Hallucination Detection in Biomedical QA Systems Using Flan-T5-Large

Introduction

This report details my progress made in detecting and mitigating hallucinations in biomedical QA text generated by the Flan-T5-Large language model. The project explores fine-tuning techniques, specifically LoRA (Low-Rank Adaptation), to improve the model's ability to generate factual and accurate medical information, thereby reducing the occurrence of hallucinations. The work leverages the Hugging Face ecosystem for model management, fine-tuning, and evaluation.

Initial Benchmarking: Default vs. Parameterized Model

The initial phase involved benchmarking the performance of the base Flan-T5-Large model against a version where generation parameters and a specific prompt were applied. The goal was to see if simple prompt engineering and adjusting parameters like temperature, num_beams, and repetition_penalty could improve the quality and reduce potential hallucinations in medical question answering.

Using Hugging Face's built-in APIs and tools facilitated this process, providing a standardized way to load models, tokenizers, and evaluation metrics. A challenge encountered during local execution on a Mac was the difficulty in fully utilizing GPU power, necessitating limitations on the scale of initial experiments.

The first attempt with the default model and no specific parameters resulted in very short and often uninformative answers. By incorporating a detailed prompt (instructing the model to use evidence-based sources and provide reasoning) and tuning parameters, the model's responses became longer and better structured, although the factual accuracy still required further improvement.

The metric results for this initial comparison were as follows:

Default Model Results:

- BLEU: 0.0038
- METEOR: 0.0577
- ROUGE-1: 0.1205
- ROUGE-L: 0.1086
- BERTScore-F1: 0.8448

Parameterized Model Results:

- BLEU: 0.0512
- METEOR: 0.2239
- ROUGE-1: 0.2532
- ROUGE-L: 0.1998
- BERTScore-F1: 0.8727

These results show a clear improvement across all metrics when using the parameterized approach compared to the default generation settings.

Fine-tuning with LoRA

To further enhance the model's performance and address hallucinations, fine-tuning was performed using the LoRA technique. LoRA was chosen because it is significantly quicker and more memory-efficient compared to traditional full model fine-tuning, making it more suitable for environments with limited GPU resources, such as the local setup used in this project.

The LoRA configuration involved setting parameters like r (rank), lora_alpha, target_modules, lora_dropout, and task_type to adapt the model's weights efficiently. The model was trained for 20 epochs on a limited dataset (40 training examples) due to computational constraints.

Evaluation Metrics

To quantify the performance of the models, five standard metrics commonly used in natural language generation tasks were employed:

- **BLEU (Bilingual Evaluation Understudy):** Measures the n-gram overlap between the generated text and reference text, indicating the precision of the generated text.

- **METEOR (Metric for Evaluation of Translation with Explicit Ordering):** Considers precision, recall, and word-to-word matches with some linguistic variations, providing a more comprehensive measure of semantic similarity.
- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** Focuses on the overlap of n-grams, word sequences, and word pairs between the generated text and the reference text, particularly useful for evaluating summaries. ROUGE-1 measures unigram overlap, and ROUGE-L measures the longest common subsequence.
- **BERTScore:** Leverages pre-trained BERT embeddings to compute a similarity score between the generated and reference sentences, capturing semantic meaning beyond simple word overlap.

Results and Interpretation

Benchmarking was conducted on both a subset of the training data (recheck_train) and a separate test dataset.

The comparison between the default model and the parameterized model on an initial dataset showed improvements across all five metrics for the parameterized version. This indicated that prompt engineering and parameter tuning had a positive impact on the structure and content length of the generated responses, aligning them better with the desired output format.

The final tuning with LoRA involved comparing the default model and the LoRA-tuned model on both the recheck_train and test datasets. The results showed improvement across all metrics for the LoRA-tuned model on both datasets, with more significant gains observed on the recheck_train set.

