CS 182/282A "Deep Learning"
Spring 2022 Marvin Zhang

Midterm 2

- Do not open the exam before you are instructed to do so.
- **Electronic devices should be turned off for the entire exam duration**, including cell phones, tablets, headphones, and laptops. Turn your cell phone off, or risk getting a zero on the exam.
- The exam is closed book, closed notes except your one-page cheat sheet. You are allowed one double-sided 8.5x11 inch cheatsheet.
- You have 1 hour and 50 minutes (unless you are in the DSP program and have an allowance of 150% or 200% time).
- Please write your initials at the top right of each page after this one (e.g., write "JD" if you are John Doe). Finish this by the end of your 1 hour and 50 minutes.
- Mark your answers on the exam itself in the space provided. Do **not** attach any extra sheets.
- For the questions in Section 2, you should mark true or false for each statement. For each question, there may be more than one true option, but there is always at least one true option. Note that there is a penalty for marking the incorrect option, but there is no penalty for not marking either option.

First name	
Last name	
SID	
First and last name of student to your left	
First and last name of student to your right	

1 Multiple Choice (Single Answer)

1. (1 point) Suppose that when performing attention, we have the following keys and values:

   Keys:

   $$\left\{\begin{bmatrix}-3\\0\\1\end{bmatrix}, \begin{bmatrix}1\\1\\-1\end{bmatrix}, \begin{bmatrix}0\\1\\1\end{bmatrix}, \begin{bmatrix}1\\0\\0\end{bmatrix}\right\}$$

   Values:

   $$\left\{\begin{bmatrix}2\\1\\1\end{bmatrix}, \begin{bmatrix}1\\2\\2\end{bmatrix}, \begin{bmatrix}0\\3\\1\end{bmatrix}, \begin{bmatrix}-1\\0\\2\end{bmatrix}\right\}$$

   We want to compute the attention embedding using these keys and values for the following query:

   $$\begin{bmatrix}2\\1\\-1\end{bmatrix}$$

   Which of the following is the correct attention embedding? **To simplify calculations, replace `softmax` with `argmax`. For example, `softmax`$([-1, 1, 0])$ would instead be `argmax`$([-1, 1, 0]) = [0, 1, 0]$.**

   ○ $\begin{bmatrix}2\\1\\1\end{bmatrix}$

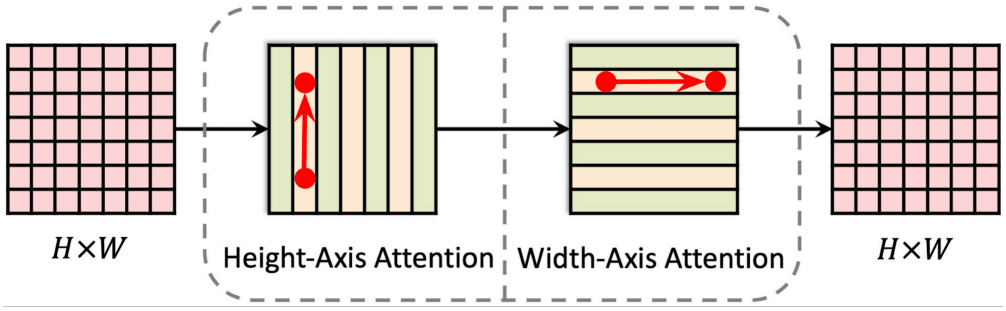
   ○ $\begin{bmatrix}1\\2\\2\end{bmatrix}$

   ○ $\begin{bmatrix}0\\3\\1\end{bmatrix}$

   ○ $\begin{bmatrix}-3\\0\\1\end{bmatrix}$

   Solution: B

2. (1 point) Vision transformers are computationally expensive for high resolution images. One idea of reducing the amount of computation is to use axial self-attention instead of normal self-attention. The idea of axial attention is very straightforward: for each patch, we only allow the network to attend to other patches in the same row (width axis) or same column (height axis). We interleave the row attention and column attention layers to obtain an axial vision transformer architecture. The architecture can be visualized in the following diagram.



