Efficient Fine-tuning of Large Language Models

TL;DR: a one-sentence project summary

Light-weight while effective fine-tuning method for language models.

Brief Project Description

Fine-tuning with language models (LMs) often brings in performance gains over NLP benchmarks. As LMs getting larger, fine-tuning with ever-growing LMs needs to be more efficient. Recently, prompting-based and adapter-based methods shed lights into parameter-wise efficient fine-tuning. For instance, recent works claim that only tuning with 0.1% of parameters of BERT marks performance results close to as if tuning with the full model. This reduce the fine-tuning computation costs by order of magnitude. Can this be even more effective? Can we do efficient tuning that is task agnostic? What if we add in light-weight adaptive attention layers and only tune on those layers? These are the questions I am interesting in investigating.

Mentor Name

Zhengxuan Wu

Mentor Contact Email

wuzhengx@stanford.edu

Neural-Augmented Retrieval for Open-Domain Dialogue Systems

TL;DR: a one-sentence project summary

Add retrieval-augmented generation to an open-domain chatbot for fun and profit!

Brief Project Description

Mixed neural-handwritten open-domain dialogue systems such as Stanford's Chirpy Cardinal (https://stanfordnlp.github.io/chirpycardinal/) have recently achieved success in the wild in evaluations such as the Alexa Prize (https://developer.amazon.com/alexaprize). In particular, Chirpy Cardinal has been open-sourced as an extensible framework for open-domain social dialogue (https://github.com/stanfordnlp/chirpycardinal). The goal of this project is to add retrieval-augmented generation, or RAG (https://arxiv.org/abs/2005.11401) to Chirpy Cardinal; currently, we have a rudimentary system. If successful, we hope to deploy this system in Chirpy's online demo (https://anonchirpy.github.io/).

Mentor Name

Ethan Chi

Mentor Contact Email

ethanchi@cs.stanford.edu

Other Comments

This project is likely to be intensive and the mentor hopes to meet with the students weekly.

Initiative Detection in an Open-Domain Neural Conversational Agent

TL;DR: a one-sentence project summary

In an open-domain neural chatbot, detect whether people are interested and modify what you're saying accordingly!

Brief Project Description

Mixed neural-handwritten open-domain dialogue systems such as Stanford's Chirpy Cardinal (https://stanfordnlp.github.io/chirpycardinal/) have recently achieved success in the wild in evaluations like the Alexa Prize (https://developer.amazon.com/alexaprize). In particular, Chirpy Cardinal has been open-sourced as an extensible framework for open-domain social dialogue (https://github.com/stanfordnlp/chirpycardinal/).

An important part of any dialogue system is *initiative*, i.e. who's driving the conversation. To achieve an interesting and varied conversation, one aims to achieve mixed initiative, with both the bot and user taking turns. Previous work (Hardy et al. 2021; https://aclanthology.org/2021.sigdial-1.11.pdf) demonstrated that simple linguistic metrics such as # noun phrases correlate strongly with human judgements of initiative.

The goal of this project is to (1) collect a dataset for initiative; (2) build a neural model to automatically detect and classify this initiative; (3) implement a "global initiative policy" which modifies the dialog policy conditioned on the initiative observed by this neural model. For example, if the user has extremely low initiative turn-to-turn, we might go on longer and longer rants in an effort to entertain the user; but if the user has extremely high initiative, we should often ask the user if they want to change the topic.

If successful, we hope to open-source this project as part of Chirpy Cardinal.

Mentor Name

Ethan Chi

Mentor Contact Email

ethanchi@cs.stanford.edu

Other Comments

This project is expected to be intensive and students should plan to commit at least 10 hours per week each during the duration of the project.

Linguistically Guided Retrieval

TL;DR: a one-sentence project summary

Build new document retrieval systems using pretrained LMs and linguistics, sorta

Brief Project Description

Retrieval of documents relevant to an input query -- think Google -- has been advanced by the use of pre-trained language models to represent documents and queries. Documents retrieved in this way often are similar to the query in terms of _topic_. But we might want to retrieve documents that are similar to the query in other ways -- similar syntax, similar sentiment, etc. In this project, we'll repurpose methods developed for analyzing what pretrained LMs learn to instead build document retrieval systems that focus on specific aspects of language.

