network. To do so, the authors trained deep learning models on true labels and on randomized labels. They then plotted various properties trained models exhibit. The data sets used in the experiment are CIFAR10 and ImageNet. Authors tested the Inception V3 architecture on ImageNet and a smaller version of Inception, Alexnet and MLPs on CIFAR10.

What is plotted:

Figure 1a depicts the change in average training loss as step count increase. Experiments using true labels, random labels, uniform partially shuffled pixels, random partially shuffled pixels, and images generated only by Gaussian noise are performed and plotted.

What we observe:

We observe that each model converges to zero training loss no matter what data set it is trained on. We see that with shuffled pixels, random pixels and Gaussian noise the model takes a little longer time to converge. It is astonishing that model was able to perfectly fit with random labels even though the relationship between image and labels is completely destroyed.

Interpretation:

This tells us that the neural net used is able to completely memorize the input no matter how arbitrary the labels are. Each data set takes a different amount of time to fully converge with the true labels converging the fastest and randomly generated labels taking the longest. It is interesting that the model is able to learn labels associated with random Gaussian noise quicker than it can learn randomly assigned labels. It is plausible that the random loss associated with the Gaussian images lead the SGD down a uniform path to full memorization with no meaningfully generalized solutions. While with random labels, some of the images are likely to have some potential correlation to cause the algorithm to go down a certain path. However further images will immediately contradict this correlation and cause the solution to diverge and take longer.

From a generalization perspective, this experiment proves that a model can learn any arbitrary data set and should thus tend to over-fit and perform poorly on test data. The fact that neural networks still tend to implicitly regularize to a well generalized solution is remarkable.

Question 8. *Explain Figures 1bc.*

Response:

Context:

The goal of the paper is to understand the effective model capacity of feed-forward neural networks. Authors choose a methodology to train standard networks on both true data and on random labeled data where there is no relationship between instances and labels and observe the effects on training time and testing error.

What is plotted:

Figure bc shows how Inception, AlexNet and MLP 1x512 neural network architectures

4 Question 7 **1 / 1**

✓ **- 0 pts** Correct

- 0.5 pts Minor inaccuracies or not clear answer

- 1 pts Click here to replace this description.

Great job!

behaves when plotted between different levels of label noise against the time taken to overfit (fig b), and test error with zero train error (fig c).

What we observe:

We observe that in fig b, with the Inception model, there seems to be a steady increase in converging time as the label corruption ratio increases. The other two architectures are converging in less time as labels are getting noisier. In fig c, the test error steeply increased for all the models as true labels were replaced by random labels. It is interesting to observe that all the models started at different test error rates with MLP as highest at 0.5 although it is trained on true data, which means that shallow neural networks lack the ability to handle complicated problem such as image classification. The upper bound of testing error 0.9 indicates that the models are producing random labels (assuming the testing data are uniformly distributed over 10 classes we have) since a random guessing over 10 classes will have 1/10 accuracy.

Interpretation:

Networks takes a little bit longer and vary by a constant factor of time to fit to the data when its completely random labels (i.e. when there is no relationship between the label and Image) but still able to perfectly fit them to zero training error. In fig c, Network choice seems to make a big difference when it comes to reducing generalization error. Overall, it is fascinating how deep neural networks are able to fit perfectly even when we completely destroy the relationship between label and Image. This happens when a neural network starts to memorize the millions of Image label combinations it is fed to. On contrary to popular belief, neural networks not only learns but also may memorises the abstract relationship between Image and labels at relatively same time

Question 9. *In Section 6, the authors write: " The experiments we conducted emphasize that the effective capacity of several successful neural network architectures is large enough to shatter the training data. Consequently, these models are in principle rich enough to memorize the training data." What is meant by 'capacity of ... neural network architectures'? What is meant by 'shatter the training data'? What is meant by 'memorize the training data'?*

Response:

Capacity:

Capacity can be seen as the ability of model to discover a function taken from a family of functions, and can be measured by the number of training examples that the model could always fit correctly. Take experiment above for example, the training loss could reach zero for certain deep neural network model on the given dataset.

Shatter:

Shatter means, the training data with random labels can also achieve zero training error, which essentially means that the model is capable of fitting an arbitrarily complicated function that has enough expression ability to represent any possible combination in the data-label space. ✓

Memorize:

For random labels, we can train a deep neural network model to shatter the training data. Also, what we can see from the experiment is that given completely random pixels, the training error again could reach zero. These two experiments mean that even when there's no correlation between data and label, the models are still able to predict ”correct” label for the training samples. In other words, the model is memorizing which data is paired with which label for the entire training set.

5 Question 9 0.5 / 1

- **- 0 pts** correct
- **- 0.2 pts** Some inaccuracies in "Memorize"
- **- 0.2 pts** Some inaccuracy in "Shatter"
- ✓ **- 0.5 pts** **Some inaccuracies**
- **- 0.8 pts** Major inaccuracies