Now suppose our input image is HxW. What is the computational complexity of our axial vision transformer?

- ◯ $O((HW)^2)$
- ◯ $O(H^2W + W^2H)$
- ◯ $O((H + W)^2)$
- ◯ $O(HW)$

Solution: B

3. (1 point) Which statement is true about softmax temperature scaling for model calibration?
   - ◯ The temperature is usually a parameter we learn during training.
   - ◯ As the temperature goes to zero, our model becomes equally confident across all classes.
   - ◯ The temperature is usually tuned to maximize log likelihood on a validation set.
   - ◯ Since the temperature changes the model output probabilities, poorly learning/tuning the temperature can affect the accuracy of the model.

Solution: C

4. (1 point) Which of the following is the correct formula for an importance sampled weighted objective? Let θ be the parameters of the model we are optimizing, X be the random variable for our inputs, and Y be the random variable for their labels.
   - ◯ $\mathbb{E}_{\text{source}}\left[\frac{p_{\text{target}}(X,Y)}{p_{\text{source}}(X,Y)}\ell(\theta; X, Y)\right]$
   - ◯ $\mathbb{E}_{\text{target}}\left[\frac{p_{\text{target}}(X,Y)}{p_{\text{source}}(X,Y)}\ell(\theta; X, Y)\right]$
   - ◯ $\mathbb{E}_{\text{source}}\left[\frac{p_{\text{source}}(X,Y)}{p_{\text{target}}(X,Y)}\ell(\theta; X, Y)\right]$

○ $\mathbb{E}_{\text{target}}\left[\frac{p_{\text{source}}(X,Y)}{p_{\text{target}}(X,Y)}\ell(\theta; X, Y)\right]$

Solution: A

5. (1 point) You are training a neural net to be used for a robotics application. During training, the robot imagines goal images of different configurations of objects in the scene before it and practices rearranging the objects to achieve the goal. You have a dataset of images of the scene, which you will use to train a generative model to produce these goal images.

   Your design must have the following capabilities:

   - Generate reasonable goal images in the robot's environment.
   - Generate quickly so that generation time does not slow down the robot's training.
   - Encode images of the current scene into a latent space (to check if the current scene is close to the goal)

   Which generative model is preferred for this application?

   ○ Standard GAN
   ○ VAE
   ○ Autoregressive model
   ○ Diffusion model

   Solution: B

6. (1 point) You are trying to fine-tune a model trained with self-supervised learning for use on a supervised downstream task. You are deciding between two options:

   Linear probing: keep the pretrained weights frozen, only train the final linear layer
   Full fine-tuning: fine-tune all parameters

   When might you prefer linear probing over full fine-tuning?

   ○ You want the model to be robust to adversarial examples.
   ○ Your pretrained model is small.
   ○ Your supervised dataset is small.
   ○ Your pretrained model was trained on data from a very different distribution.

   Solution: C

7. (1 point) What is the time complexity of using beam search to generate a sequence of length T, if we use a beam width of K? Treat the vocabulary size and run time for each forward pass of our model as constants. Beam width refers to the number of candidate sequences we maintain.

   ○ $O(K^2T)$
   ○ $O(K^T)$

◯ $O(T^K)$

◯ $O(KT)$

Solution: D

8. (1 point) Which of the following statements best characterizes the difference between meta-learning and multi-task learning?

   ◯ In meta-learning, you typically only have one task available for training, whereas multi-task learning assumes many training tasks.

   ◯ Multi-task learning is capable of handling novel classes at test time, whereas meta-learning is not.

   ◯ Meta-learning aims to use the training tasks to learn how to quickly solve a new task, rather than just aiming to perform well on the training tasks themselves.

   ◯ Meta-learning has not yet been shown to be effective when combined with deep neural networks, whereas multi-task learning has.