Mentor Name

John Hewitt

Mentor Contact Email

johnhew@stanford.edu

Word Embedding Initialization

TL;DR: a one-sentence project summary

Let's figure out how best to expand the vocabulary of pretrained LMs

Brief Project Description

Pretrained language models come with a fixed, pre-specified vocabulary. But maybe that means they split the text you're processing into a ton of word pieces, oh no! So you add new words to the vocabulary of the LM. But now these new embeddings have to be initialized and then trained. What's the best way to do this? Turns out, just averaging all the old embeddings is a nice baseline, as I discuss in this blog post: https://nlp.stanford.edu/~johnhew/vocab-expansion.html . Let's figure out how to do even better.

Mentor Name

John Hewitt

Mentor Contact Email

johnhew@stanford.edu

Reverse Dictionary

TL;DR: a one-sentence project summary

Reverse Dictionary: a method to find a word you can't remember by describing its meaning

Brief Project Description

The goal is to produce a reverse dictionary web app, which allows you to find a word you can't remember by describing its meaning. The goal will be to compare and contract multiple possible approaches and to find the best way to do this, both in terms of accuracy and in terms of computational workload and speed. Ideally, another outcome of the project will be a working website hosted in the cloud and available for anyone to use.

Mentor Name

Andrey Kurenkov

Mentor Contact Email

andreyk@stanford.edu

Other Comments

I will be happy to meet once a week and be available over email for giving advice!

CoAuthor: Human-AI Collaborative Writing

TL;DR: a one-sentence project summary

Explore how humans and language models collaborate and complement each other in interactive settings, by analyzing rich interactions captured by CoAuthor

Brief Project Description

How can language models help humans think out of the box and write creatively? Can they give individuals the nudge to write things that they would not typically think of? Which points in the writing process are most challenging for humans and when might they benefit from help?

We would like to deepen our understanding of how humans and language models collaborate and complement each other in interactive settings, by analyzing rich interactions captured by CoAuthor. CoAuthor is a new, replayable dataset consisting of detailed recordings of people using GPT-3 to write creative stories and argumentative essays (https://coauthor.stanford.edu/). Another more technical approach is to utilize the dataset to fine-tune and/or evaluate existing language models' ability to support humans in the writing process.

If you are interested in this direction, please take a look at our paper (https://arxiv.org/pdf/2201.06796.pdf) and email Mina Lee (minalee@cs.stanford.edu) with a few ideas you would like to try out in the final project.

Mentor Name

Mina Lee

Mentor Contact Email

minalee@cs.stanford.edu

Other Comments

I am happy to work with multiple teams as long as they are committed and passionate about the topic! Based on the availability of students, we can also consider submitting the final report to The First Workshop on Intelligent and Interactive Writing Assistants @ ACL 2022 (https://in2writing.glitch.me/).

Reasoning with natural language and knowledge graphs

TL;DR: a one-sentence project summary

Develop new methods to incorporate knowledge graphs into NLP models for robust question answering. Alternatively, explore new problems and tasks where existing knowledge-aware NLP models can be applied to.

Brief Project Description

Written natural language (e.g., stories, news articles, conversations) elides implicit knowledge that is fundamental to the understanding of the situation being described. Relationships among different knowledge categories (common sense, facts, etc.) are often not explicitly stated in text, but are nonetheless immediately gleaned or inferred by human readers. For NLP agents to understand text with similar depth, they must be able to represent these underlying characteristics. Knowledge graphs provide a promising opportunity to better capture such background knowledge. The goals of projects include: (1) explore new problems and tasks that can leverage knowledge graphs to fill in the gaps missing from written textual context, and (2) design new approaches for reasoning over written language and underlying knowledge (knowledge graphs) to perform robust question answering or understanding of situations described in language.

Mentor Name

Michihiro Yasunaga

Mentor Contact Email

myasu@stanford.edu

Improving Semantic Affordance Mining

TL;DR: a one-sentence project summary

Improve the detection on whether an action can be implemented on an object.

Brief Project Description

Semantic affordances are fundamental object attributes describing whether an activity can be associated with an object. Mining semantic affordance has attracted researchers from both natural language processing and the computer vision communities. It has been considered crucial for robotics and other areas. Following the previous work (https://www.cs.princeton.edu/~jiadeng/paper/chao_cvpr2015.pdf), improve the semantic affordance mining with the use of modern pretrained language representation models or improved traverse on the lexical bases, such as WordNet.

Mentor Name

To be determined!

Mentor Contact Email

em7@stanford.edu

Other Comments

A simple task with very old baseline. Students might be able to try multiple interesting methods since the formation of the task is simple (binary classification or regression with 2 words as the input).

