Project Proposal: Comparative Attention Analysis on Medical vs. Legal QA

Title:

Comparative Attention Analysis of Generative Language Models on Medical vs. Legal Question Answering

Description:
This project explores how generative language models (such as GPT-2 or T5) attend to different words when generating answers to questions in the medical versus legal domains. Using attention weights extracted from decoder layers, we aim to visualize and quantify how the model focuses shifts based on domain-specific content. This interpretability-focused study seeks to provide insights into domain adaptation, attention span, salience in complex question- and linguistic answering tasks.

Problem Statement

Large language models are increasingly utilized in sensitive fields such as medicine, and internal decision-making processes remain opaque, but their use remains legal. While these models can perform well across domains, we don't fully understand how they prioritize or focus on different parts of the input text in varying contexts. This project investigates:

- How attention is distributed over key words or phrases when answering questions in different domains.
- Whether medical and legal questions elicit different attention patterns, focal points, or reasoning paths.

This is interesting from both an interpretability and model design standpoint, especially when building trustworthy AI systems in high-stakes fields.

Background

Previous work on attention visualization and interpretability includes:

- Vaswani et al. (2017): Introduced the transformer architecture and the concept of self-attention, which is widely used in most large language models.
- Vig (2019): Developed BertViz for visualizing attention heads.

- Talmor et al. (2020): Studied multi-hop reasoning and focus shifts in QA tasks.
- BioBERT and LegalBERT: Domain-adapted transformer models, but little comparative work exists analyzing *attention behavior* across domains.

This project extends interpretability work by comparing raw attention weights between generic models (e.g., GPT-2) and domain-specific models that reason in a particular domain.

Methodology

1. **Dataset Sampling**

- **Medical domain**: MedQA (USMLE-style questions)
- **Legal domain**: COLIEE or synthetic legal QA pairs
- Create 20–50 QA pairs per domain for analysis.

2. **Model Selection**

- Use **GPT-2** or **T5 (small)** models from Hugging Face Transformers.
- Enable decoder attention tracking: `output_attentions=True`

3. **Inference and Attention Capture**

- Generate answers using pre-trained models.
- Extract decoder self-attention and cross-attention weights.
- Annotate domain-relevant keywords (e.g., "diagnosis", "plaintiff") using `spaCy` or `SciSpacy`

4. **Analysis and Visualization**

- Plot attention heatmaps for selected tokens.
- Compare keyword-focused attention, entropy, and distance.
- Use `matplotlib`, `bertviz`, or dimensionality reduction (e.g., PCA on attention vectors)

5. **Evaluation**

- Quantify:
    - Avg. attention on domain-specific keywords

- Entropy of attention distribution
- Focus span (token range/distance)

Expected Outcomes

- Attention heatmaps showing differences between medical and legal QA tasks
- Tables comparing metrics: entropy, keyword salience, span
- A short paper/report with figures, findings, and discussion
- Insights into how generative models implicitly adapt (or fail to adapt) their attention patterns across domains

---

Team Roles

Team Member	Primary Role	Secondary Role
1	Dataset curation, keyword tagging (spaCy, manual review)	Assist with ambiguity annotation and data preprocessing
2	Model loading, inference, and attention extraction using Hugging Face Transformers	Support visualization and debugging
3	Statistical analysis: entropy, attention span metrics	Review patterns and assist in interpreting trends
4	Visualization: attention heatmaps, PCA, comparative plots	Lead report writing and formatting

Everyone will contribute to:

- Final report formatting and interpretation
- Weekly updates and demo preparation