   Solution: C

2 Multiple Choice (True False)

Fill in either true or false for all statements: there may be more than one true option, but there is always at least one true option. Each question is worth two points, and each statement is worth 0.5 points. Marking the incorrect option will result in a penalty of -0.5 points, whereas not choosing an option does not have a penalty. Your score for each question will be ReLU(points).

9. (2 points) Mark true for all of the following statements that are correct about LSTMs and vanilla RNNs, and false otherwise.

    ◯ T ◯ F : RNNs use backpropagation through time to compute gradients because a forward pass involves a time component.

    ◯ T ◯ F : Using gating functions in LSTMs help prevent vanishing gradients.

    ◯ T ◯ F : Gradient clipping cannot be used to remedy the exploding gradient problem in vanilla RNNs.

    ◯ T ◯ F : LSTMs can model long-range dependencies that vanilla RNNs cannot.

    Solution: A and B are true

10. (2 points) Consider a modified LSTM with an additional layer applied to the cell state output:

```
# Standard LSTM equation. Assume f is your forget gate,
# i is your input gate, prev_c is the past cell state,
# and g is the content you are adding to the cell state.
next_c = f * prev_c + i * g

# Modified LSTM equation, introducing additional linear layer W_new
modified_next_c = np.dot(f * prev_c + i * g, W_new)
```

    Mark true for all of the properties you expect this modified LSTM to have, compared to a standard LSTM, and false otherwise.

    ◯ T ◯ F : The content in the cell state would change more rapidly.

    ◯ T ◯ F : The network is more likely to encounter vanishing and exploding gradients.

    ◯ T ◯ F : The function classes are equally expressive - i.e. every function that can be represented using a standard LSTM can also be represented with a modified one, and vice versa.

    ◯ T ◯ F : In both the original and modified LSTM, processing one sentence takes $O(M^2)$ computations, where M is the input sequence length.

    Solution: A and B are true.

11. (2 points) Mark true for all of the following statements that are correct about transformers, and false otherwise.

    ◯ T ◯ F : Unlike with RNNs, the amount of learnable parameters in a transformer scales with the maximum sequence length of inputs it is trained on.

    ◯ T ◯ F : If we remove all of the feedforward layers in a standard transformer, each output of our model at each timestep is a linear combination of the inputs.

    ◯ T ◯ F : Without positional encodings, if you permute the input sequence to a transformer encoder, the resulting output sequence will be the output sequence of the original input, except permuted in the same way.

    ◯ T ◯ F : In a single multi-head attention layer, the operations for each head can be run in parallel to the other heads (e.g. the operations for one head do not depend on the others).

    Solution: B, C and D are true.

12. (2 points) Mark true for all of the following statements that are correct about the difference between transformers at training time and inference time, and false otherwise.

    ◯ T ◯ F : For a decoder-only transformer, at training time, you can predict an entire target sequence with one forward pass, whereas at inference time you must predict tokens one by one.

    ◯ T ◯ F : Beam search is used both during training and inference.

    ◯ T ◯ F : At training time, layer norm statistics are computed separately on each mini-batch, but at test time you use fixed values.

    ◯ T ◯ F : At training time, the runtime scales linearly in the output sequence length, but at inference time it scales quadratically.

    Solution: A is true.

13. (2 points) You want to build a text classifier that takes in a sequence of text and outputs its sentiment. You want to use a pretrained model to extract a fixed-size representation from the input text sequence, and train a linear classifier on this representation. You would like this representation to capture information from the entire input sequence. For an input sequence with 10 tokens (including start/stop tokens), mark true for the following representations which could satisfy this property, and false otherwise.

    ◯ T ◯ F : The 5th embedding in the final layer of embeddings of BERT.

    ◯ T ◯ F : The 10th embedding in the second layer of embeddings of BERT.

    ◯ T ◯ F : The 5th embedding in the final layer of embeddings of GPT-3.

    ◯ T ◯ F : The 10th embedding in the second layer of embeddings of GPT-3.