Bi-encoder structure for data explanation matching

TL;DR: a one-sentence project summary

matching data with their explanation using bi-encoder structure

Brief Project Description

Explainable Natural Language Processing has acquired increasing attention from the community. Multiple datasets focusing on collecting human-annotated textual explanations have been proposed (https://openreview.net/pdf?id=ogNcxJn32BZ). However, it remains uncertain to what extent the models can understand the explanation. In this project, we examine the usage of the bi-encoder architecture (https://aclanthology.org/2020.acl-main.95/) on matching the data with their explanation text.

Mentor Name

To be determined!

Mentor Contact Email

em7@stanford.edu

Other Comments

Students are able to make their own explainability matching dataset sourcing from various previous work and studying the binary classification task with an architecture that works on multiple tasks with two kinds of heterogeneous data.

Understanding the Acceptance Criteria and Innovation Concepts of the US Patent Applications Using Large Language Models

TL;DR: a one-sentence project summary

This project seeks to use large language models to answer some of the ambitious, open-ended questions in patent analysis and to examine how the US patent applications have been evolving in the past twenty years across different technology areas.

Brief Project Description

Innovation is a major driver of economic and social development, and information about many kinds of innovation is embedded in semi-structured data from patents and patent applications. Though the impact and novelty of innovations expressed in patent data are difficult to measure through traditional means, machine learning offers a promising set of techniques for evaluating novelty, summarizing contributions, and embedding semantics. This project seeks to use large language models to explore the textual characteristics of the 4.5 million patent applications filed to the USPTO between 2004 and 2018 and answer some of the open-ended questions in the field of patent analysis.

Depending on the intellectual interests and involvement of the students, the project can focus on certain tasks, experiments, or questions. The following list contains some of the questions that the students can explore in this project, but it should be noted that the students need not limit themselves to these questions—we can definitely look at some other aspect(s) of patent applications using NLP methods: (i) How do accepted patent applications differ from rejected patent applications in terms of their semantic contents and syntactic structures? (ii) Can we train neural classifiers to predict the acceptance likelihood of a patent application using its abstract/claims section and metadata information? (iii) What are the textual characteristics of “super-star” inventions—do valuable patent inventions indeed combine ideas from different technology areas? (Can we predict the market value of a patent application at the time of its filing?) (iv) How have the patent acceptance criteria evolved across different technology areas and over time? (v) Can we determine and quantify the impact of Alice v CLS Bank International (2014) on software/tech patent applications after 2014? (vi) Given a patent application, can we find similar patents or patent applications? (Though this retrieval task might sound rather simple, it is actually quite difficult; it is also a crucial task for patent applicants and patent examiners alike.)

Mentor Name

Mirac Suzgun

Mentor Contact Email

msuzgun@stanford.edu

Designing Automatic Metrics for Story Generation

TL;DR: a one-sentence project summary

Design robust automatic evaluation metrics for story generation tasks.

Brief Project Description

Large pre-trained language models have achieved many success in language generation tasks, including open-ended story generation. However, most of the story generation evaluation relies on human annotation, which could be expensive and hard to scale. Automatic evaluation of stories would allow cheap and replicable evaluations. Some recent efforts include [https://arxiv.org/abs/2009.07602 , https://aclanthology.org/2021.naacl-main.343]

The goal of the project is to design robust automatic evaluation metrics for stories. We could show how these metrics correlate with human judgment .We could also exploit these metrics, and hope to design better search algorithms for story generation.

Mentor Name

Lisa Li

Mentor Contact Email

xlisali@stanford.edu

Estimating pretrained transformers' knowledge of multi-subword entities.

TL;DR: a one-sentence project summary

Evaluating pretrained transformer models' knowledge of entities that model tokenizers split into multiple subwords is difficult but potentially useful. How can we do it efficiently?

Brief Project Description

When doing interpretability or analysis work in NLP, we often want to compare the probability of one token to another. For example, in syntactic evaluation, we might want to know if the a model puts more probability mass on a singular verb compared to a plural one. Or when determining what factual information a model knows, we want to check if a model puts the most probability mass on the correct entity (e.g. predicting "Ottawa" in the context "The capital of Canada is [MASK]."). However, there are some tokens that are split into multiple subwords by the model's tokenizer. We can't use these tokens during evaluation because it's either difficult or impossible to compare the probability of a vocabulary item to one not in the vocabulary[1][2]. Usually, these tokens are just thrown out, but this is unsatisfying: there is potentially interesting knowledge that we can't test if our model knows. A natural method for addressing this problem is to add the subword to the vocabulary, however this requires re-training the model (or at least continuing pretraining), which can be expensive for large models. Can we come up with a more efficient way to estimate models' knowledge on multi-subwords tokens?

