CIS 620, Advanced Topics in Deep Learning, Spring 2024

Post-Quiz Week 2
Due: Sunday, February 4, 11:59pm
Submit to Gradescope

This 'quiz' should take less than 10 minutes; the goal is just to get you to remember what you read or discussed in class. (but look up things you don't know) Answers can each be a couple of sentences.

1. What should we do to make this class better?

Ans: I personally feel that discussing the technical aspects of the paper during class will be very beneficial but the pre-quiz really gives me a sense of what I am expected to learn before class, so it is better now compared to the first few classes.

Maybe an overview of what will be covered in class would also be nice. I feel like I am putting so much effort learning everything the hard way when I can just come to class and listen and learn the easy way and I can use my time outside class to do other work.

2. How would Yolo's performance be on a semantic segmentation task? Think about the grid size, speed of training and inference, and accuracy of the output.

Ans: YOLO, primarily being an object detection framework focussed on speed and real-time performance, might not be optimal for semantic segmentation tasks. It is designed for detecting objects in a grid and classifying them, which is different from the pixel-wise classification required in semantic segmentation.

3. What are *invariance* and *equivariance* (and how are they different)? Where have you seen the implementation of translational equivariance, translational invariance and elastic invariance in the papers of this week?

Ans: Invariance refers to a model's response being unchanged under transformations of the input. For example, an object’s identity remains the same regardless of its position in an image.
Equivariance means the output transforms in a predictable way as the input is transformed. For example, the position of an object in an output feature map changes as the object moves in the input image.

Convolutional layers are naturally translationally equivariant, meaning that if the input image is shifted, the features extracted by these layers will also shift in the same way. The U-Net paper describes the use of random elastic deformations of the training samples as a key concept for training the segmentation network with very few annotated images. This strategy allows the network to learn elastic invariance. Additionally, pooling layers in the U-Net architecture contribute to translational invariance, where the output remains relatively unchanged despite shifts in the input.

4. Does U-Net overfit? What applications would U-Net perform poorly on?

Ans: Yes, because U-Net is designed for biomedical image segmentation and may overfit if trained on small, highly specific datasets without enough variation. It could perform poorly on applications where context outside the local area of focus is crucial, given its design focuses on local features.

5. Why do segmentation tasks use *focal loss* rather than *cross-entropy*?

Ans: Focal loss is used to address class imbalance in segmentation tasks. It applies a modulating factor to the standard cross-entropy loss, reducing the relative loss for well-classified examples and focusing more on hard, misclassified examples.

6. What is a foundational model and why is Segment Anything a foundational model?

Ans: Foundational models are large-scale, versatile models that can be adapted to a wide range of tasks and data types. Segment Anything Model is considered to be a foundational model because of its general applicability to various segmentation tasks and its ability to adapt to new, unseen data distributions.

7. Why do you think Segment Anything was designed in a way that the image encoder can take longer than the other components such as the prompt encoder?

Ans: The design choice where the image encoder takes longer than other components like the prompt encoder in the Segment-Anything architecture is likely to facilitate flexibility and efficiency. The image encoder's complexity allows it to extract rich, reusable features from images, while other components can operate faster, focusing on specific tasks or prompts.

8. How is *cross-attention* used in Segment Anything?

Ans: Cross-attention mechanisms in “Segment Anything” enable the interaction of features from different sources like the image encodings and textual prompts, allowing the model to focus on relevant parts of the image for a specific task or prompt.

9. Very briefly: what are some important scaling laws in NLP.

Ans: Some key scaling laws in NLP include the observation that larger models tend to perform better, improvements in task performance with more training data, and the importance of efficient training regimes to manage computational costs.

10. What are examples of zero-shot and few-shot learning in vision?

Ans: In class, one student spoke about, I am not sure if I heard it correctly, how a few pictures of a dog can be given along with appropriate masks of a dog and how a model can learn from that and identify that particular dog in other photos.

SAM is an example of a zero shot model in vision.