Homework 4: Machine Translation

CS 1470/2470

Due October 30, 2020 at 11:59pm AoE

1 Conceptual Questions

1. What is the purpose of the positional encoding in the Transformer architecture? (2-3 sentences)
2. What are the limitations of RNNs that Transformers solve? (3-6 sentences)
3. Consider the parameters for two different attention heads. Is it necessary that they be initialized randomly, or could we just start them all with the same vector of values? (2-4 sentences)
4. In an encoder, that has attention from right to left, and left to right is the attention that word 5 pays to word 10, is it the same as that word 10 pays to word 5? (3-5 sentences)
5. (Optional) Have feedback for this assignment? Found something confusing? We'd love to hear from you!

2 Ethical Implications

Sustainable Deep Learning: In lecture, we explored the environmental impacts of deep learning, and examined how transformers are some of the most energy-intensive models to train. We also discussed suggestions for ways in which the field can keep sustainability in mind moving forward. For further reading, check out this Forbes article or this paper on quantifying carbon emissions.

1. What is network pruning? How and why does it work? How can it reduce long-term environmental effects of deep learning? (3-5 sentences)

2. What are at least two ways deep learning engineers like yourself can minimize the environmental impact of your models? For each method, explain how and why it preserves the environment. (3-6 sentences)

3. What are at least two ways can we hold engineers, in research and/or industry settings, accountable to minimize the environmental impact of their deep learning work? (3-6 sentences)

3 CS2470-only Questions

1. What requires more parameters: single or multi-headed attention? Explain. Does this mean one necessarily trains faster than another? (3-5 sentences)

2. For the last homework, we asked you to consider convolutional architectures for language modeling, and to weigh their trade-offs against RNN-based architectures.

   Transformers can also be used for language modeling. In fact, they are current state of the art.
   (see https://openai.com/blog/better-language-models/).

   How are transformer-based language models similar to convolution? What makes them more suited for language modeling? (5-10 sentences)

3. Read about BERT, a state-of-the-art transformer-based language model here: https://arxiv.org/pdf/1810.04805.pdf, and answer the following questions.

   (a) What do the researchers claim is novel about BERT? Why is this better than previous forms of language modeling techniques?

   (b) What is the masked language model objective? Describe this in 1-2 sentences.

   (c) Pretraining and finetuning both are forms of training a model. What's the difference between pretraining and finetuning, and how does BERT use both techniques?