[1] In particular, it's not possible for bidirectional-MLM models like BERT, and is sometimes possible for causal-LM models like GPT2.

[2] For example, we couldn't ask "The capital of the US is [MASK]" because "D.C." is not a single token (and "Washington D.C." is multiple words)

Mentor Name

Benjamin Newman

Mentor Contact Email

blnewman@stanford.edu

Mental health analysis of social media posts during the COVID-19 pandemic

TL;DR: a one-sentence project summary

Analysis of mental health disorders using social media posts on a variety of platforms over the course of the COVID-19 pandemic

Brief Project Description

As the COVID-19 pandemic settled in, reports of deteriorating mental health only seemed to rise. Without the ability to socialize with friends and colleagues in-person, people turned to popular social media platforms to find solace. As mental health still remains a taboo topic in a lot of households, many people are unwilling to directly admit their mental health issues and seek help to treat them. Rather, they write (sometimes, anonymized) coded posts about what they are feeling on their social media pages. We can use a deep learning model to analyze these posts and extract relevant information to detect whether the author potentially suffers from a mental health disorder. This can be used to later direct advertisements to the author that feature resources from which they may seek help.

Some recent work has been done to detect mental health illnesses from social media posts. However, few works have analyzed how the coronavirus affects mental health and how that is translated through social media. Moreover, most literature has been restricted to a single platform. Thus, it would interesting to see if a deep learning model trained on Twitter tweets, for example, can transfer well to Reddit posts. Additionally, many works have studied social media posts independent of time. Therefore, it would be insightful to understand how people's mental health has changed over the course of the pandemic.

Mentor Name

Elaine Sui

Mentor Contact Email

esui@stanford.edu

Chest X-ray radiology report generation

TL;DR: a one-sentence project summary

Generate chest X-ray radiology report by generating text or modifying template text.

Brief Project Description

The project can be building a model that takes X-ray images as input and generates the corresponding diagnostic radiology report, either by generating free text or by modifying existing template text. The model will be a vision-language model, and it can be evaluated on CheXpert: https://stanfordmlgroup.github.io/competitions/chexpert/.

Mentor Name

Kathy Yu

Mentor Contact Email

fyu9@stanford.edu

Understanding the learning dynamics of word2vec

TL;DR: a one-sentence project summary

Understand the training dynamics of word2vec

Brief Project Description

Considerable study has been devoted to understanding the properties of fully-trained word2vec vector sets. Word analogies, semantic similarity and relatedness, general utility to a wide range of downstream tasks -- word2vec vectors helped lead to an explosion of NLP progress. But what does the learning process of word2vec look like? What can we learn from theoretical and/or empirical analysis of word2vec word vectors just a few (or a few hundred (or a few thousand)) gradient steps into the training process?

In this project, we'll explore and finally settle on one of a small set of theoretical and/or empirical studies in understanding word2vec's vectors' properties during training. For example, students in 224n have pointed out that, disconcertingly, the estimated loss of word2vec (on our provided dataset in assignment 2) initially increases before eventually decreasing and converging. Can we confirm empirically that this is a general behavior? Can we prove theoretically some conditions under which it will occur? As another example, it's been shown that word vectors encode basic syntactic properties like whether a word is a noun or a verb, along with more famous properties like word analogy linear offsets. When during training do different well-known properties emerge? Do performances on these metrics measuring these properties grow linearly with training time or is there some form of phase transition?

Mentor Name

John Hewitt

Mentor Contact Email

johnhew@stanford.edu

Characterizing Changes in the Language of Jobs

TL;DR: a one-sentence project summary

Defining a measure of tone in job postings, and measuring differences in tone across ten years of job postings

Brief Project Description

One area in which there is a wealth of new text data is job search. Firms post job advertisements, highlighting characteristics of the firm, the role, and the amenities. Though the explicit content of a posting has been studied, little research has been done on the tone of postings and how they differ across places, industries, and roles.

In this project, you will use natural language processing techniques on ten years of job postings to measure differences in tone across postings. You can define tone based on previous research or your own novel methodology. Some potential questions include: have differences in tone grown over time? Are there notable differences in tone even within the same firm? Are there clear leaders and followers in the diffusion of tone in job postings? What elements of postings exhibit the greatest differences in tone?

