CS 1678/2078: Deep Learning

Second Exam: Review Guide

April 10, 2024

General info: The exam will include all lectures from March 18 to April 17 (inclusive). The format will include true/false questions, multiple choice questions, and short answer questions. You do not need to (and shouldn’t) bring calculators or other aids. You don’t need to memorize equations but should understand what they do at a high level. In addition to being tested on understanding the lectures by the instructor, you will also be tested on having understood some of the project presentations given by the students in CS 2078 (see example questions below). The exam will be on April 24, 2-3:15pm, in 358 Cathedral of Learning.

Concepts and algorithms to review:

1. Transformers
   1. What are some problems with using discrete symbols and WordNet to represent word meaning?

Synonyms depending on context like bass do i mean the fish or the instrument

Also words with double meaning in general.

Discrete symbols means that synonyms are mostly ignored

* 1. What is a one-hot representation for the word “hotel”?

0 0 0 0 0 0 0 1 0 0 0 0 0 0 0

* 1. What is the relationship between word context and word meaning?

A words meaning is context dependent like duck can be the animal or the action depending on context

* 1. What objective does Word2Vec use to learn word embedding vectors?

It uses the similarity of the word vectors to calculate probability of one given another. Its adjusted to maximize this probability

Uses objective funtion

* 1. How is attention formulated mathematically, in its most basic form?

Attention takes the weighted sum of encoder hidden states its output contains information from hidden tstates with high attention. As a formula it uses dot product sum\_i (e^(q \* k\_i) / sum\_j (e^q \* k\_j) ) \*v\_i

* 1. What are queries, keys and values?

Query is the text in a search bar keys would be video titles or descriptions and values would be the best matched videos

Attention treats Each words representation as a query, Attention operates on queries keys and values each being a seperate matrix

Queries and keys share dimensionality with each other but not necessarily v

* 1. What makes queries, keys and values differ even though they are based on the same input x?

They all represent different parts of x vectors to be used/emphasized

* 1. How many query matrices will be there be if we use 8-headed attention?

8

* 1. How is self-attention different than encoder-decoder attention we saw for RNNs?

In self attention queries keys and values are drawn from the same source in encoder decoder they are all seperate.

* 1. What are positional encodings/representations?

Concatenete position encoding p to each input vector when p = pos(j)

* 1. When might we need some attention scores to 0?

When a value is not considered

* 1. What attention do transformers use (self-attention, cross-attention, or both)?

Cross attention

* 1. What is stacking of attention?

It is a way of capturing a hierarchy of relationships within data

* 1. In what ways is computing attention expensive?

It adds more calculations to a task now needing to consider and apply weights to each input

* 1. What is pretraining?

Pretraining models is a way of initializing modedls. Parts of the input are hidden and the model is trained to reconstruct it

* 1. How do we use masking for pretraining?

We mask the hidden parts that need to be trained

* 1. What is BERT?

Bidirectional Encoder Representations from Tranformers its a popular pretrained transformer model

* 1. What task might decoders use for pretraining?

They use a language model predicting missing words. The goal is to be able to generate new sentances

* 1. What tokens are used for image transformers?

Smaller parts of the image

1. Self-supervised learning
   1. How is self-supervised learning useful, given that we don’t know the labels of categories? How is it usually applied in practice?

It generates labels automatically and is generaly good at supervised learning objectives like classification and regression. Not all categories need worded labels as long as they can be reasonably grouped

* 1. What are five pretext tasks we can use for self-supervised learning? What information that is already present in data (i.e. that doesn’t require human labels) can you “teach” a network to predict, such that it learns a useful visual representation?

Image completion, rotation prediction, “jigsaw puzzle” ie the image is split and the pieces are shuffled, colorization, missing pixels

* 1. Is supervised representation learning always an upper bound to unsupervised learning? If not, why not?

NO it depends on training samples size

* 1. How is the relative position task (predict location of a patch relative to another patch) similar to a language modeling (next word prediction) task? How is it different?

Its similar in function where you read the before and after and fill in the blank. Its different in being a 2d image where sentences are a line and that you know its a piece and are putting it in a place.

* 1. In what way is most learning in computer vision static, compared to human/child learning?

Children tend to learn by acting and moving in the world while computers only see.

* 1. What is equivariance in learning the relation between how an agent acts, and what it observes? How do we use it to learn representations?

Equivariance is predictably responsive to class transformations organizing it. If 2 things move in the same way they should be related.

* 1. What are positive and negative pairs in contrastive learning, like SimCLR?

It works well at predicting if 2 things are different sometimes to well whete 2 different cats can be seen as different which you may not want. Simclr requires large batch sizes

* 1. How is data augmentation used in SimCLR?

We use data augmentation to artificially increase the amount of data by showing the same thing in different ways.

1. Foundation models and prompting
   1. What is instruction tuning?

Its a collection of instuction and output pairs used to train the language model

* 1. Why is instruction tuning needed?

Instruction tuning makes it more human understandable and understand what we want as a response.

* 1. What are some problems with instruction tuning?

Hard to collect demonstration expensive. Mismatch between lm objective and human preference

* 1. What is reinforcement learning using human feedback (RLHF)?

This is maximizing for human preference for question answer pairs

* 1. What sort of feedback could humans give?

Humans can give feedback like what they would answer or how good an answer was

* 1. What is a reward model and how does it help with the expense of collecting human feedback?

