Homework 2: Prompting & Generation with LMs (50 points)

Contents

The second homework zooms in on the following skills: on gaining a deeper understanding of different state-of-the-art prompting techniques and training your critical conceptual thinking regarding research on LMs.

Logistics

- submission deadline: June 2nd th 23:59 German time via Moodle
  - please upload a **SINGLE .IPYNB FILE named Surname_FirstName_HW2.ipynb** containing your solutions of the homework.
- please solve and submit the homework **individually**!
- if you use Colab, to speed up the execution of the code on Colab, you can use the available GPU (if Colab resources allow). For that, before executing your code, navigate to Runtime > Change runtime type > GPU > Save.

Exercise 1: Advanced prompting strategies (16 points)

The lecture discussed various sophisticated ways of prompting language models for generating texts. Please answer the following questions about prompting techniques in context of different models, and write down your answers, briefly explaining them (max. 3 sentences). Feel free to actually implement some of the prompting strategies to play around with them and build your intuitions.

> Consider the following language models:
>
> - GPT-2, GPT-4, Vicuna (an instruction-tuned version of Llama) and Llama-2-7b-base.
>
> Consider the following prompting / generation strategies:
>
> - beam search, tree-of-thought reasoning, zero-shot CoT prompting, few-shot CoT prompting, few-shot prompting.
>
> For each model, which strategies do you think work well, and why? Do you think there are particular tasks or contexts, in which they work better, than in others?

Exercise 2: Prompting for NLI & Multiple-choice QA (14 points)

In this exercise, you can let your creativity flow – your task is to come up with prompts for language models such that they achieve maximal accuracy on the following example tasks. Feel free to take inspiration from the in-class examples of the sentiment classification task. Also feel free to play around with the decoding scheme and see how it interacts with the different prompts.

TASK:

> Use the code that was introduced in the Intro to HF sheet to load the model and generate predictions from it with your

Skip to main content

> - Please provide your code.
> - Please report the best prompt that you found for each model and task (i.e., NLI and multiple choice QA), and the decoding scheme parameters that you used.
> - Please write a brief summary of your explorations, stating what you tried, what worked (better), why you think that is.

- Models: Pythia-410m, Pythia-1.4b
- Tasks: please **test** the model on the following sentences and report the accuracy of the model with your best prompt and decoding configurations.
  - Natural language inference: the task is to classify whether two sentences form a “contradiction” or an “entailment”, or the relation is “neutral”. The gold labels are provided for reference here, but obviously shouldn’t be given to the model at test time.
    - A person on a horse jumps over a broken down airplane. A person is training his horse for a competition. neutral
    - A person on a horse jumps over a broken down airplane. A person is outdoors, on a horse. entailment
    - Children smiling and waving at camera. There are children present. entailment
    - A boy is jumping on skateboard in the middle of a red bridge. The boy skates down the sidewalk. contradiction
    - An older man sits with his orange juice at a small table in a coffee shop while employees in bright colored shirts smile in the background. An older man drinks his juice as he waits for his daughter to get off work. neutral
    - High fashion ladies wait outside a tram beside a crowd of people in the city. The women do not care what clothes they wear. contradiction
  - Multiple choice QA: the task is to predict the correct answer option for the question, given the question and the options (like in the task of Ex. 3 of homework 1). The gold labels are provided for reference here, but obviously shouldn’t be given to the model at test time.
    - The only baggage the woman checked was a drawstring bag, where was she heading with it? [“garbage can”, “military”, “jewelry store”, “safe”, “airport”] – airport
    - To prevent any glare during the big football game he made sure to clean the dust of his what? [“television”, “attic”, “corner”, “they cannot clean corner and library during football match they cannot need that”, “ground”] – television
    - The president is the leader of what institution? [“walmart”, “white house”, “country”, “corporation”, “government”] – country
    - What kind of driving leads to accidents? [“stressful”, “dangerous”, “fun”, “illegal”, “deadly”] – dangerous
    - Can you name a good reason for attending school? [“get smart”, “boredom”, “colds and flu”, “taking tests”, “spend time”] – “get smart”
    - Stanley had a dream that was very vivid and scary. He had trouble telling it from what? [“imagination”, “reality”, “dreamworker”, “nightmare”, “awake”] – reality

Exercise 3: First neural LM (20 points)

Next to reading and understanding package documentations, a key skill for NLP researchers and practitioners is reading and critically assessing NLP literature. The density, but also the style of NLP literature has undergone a significant shift in the recent years with increasing acceleration of progress. Your task in this exercise is to read a paper about one of the first successful neural langauge models, understand its key architectural components and compare how these key components have evolved in modern systems that were discussed in the lecture.

> Specifically, please read this paper and answer the following questions: Bengio et al. (2003)
>
> - How were words / tokens represented? What is the difference / similarity to modern LLMs?
> - How was the context represented? What is the difference / similarity to modern LLMs?
> - What is the curse of dimensionality? Give a concrete example in the context of language modeling.
> - Which training data was used? What is the difference / similarity to modern LLMs?
> - Which components of the Bengio et al. (2003) model (if any) can be found in modern LMs?

Skip to main content

> - Please formulate one question about the paper (not the same as the questions above) and post it to the dedicated **Forum** space, and **answer 1 other question** about the paper.

Furthermore, your task is to carefully dissect the paper by Bengio et al. (2003) and analyse its structure and style in comparison to another more recent paper: Devlin et al. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

TASK:

> For each section of the Bengio et al. (2003) paper, what are key differences between the way it is written, the included contents, to the BERT paper (Devlin et al., 2019)? What are key similarities? Write max. 2 sentences per section.

‹ Previous
Homework 1: Language models (50 points)