Mentor Name

Sarah Bana

Mentor Contact Email

sbana@stanford.edu

Characterizing Elements of Job Postings

TL;DR: a one-sentence project summary

Tag various parts of job postings based on deep learning methods

Brief Project Description

One area in which there is a wealth of new text data is job search. Firms post job advertisements, highlighting characteristics of the firm, the role, and the amenities. Though most postings have critical elements (Education requirements, benefits), these elements may not exist in each posting. Schemas, such as one defined on Schema.org/JobPosting, have been developed, but are not well utilized.

In this project, you will use natural language processing techniques on ten years of job postings to develop an approach to tag the various parts of job postings. Has the amount of content devoted to different categories changed over time? Do different types of organizations devote more text to one type of content versus another?

Mentor Name

Sarah Bana

Mentor Contact Email

sbana@stanford.edu

The Diffusion of Concepts in Job Postings

TL;DR: a one-sentence project summary

Use NLP to identify new concepts diffusing across the labor market in job postings.

Brief Project Description

One area in which there is a wealth of new text data is job search. Firms post job advertisements, highlighting characteristics of the firm, the role, and the amenities. Because these are publicly available, it is quite possible to use others' postings to craft a new posting. However, there has been no systematic research highlighting diffusion of language in postings.

In this project, you will use natural language processing techniques on ten years of job postings to identify patterns of diffusion of new concepts. These concepts can be technological, skills-based, or socio-cultural. Are there leaders and laggards in these concepts? Do they vary by geography, industry, or occupation?

Mentor Name

Sarah Bana

Mentor Contact Email

sbana@stanford.edu

A Natural Language interface to exploring datasets

TL;DR: a one-sentence project summary

Leverage Language models for exploring datasets based on language descriptions

Brief Project Description

Most interfaces for exploring text-based datasets are based on either regexes or tree-regexes (regex over syntax trees), or SQL. The first two require writing regex patterns that have low expressivity while the third requires SQL which may have higher expressivity but is still unable to capture certain aspects we might want to retrieve based on. For instance, consider an ML practitioner wanting to explore a sentiment classification dataset. Possible queries might include:
- retrieve all examples where the movie plot is bad
- retrieve all examples where the plot is good, but the casting is bad.
- retrieve examples where the review is about italian, indian or chinese food.
- retrieve examples where reviewer speaks with the manager
- retrieve examples where the reviewer's friends liked the restaurant but the reviewer did not.

Why language: Language is a natural & powerful way for people to communicate abstractions and can be used to provide arbitrarily complex descriptions for dataset exploration.

While one could write a complex SQL for a subset of these, we'd like to leverage modern NLP machinery to increase coverage (possibly at the cost of reduced precision). Concretely, the goal is to design a retriever function that takes as input a *language description* of the salient features that practitioner wants to retrieve based on, and outputs al examples that match the language description. In a way, the LM can be thought of as implementing a soft SQL and then executing it against the dataset to retrieve relevant queries (in this case, data points).

Possible use cases:
- Dataset exploration
- Selectively removing toxic content from datasets ('retrieve all examples where the reviewer uses foul language")
- Selectively upweighting examples based on knowledge of phenomenon in the downstream application ("retrieve all examples that require counting")

Mentor Name

Shikhar Murty

Mentor Contact Email

jsmurty@stanford.edu

Measuring Domain Generalization of Fewshot learning methods

TL;DR: a one-sentence project summary

We aim to measure robustness of various fewshot finetuning methods to distribution shifts.

Brief Project Description

In the last year, several approaches have been proposed for adapting pre-trained LMs to solve downstream tasks. This including "light-weight finetuning" approaches (Adapters, finetuning just the bias terms of the transformer etc), learning optimal discrete prompts (NullPrompt, LM-BFF), learning soft prompts, and learning optimal examples to use for in-context learning.

While all these methods give substantial gains, we are interested in studying which of these are most robust to distribution shifts. In particular, given a model that's been fewshot adapted based on a source distribution, how does performance deteriorate when the model is evaluated on a different set of target distributions? Are there qualitative differences between these methods? How does this robustness change as we provide more data from the source distribution for adapting the LMs?

Finally, is there a tradeoff between source distribution fewshot accuracy vs robustness to other distributions?

Possible application: Fewshot finetuning on one of the datasets from the fewshot QA track, and measure generalization to others.