    Solution: A, B, and D are true.

14. (2 points) Mark true for all of the following statements that are correct about language models, and false otherwise.

    ◯ T ◯ F : GPT-3 is a decoder-only transformer.

    ◯ T ◯ F : BERT's training objective is to predict the next token in a sentence.

    ◯ T ◯ F : The final token embeddings from BERT are context-dependent, meaning the same token could have a different embedding depending on the other tokens in the sequence.

    ◯ T ◯ F : ELMo is a transformer with an encoder and decoder.

    Solution: A and C are true.

15. (2 points) Mark true for all of the following statements that are correct about GPT-3, BERT, and ViT, and false otherwise.

    ◯ T ◯ F : Both GPT-3 and BERT can be directly used for generation tasks, e.g., generating a novel.

    ◯ T ◯ F : GPT-3 uses a stack of transformer decoders while BERT uses a stack of transformer encoders.

    ◯ T ◯ F : Positional embeddings are not typically used in ViT, but they are necessary for text data.

    ◯ T ◯ F : For ViT, for the same size input image, using smaller image patches requires higher memory usage.

    Solution: B and D are true.

16. (2 points) Suppose we want to extend our vision transformer from image to videos. One simple method is to split each frame of the video into a sequence of patches and treat the patches in all the frames as a sequence. Mark true for all of the following statements that are correct about this method, and false otherwise.

    ◯ T ◯ F : The computation complexity for our video transformer scales linearly with the length (number of frames) of the video.

    ◯ T ◯ F : We may choose to use either learned or fixed (such as sin/cos) positional embeddings in our model.

    ◯ T ◯ F : For the same video, if we keep the hidden dimension of our video transformer the same, enlarging the patch size would increase the computation cost.

    ◯ T ◯ F : We can reduce the computational cost of our model if we stack up K frames in the channel dimension and split the K-frame tensor into patches instead of splitting each frame into patches

    Solution: B and D are true.

17. (2 points) Imagine you have a poorly calibrated language model, and you want to improve the calibration after training by dividing the logits by a positive temperature scalar. Mark true for all of the following statements that are correct about how the model may change, and false otherwise.

    ◯ T ◯ F : The sentence chosen if you select model outputs using beam search may change.

    ◯ T ◯ F : If you rank the probabilities of next tokens at a particular timestep, the rank order may change.

    ◯ T ◯ F : After the change, it will be harder to create targeted adversarial attacks.

    ◯ T ◯ F : If the temperature scalar is large, then if you sample a list of several sentences from the language model, the list is more likely to contain unusual sentences.

    Solution: A and D are true.

18. (2 points) Mark true for all of the following statements that are correct about machine translation, and false otherwise.

    ◯ T ◯ F : If we use the transformer architecture for machine translation, since we already have positional encodings in the transformer encoder input and the transformer encoder's output is fed into the transformer decoder, it is not necessary to use additional positional encodings for the transformer decoder.

    ◯ T ◯ F : If we encounter a rare word in the dataset, it is often good practice to replace it with an unknown <UNK> token.

    ◯ T ◯ F : If we use the byte-pair encoding algorithm for sub-word tokenization, all of the single letters in our original vocabulary will appear in our final set of tokens.

    ◯ T ◯ F : One advantage of using a transformer over an LSTM for machine translation is that during test time, it can generate many words at the same time in parallel, whereas LSTMs can only generate one word at a time.

    Solution: C is true.

19. (2 points) Mark true for all of the following statements that are correct about data augmentation, and false otherwise.

    ◯ T ◯ F : It can improve robustness to distribution shifts.

    ◯ T ◯ F : For a classification task, it can improve accuracy on the validation set.

    ◯ T ◯ F : Some augmentations, such as CutMix, can improve robustness to adversarial attacks.

    ◯ T ◯ F : Image augmentations are typically chosen carefully so as to match the distribution of data points which will be seen at test time.