Recheck Train Comparison:

- BLEU: Default Model - 0.0000, Retrained Model - 0.0631, Δ +0.0631 (Improvement: Yes)
- METEOR: Default Model - 0.0618, Retrained Model - 0.2043, Δ +0.1425 (Improvement: Yes)
- ROUGE-1: Default Model - 0.1113, Retrained Model - 0.2174, Δ +0.1061 (Improvement: Yes)
- ROUGE-L: Default Model - 0.0929, Retrained Model - 0.1689, Δ +0.0760 (Improvement:

Yes)

- BERTScore-F1: Default Model - 0.8509, Retrained Model - 0.8664, Δ +0.0155 (Improvement: Yes)

Test Data Comparison:

- BLEU: Default Model - 0.0000, Retrained Model - 0.0524, Δ +0.0524 (Improvement: Yes)
- METEOR: Default Model - 0.0579, Retrained Model - 0.2235, Δ +0.1656 (Improvement: Yes)
- ROUGE-1: Default Model - 0.1100, Retrained Model - 0.2104, Δ +0.1004 (Improvement: Yes)
- ROUGE-L: Default Model - 0.0924, Retrained Model - 0.1695, Δ +0.0771 (Improvement: Yes)
- BERTScore-F1: Default Model - 0.8489, Retrained Model - 0.8612, Δ +0.0123 (Improvement: Yes)

While the metrics showed improvement, particularly on the training data, manual inspection of the generated answers on new test data revealed that despite better sentence structure and length, the factual accuracy (and thus hallucination reduction) was not significantly improved in all cases.

Explaining the Discrepancy

The observed discrepancy where metrics improved, especially on the trained data, but the quality of answers on new test data wasn't consistently better can be attributed to several factors:

1. **Limited Training Data:** The small size of the training dataset (40 examples) is likely the primary reason. The model may have overfitted to the specific examples in the training set, learning to generate responses that score well on those particular questions and reference answers according to the metrics, without truly generalizing its understanding of factual medical information.
2. **Metric Limitations:** While the chosen metrics are standard, they primarily measure lexical and semantic similarity to the reference answers. They may not fully capture the

nuances of factual accuracy and the presence of subtle hallucinations, especially in a domain as complex as medicine. A response might be grammatically correct and semantically similar to a true statement but still contain a factual error.

3. **Complexity of Medical Knowledge:** Medical information is vast and requires deep understanding and access to up-to-date evidence. Fine-tuning on a small dataset, even with an advanced model like Flan-T5-Large, may not be sufficient to instill the necessary factual knowledge and reasoning abilities to avoid hallucinations consistently.

Future Directions

To further improve hallucination detection and reduction in medical text generation, several avenues can be explored:

- **Retrieval Augmented Generation (RAG):** Implementing a RAG system could significantly enhance factual accuracy. This involves retrieving relevant information from a trusted medical knowledge base or reliable sources based on the user's query and then using this retrieved information to guide the language model's generation process. This grounds the responses in external evidence, reducing reliance on the model's internal, potentially inaccurate, knowledge.
- **Larger and More Diverse Training Data:** Acquiring and utilizing a significantly larger and more diverse dataset of high-quality, fact-checked medical question-answer pairs would be crucial. Training on more extensive data would help the model generalize better and reduce overfitting.
- **Domain-Specific Pre-training or Fine-tuning:** Further fine-tuning on a large corpus of medical text (e.g., medical journals, textbooks, clinical guidelines) could improve the model's understanding of medical terminology and concepts.
- **Advanced Evaluation Metrics:** Exploring or developing evaluation metrics specifically designed to assess factual consistency and detect hallucinations in domain-specific text could provide a more accurate measure of progress.

References

- BERTScore: Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019).

BERTScore: Evaluating Text Generation with BERT. *arXiv preprint arXiv:1904.09675*. Available at: https://ar5iv.labs.arxiv.org/html/1904.09675

- BLEU: Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics* (pp. 311-318). Available at: https://www.researchgate.net/publication/2588204_BLEU_a_Method_for_Automatic_Evaluation_of_Machine_Translation