Mentor Name

Shikhar Murty

Mentor Contact Email

jsmurty@stanford.edu

Multi-Domain Retrieval in Question Answering

TL;DR: a one-sentence project summary

In this project we'll study how to design a robust retriever for multiple data distributions in open domain question answering.

Brief Project Description

Open-domain question answering (ODQA) entails providing an answer to a factoid question expressed in natural language, about nearly anything and without explicitly provided context from which to find the answer. Prevailing methods for ODQA collect a large collection of documents and follow a retrieve-and-read approach where the retriever model retrieves a small set of relevant documents from the collection, and the reader model extracts an answer from the subset of documents. The answer is typically a span from one or more of the retrieved passages. Recent work demonstrates that dense passage retrieval methods, i.e., encoding the questions and passages as dense embeddings and using embedding distances to select retrieved passages, exceeds the performance of legacy information retrieval methods methods, which used sparse information theoretic features of the questions and passages to determine which passages to retrieve for a given query.

Existing work largely focuses on training and retrieving using a single distribution of passages (e.g., Wikipedia passages). However realistically passages come from diverse distributions and the process of simultaneously retrieving over in distribution and out-of-distribution datasources is poorly understood. The goal of this project is, given questions that can be answered by passages from multiple distributions (e.g., Wikipedia and Stack Overflow data), characterize the performance of using a single retriever for both, using retrievers fine-tuned on each, and then investigate how to design a single robust retriever.

Mentor Name

Simran Arora

Mentor Contact Email

simarora@stanford.edu

Helping Humans Express Emotions Through Emojis

TL;DR: a one-sentence project summary

Predict emojis for a message & generalize zero-shot to new emojis

Brief Project Description

Using emojis on Slack is challenging because new emojis keep being added, and users can't keep track of them all. However, many people depend on emojis to accurately convey emotions through text messages. This is especially important when working from home, a setting where several dimensions of expression are lost.

Given a dataset of messages, each annotated with an emoji, can you learn to predict emojis for new messages? Unlike in the traditional text classification setting, we want to be able to generalize zero-shot to new classes at test time. For example, if I say a new emoji has been added, and it's textual representation is :water-bottle:, your model should be able to detect that it is a good match for a message like "Done with morning run, time to hydrate!", but not "Wow look at this sunset". Since emojis also have associated images, a stretch goal would be to explore multi-modal embeddings from models such as CLIP.

Mentor Name

Gabriel Poesia

Mentor Contact Email

poesia@stanford.edu

Other Comments

Dataset: https://www.kaggle.com/rexhaif/emojifydata-en

Building language models that actually have consistent beliefs

TL;DR: a one-sentence project summary

Make models that behave in a more trustworthy, logically-consistent manner!

Brief Project Description

Current LMs are not consistent; for example, they may answer "is oxygen colorless" with "yes" and "what gas do plants produce?" with "oxygen", and then happily go on to answer "what color is the gas plants produce?" with "green." A logically consistent model should be able to realize that "green" is not a valid answer to the third question, given the past responses.

In this project, we will augment a pre-trained question-answering model with an external memory, to which we will store past model predictions. We will then add a pre-trained NLI model to enforce a soft consistency constraint between past model predictions and future predictions, measuring if this memory + constraint checker can improve model consistency without sacrificing accuracy. See this paper for reference: https://arxiv.org/abs/2109.14723

Mentor Name

Eric Mitchell

Mentor Contact Email

em7@stanford.edu

Other Comments

This project could turn into a more serious research project if the team is interested & committed

Improving language model consistency with sparse activation subspaces

TL;DR: a one-sentence project summary

Can we make language models behave more consistently by regularizing their internal activations?

Brief Project Description

Language models tend to give inconsistent predictions for related inputs; for example, a model might answer the question "who is the UK PM?" with "Boris Johnson" but the question "the prime minister of the UK is who?" with "Theresa May." We hypothesize that this failure mode occurs because the model does not correctly understand that these two inputs are actually asking the same question. In the terminology of this paper, the model has low $\mathbf{P}_{\text{reuse}}$: https://openreview.net/pdf?id=7uVcpu-gMD , or, insufficient "reusing" of the same components of the model's parameters to process similar inputs.

In this project, we will explore methods for increasing $\mathbf{P}_{\text{reuse}}$ in a pre-trained fact-checking model, particularly through inducing sparsity in the model's activations.

Mentor Name

Eric Mitchell

Mentor Contact Email

em7@stanford.edu

Other Comments

Mentor can meet weekly if the team is ambitious