    Solution: A, B, and C are true.

20. (2 points) Mark true for all of the following statements that are correct about domain adaptation, subpopulation shift, and domain generalization, and false otherwise.

◯ T ◯ F : Methods developed for any one of the frameworks are unlikely to work in any of the other frameworks.

◯ T ◯ F : In domain generalization, we are trying to account for domains for which we have no training data.

◯ T ◯ F : Upsampling rare domains and/or downsampling common domains is often an effective way to address subpopulation shift.

◯ T ◯ F : If we learn invariant features between one source and one target domain, our invariant features will work on other target domains.

Solution: B and C are true.

21. (2 points) Once again, mark true for all of the following statements that are correct about domain adaptation, subpopulation shift, and domain generalization, and false otherwise.

◯ T ◯ F : In subpopulation shift, we have multiple domains in our training data and want to generalize to completely new domains at test time.

◯ T ◯ F : Importance weighting is typically impractical in deep learning because estimating distributions over high-dimensional data is difficult.

◯ T ◯ F : In domain generalization, at training time we have a large amount of data from a source domain and a small amount of data from a single target domain that we wish to generalize to at test time.

◯ T ◯ F : Learning domain invariant features is typically impractical in deep learning because deep learning tends to overfit to one domain, making learning features that are invariant across domains difficult.

Solution: B is true.

22. (2 points) Mark true for all of the following statements that are correct about distribution shift datasets, and false otherwise.

◯ T ◯ F : ImageNet challenge datasets contain image classes that are different from the ones in ImageNet. Thus, they are a good stress test of whether models trained on ImageNet can also generalize to different classes.

◯ T ◯ F : WILDS aims to curate a suite of problems that faithfully represent how distribution shift manifests in real world applications, e.g. shifts caused by deploying models into different countries or different hospitals.

◯ T ◯ F : The ImageNet challenge datasets come in many styles of images (e.g. realistic, paintings, sketches).

◯ T ◯ F : The ANLI dataset consists of adversarial examples for natural language inference that were generated through a GAN.

Solution: B and C are true.

23. (2 points) Mark true for all of the following statements that are correct about uncertainty estimation, and false otherwise.

◯ T ◯ F : Let $\hat{y}$ be the class prediction for $\mathbf{x}$ and let $\hat{p}(\hat{y} \mid \mathbf{x})$ be its associated confidence. Let y be the true label for $\mathbf{x}$. If a model is calibrated, $\mathbb{P}(\hat{y} = y \mid \hat{p}(\hat{y} \mid \mathbf{x})) = \hat{p}(\hat{y} \mid \mathbf{x})$. If the model is usually overconfident, then $\mathbb{P}(\hat{y} = y \mid \hat{p}(\hat{y} \mid \mathbf{x}))$ is likely to be *less than* $\hat{p}(\hat{y} \mid \mathbf{x})$.

◯ T ◯ F : Changing the softmax temperature to different positive values in order to calibrate models can increase accuracy.

◯ T ◯ F : An ecologist wants to estimate the frequencies of known invasive species in a river stream. The ecologist is choosing between two computer vision models with equal accuracy, one which is more calibrated according to a held out part of the training set, and one which is better at anomaly detection. The classification decisions will directly affect the running tally of the species' frequencies. The river also contains many novel species which do not appear in the models' training data. Knowing all of this, the second model (which is better at anomaly detection) is preferable.

◯ T ◯ F : The maximum softmax probability anomaly detector for a typical ResNet-50 can be used detect untargeted adversarial examples with high performance.

Solution: A and C are true.

24. (2 points) Mark true for all of the following statements that are correct about adversarial examples, and false otherwise.

◯ T ◯ F : If inputs are always blurred during preprocessing before they are passed into a neural network, the network cannot have adversarial examples.

◯ T ◯ F : Training a network to be robust to adversarial examples currently often hurts performance on clean examples.

◯ T ◯ F : Adversarial examples are only found in convolutional networks.

