Neeti Desai, Karen Zhang, Aditya Ratan Jannali

NLP Final Project Proposal

Our project includes two major phases: fine-tuning and interpretability. The first involves fine-tuning a GPT-2 model to classify emotions, using the emotion detection dataset from Kaggle. The second involves employing interpretability techniques to investigate the inner feature representation and architecture of the model. We are planning to mask different subsets of tokens (an entire attention head for example) in the transformer, and see how it affects the prediction output. We hope to be able to make meaningful conclusions on how information is represented throughout the architecture. Through masking different regions of tokens and observing the output, we can hopefully find regions which store different information, for example, one of the attention heads could mostly contain information related to joy and is crucial to the prediction of the joy label. If time permits, we can also employ other interpretability techniques like probing or visualizing the attention map.

We will be utilizing PyTorch, pandas, numpy, nltk, and scikit-learn packages to collect, clean, and process the data, and import GPT-2 models and tokenizers from HuggingFace transformer package. The main hyperparameters we will use for fine-tuning the models include (but are not limited to) - number of epochs, learning rate, and weight decay.

Model of choice:

GPT-2

a. **Language Modeling Strength**: GPT-2 is pre-trained on a massive amount of text data, which means it has learned rich representations and associations between words, making it very effective at generating meaningful context around text. This language modeling ability can be leveraged for sequence classification by training GPT-2 to generate specific "end tokens" or prompts that correlate with the class labels.

b. **Few-Shot and Zero-Shot Classification**: GPT-2 and other large decoder models can handle classification tasks even in few-shot or zero-shot settings. By formatting inputs with prompts or instructions, GPT-2 can interpret tasks like sentiment classification, topic identification, or other tasks without needing a specialized classification head.

Model Interpretability:

Masking based attention analysis for model interpretability

a. **Objective:** The intent of this technique is manipulate attention weights within specific regions (such as certain rows, full matrices, or selected heads) to assess their contributions to the model's prediction behavior for particular classes. By masking attention scores across various configurations, we aim to identify critical network components driving class-specific predictions.

b. **Class-Specific Behavior through Selective Masking**: When analyzing class-specific predictions, selective masking allows us to identify which attention heads or matrices contribute to the differentiation of classes. For instance, if masking certain rows or heads consistently reduces accuracy for a specific class, this indicates that those elements may capture features pertinent to that class, thereby highlighting their importance.

c. **Impact on Network Regions and Predictive Power**: Masking entire heads or subregions within the attention matrices can reveal how different layers and heads interact to reinforce or modify class-specific signals. By conducting this masking experiment, we can assess the model's reliance on particular heads or attention regions for prediction of certain classes, thereby mapping out an interpretability framework that emphasizes structural and functional importance within the network layers.

Model Evaluation:

a. To measure the performance of the model, we use task specific metrics like Accuracy, Precision, Recall, Specificity, ROC, and more.
b. A successful evaluation of interpretability will include feature importance, human evaluation, ablation studies, and task specific metrics (like F1, ROC, etc.)

Related Work:

1. Roles and Utilization of Attention Heads in Transformer-based Neural Language Models:
   a. The authors suggest an analysis method which helps understand where linguistic properties are learned and represented along attention heads. This is a good starting point to interpretability as it can guide us in designing masks to be applied.
2. Language Models are Unsupervised Multitask Learners
   a. We are planning to fine-tune GPT-2 on the task. In order for us to fine-tune and properly interpret model output and understand the inner representation, we would need to know the architecture of GPT-2.
3. A Practical Review of Mechanistic Interpretability for Transformer-Based Language Models
   a. This paper explains different interpretability techniques in a beginner-friendly way that can be applied to Language Models.
4. Token-level Masking for Transformers
   a. This paper covers token masking in transformers to regularize self-attention, which will provide useful background and techniques for us.
5. Masking important information to assess the robustness of a multimodal classifier for emotion recognition
   a. This is similar to what we are doing, so will also provide a good background.
6. Evaluating Emotional Detection & Classification Capabilities of GPT-2 & GPT-Neo Using Textual Data
   a. It will be interesting to read about GPT-2's general performance on emotion detection/classification tasks before we start fine tuning it on the task.