15-780 HW7 + HW8

HW7: Due 4/10, HW8: Due 4/17

We're releasing HW7 and HW8 simultaneously, because HW8 is quite open ended, and you can use additional time during the weeks to complete it.

HW7: Llama-2

In this homework, you'll implement the Llama-2 language model (`https://arxiv.org/abs/2307.09288`). The main code you'll build upon is contained in the corresponding `hw7.zip` file. In addition to the data in this file, you'll need to download the official weights of the llama-2-7b-chat model, which you can download after filling out this form: `https://llama.meta.com/llama-downloads/`.

You'll want to implement your code in the `llama.py` file, filling in the missing implementation of all the functions. The `hw7.ipynb` file then contains all the logic you'll need to run the resulting model. For reference, the official Llama-2 model definition can be found at `https://github.com/meta-llama/llama/blob/main/llama/model.py`.

HW8: A reference implementation

Pick one of the other models we have covered in class, e.g., BERT (`https://arxiv.org/abs/1810.04805`), CLIP (one of the Transformer variants, `https://github.com/openai/CLIP`), or a diffusion model (again, a Transformer version like the one here `https://github.com/facebookresearch/DiT/`).

Pick any reference implementation of any of these models (you can find any on GitHub, but you'll want to pick one where you can download the pretrained weights), and implement a version from scratch in PyTorch (in the same way we have done in class, using none of the Pytorch `nn` classes except the base `Module` class itself. Demonstrate the output of the model.