◯ T ◯ F : For real-world applications, adversarial training is most useful if you have a realistic model of the modifications an attacker can make to inputs.

Solution: B and D are true.

25. (2 points) Suppose $\mathbf{x}$ is our original input, y is its original label, θ is our model parameters, and $\mathcal{L}$ is the cross entropy loss. Let $l(\mathbf{x})_i$ be the logit (pre-softmax input) from our model at dimension i, given input $\mathbf{x}$.

Consider a "MultiTargeted" attack which targets each wrong class out of K possible classes:
For $t \in \{1, \ldots, y-1, y+1, \ldots, K\}$
$\mathcal{L}(\mathbf{x}, y; \theta) := l(\mathbf{x})_y - l(\mathbf{x})_t$
$\mathbf{x}_{\text{adv,t}} = O(t)$
$\mathbf{x}_{\text{adv}}$ is set to be whichever $\mathbf{x}_{\text{adv,t}}$ incurs the greatest loss.

Mark true for all of the following statements that are correct about this attack, and false otherwise.

◯ T ◯ F : We should set $O(t) = \text{argmin}_{\delta:\|\delta\|_p \leq \varepsilon} \mathcal{L}(\mathbf{x} + \delta, t; \theta)$.

◯ T ◯ F : We should set $O(t) = \text{argmax}_{\delta:\|\delta\|_p \leq \varepsilon} \mathcal{L}(\mathbf{x} + \delta, t; \theta)$.

◯ T ◯ F : This attack can be at least $K - 1$ times as expensive as an untargeted attack.

◯ T ◯ F : The MultiTargeted attack has some of the drawbacks of untargeted attacks, e.g., it can cause models to misclassify examples as similar classes.

Solution: A, C, and D are true.

26. (2 points) Suppose we use adversarial training to train our network. The adversarial examples are generated through the fast gradient sign method (FGSM) and are within an ε distance (according to the ℓ_∞ norm) from our original points. Mark true for all of the following statements that are correct about the resulting model, and false otherwise.

◯ T ◯ F : The network is certified to be robust against all adversarial examples that are less than an ε distance (according to the ℓ_∞ norm) from the original training points.

◯ T ◯ F : Compared to a classifier trained without adversarial training, our classifier will likely have a higher accuracy on non-adversarial examples.

◯ T ◯ F : Compared to a classifier trained without adversarial training, our classifier will likely have a higher accuracy on adversarial examples generated via FGSM.

◯ T ◯ F : The classifier is likely to be more robust to adversarial examples generated from FGSM compared to adversarial examples generated through a different method, e.g. PGD.

Solution: C and D are true.

27. (2 points) Mark true for all of the following statements that are correct about adversarial examples and adversarial defense, and false otherwise.

◯ T ◯ F : Suppose we have set up a normal training process and trained a model. In order to defend against adversarial examples, it suffices to first generate a set of adversarial examples from our already trained model, add them to the dataset we have, and then train a new model from scratch using the same training process we already have.

◯ T ◯ F : When generating adversarial examples to attack a model, if we cannot have access to that model directly, we can train a similar model with the same dataset and generate adversarial examples on this new model. The adversarial examples can sometimes be effective against the original model.

◯ T ◯ F : Suppose we want to generate an untargeted adversarial example for a given model by maximizing the loss and we want to constrain the norm of our perturbation to be less than ε. If we constrain the perturbation's ℓ_∞ norm to be less than ε, we can typically generate an stronger adversarial example that induces greater loss compared to the adversarial example we can generate if we constraint the perturbation's ℓ_2 norm to be less than ε.

◯ T ◯ F : Suppose we first train our model and then make our model non-differentiable by finely quantizing the activations for each layer without changing the model's prediction. Now since we cannot take gradients with respect to the model input, our model is guaranteed to be safe against any adversarial examples.

Solution: B and C are true.

28. (2 points) Mark true for all of the following models which are typically evaluated via negative log likelihood on an unlabeled held out dataset, and false otherwise.