The reward model is when the labeler ranks outputs best to worse and that data is used for training. It helps by not requiring a human to create the labels themselves just rank them.

This causes the problem where sounding right is more important than being right

* 1. What is in-context learning?

A task is specified by prepending examples of the task before the next example. No gradient updates are performed. The learning is done within the context of the question being asked

* 1. What is zero-shot learning/performance? What is few-shot learning/performance?

Few shot is in-context. zero shot is the same but with no examples.

* 1. What is chain-of-thought prompting?

Its a way of formatting inputs where you ask a example question and in the example answer you explain the chain of thought to get the result followed by a new question. Hopefully the new question will use the chain of thought from the example.

* 1. What is zero-shot vs manual/few-shot chain-of-thought?

Zero shot is examples and it just doing a chain of thought with step by step manual/few shot uses a few examples and is generally a bit better then zero shot

* 1. What is prefix tuning?

Adds a prefix of parameters and freezes all pretrained. The prefix is processed by the model like words. This allows each element of a batch at inference to runa a different tuned model

* 1. What is contrastive language-image pretraining (CLIP)?

Clip trains an image encoder and text encoder to predict pairings of a batch of training examples. During tests it uses zero-shot linear classifier by embedding the names or descriptions of l target classes

* 1. In what way is the data that CLIP uses noisy?

It uses massive datasets if image text pairs from the web. This data could for all we know not be releveant to each other so its inherently noisy.

* 1. How does object classification with CLIP work?

It computes the dot product between an image and a prompt for each class. The return class is the highest

* 1. How can we get a list of object descriptors from a language model, to improve object classification using CLIP?

Adding a – or a bullet point helps tell the model to give a list

* 1. What are learned prompts in computer vision?

Learned prompts are arbitrary visual prompts that are meant to prioritize certain parts of an image.

* 1. How can we use a frozen language model for vision-language tasks?

Frozen language models can be used to ensure to much of the model is not changed when run.

* 1. How can we automatically collect input-output pairs for visual instruction tuning?

Web scraping is an option like just making a bot in reddit with prompts.

Another choice is using other already made ai and training on them.

1. Reinforcement learning
   1. What needs to be defined, to set up a problem as a reinforcement learning task?

The objective the current state of the game the potential actions and the reward

* 1. What is Q-learning? What objectives do we formulate for our neural network, that we minimize/maximize at training time?

Q learning is solving for the optimal policy. It uses a Q function.. The goal is find the optimal policy u or maximizeing the expected sum of rewards

* 1. How are Q-learning methods different than policy gradient methods? What are some advantages of policy gradients as opposed to Q-learning?

Q leaning is more complicated while policy gradients are just a form of gradient descent testing a set of inputs and outputs.

1. Image and video generation
   1. What are autoencoders?

An unsupervised approach for learning a lower dimensional feature representation from unlabeled training data. Training features such that they can reconstruct original data. Encoding itself

* 1. What are variational auto-encoders (VAEs)?

They are auto encoders with a probabilistic spin. It adds an additional encoder network that approximates p\_theta(x|z). THis makes it far faster than the other method but also less accurate

* 1. What are generative adversarial networks (GANs)?

THese are networks that are trained by trying to fool others unto thinking its real. The generator network makes an image look as real as possible to fool the discriminator netork that tries to determine if its real or fake.

* 1. What are the two competing agents in a GAN, and what is the training objective?

Generator network and discriminator network . The generator wants to trick the discriminator the discriminator wants to distinguish real vs fake. The goal is to make the best generator network.

* 1. What are the two inverse processes we model in a diffusion model?

Diffusion and Denoising both markov chain based both are very expensive.

1. Bias, domain shifts, attacks
   1. What are domains and domain shift?

Domains are styles are an object. A domain shift is looking at lets say a cartoon mailbox vs a real one the cartoony artstyle is different then realistic. Another example is location something in the US may look different in Asia so going between the 2 is a domain shift.

* 1. Why are domain shifts problematic?

They are problematic because they are difficult to link together. Humans can even struggle with it not understanding what something in another country is. The computer can think that an asian mailbox vs a us mailbox are completely different since they look different despite being functionally the same.

* 1. What are examples of domains?

Domains can be nearly anything. Artstyles can be domains. Countries can be domains. Urban vs suburban vs rural can be domains.

* 1. What are some ways to align feature distributions across domains?

Firstly gathering data across domains is a good solution. Like when training to detect apples instead of only showing granny smiths showing that and golden delicious and other apple breeds.

* 1. What is an example of bias in language? Bias in vision?

Bias in language is words generally being attached to a meaning that they should not be. The most common variation is attaching gender to things like jobs. Like nurses being female. Bias in vision is the same but for images. When we see bananas we dont imagine them as green or spoiled but a spoiled banana is still a banana.

* 1. How might generative models negatively impact society?

One could deepfake someone saying/doing something offensive and try to ruin them socially or there credibility. Even when proven to be fake it generally leaves an effect. Especially considering that not all people who see the faked video will find out that its fake.

* 1. What are adversarial attacks?

Adversarial attacks are attacks targeting vulnerabilities of ai algorithms meant to decieve ai systems. They are generally imperceptible to people but cause ai to be inaccurate.

1. Course projects presented by CS 2078
   1. What was the motivation?
   2. What was the approach?
   3. In what ways was the project successful?
   4. What are some limitations or failures of the project?