◯ T ◯ F : GANs

◯ T ◯ F : VAEs

◯ T ◯ F : Autoregressive models

◯ T ◯ F : BERT

Solution: B and C are true.

29. (2 points) Mark true for all of the following statements that are correct about Contrastive Predictive Coding (CPC) and SimCLR, and false otherwise.

◯ T ◯ F : Because CPC requires image augmentations, it is only intended for image representation learning.

◯ T ◯ F : Because the SimCLR framework requires image augmentations, it is only intended for image representation learning.

◯ T ◯ F : CPC requires labeled data for learning representations.

◯ T ◯ F : SimCLR requires labeled data for learning representations.

Solution: B is true.

30. (2 points) Mark true for all of the following statements that are correct about masked autoencoders (MAEs), and false otherwise.

◯ T ◯ F : The encoder takes only unmasked patches as input which can reduce memory usage and training cost.

◯ T ◯ F : The decoder input consists of mask tokens and embeddings of unmasked images patches.

◯ T ◯ F : A high masking ratio around 75% hurts model performance.

◯ T ◯ F : MAEs use causal attention masks.

Solution: A and B are true.

31. (2 points) Consider the following plots related to scaling laws, taken from the scaling laws (left) and GPT-3 (right) papers.

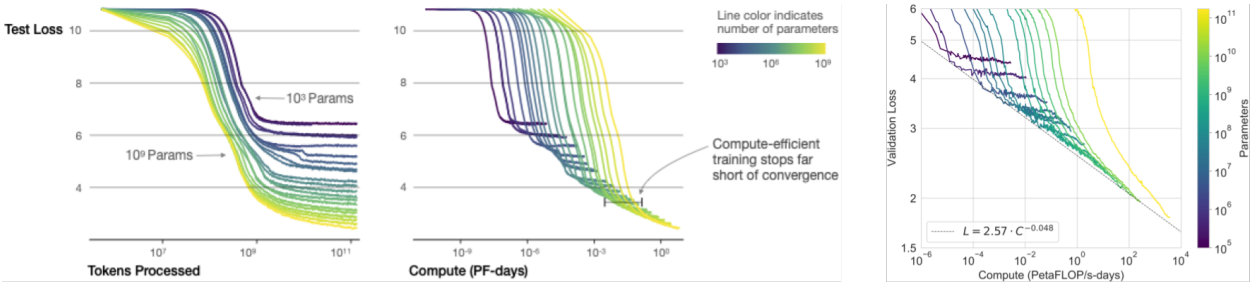


Figure 1: Scaling law plots.

Mark true for all of the following statements that are correct based on these plots, and false otherwise.

◯ T ◯ F : In both sets of plots, the first batch of the largest model uses more compute than the entire training run of the smallest model.

◯ T ◯ F : In the right plot, the loss lower bound is $L = 2.57 \cdot C^{-0.048}$. If the 2.57 were replaced by 3, the slope of the line representing L would change.

◯ T ◯ F : Some of the models in these plots are showing signs of overfitting.

◯ T ◯ F : The left plot shows that larger models require fewer training tokens to reach the same performance as smaller models.

Solution: A and D are true.

32. (2 points) Mark true for the two most typical examples (out of the examples below) of distribution shift in reinforcement learning, and false otherwise.

◯ T ◯ F : In behavioral cloning, the distribution of states or observations visited by the policy may drift away from the demonstration state distribution due to compounding errors in the policy.

◯ T ◯ F : In Q-learning, searching for actions which maximize the learned Q-function is analogous to the distribution shift caused by adversarial examples.

◯ T ◯ F : For robotic grasping tasks, humans may have difficulty providing the ground truth labels that correspond to successful grasps.

◯ T ◯ F : Supervision in reinforcement learning often is first specified in the form of abstract goals, rather than concrete ground truth labels, which then have to be translated into a reward function.

Solution: A and B are true.