CSEN 1076

Milestone 3

Report

Khadiga Elzafarany 52-1145
Nayera Mahran 52-26764
Mariam Wael 52-1647

Overview:

In this milestone, we developed a question-answering chatbot using two transformer models: **BERT** and **FLAN-T5-base**. The BERT model (bert-large-uncased-whole-word-masking-finetuned-squad) is already fine-tuned on the SQuAD v1.1 dataset and was used as a strong baseline. In contrast, we applied **transfer learning** to the **FLAN-T5-base** model to enhance its ability to answer questions and better generalize unseen data.

The chatbot retrieves relevant context documents using sentence embeddings and then passes the question-context pair to each model to generate answers. This setup allowed us to compare the performance of a pre-finetuned model (BERT) and a transfer learning model (FLAN-T5-base). We evaluated the chatbot's performance using standard QA metrics such as **Exact Match (EM)** and **F1 score**.

1. BERT Model:

In this experiment, we built a chatbot using the BERT architecture, specifically the bert-large-uncased-whole-word-masking-finetuned-squad model, which is already fine-tuned on the SQuAD dataset. The system follows a retrieval-based pipeline where input questions are semantically matched to relevant contexts using the all-MiniLM-L6-v2 model from Sentence Transformers. After retrieving the top relevant paragraph from the context pool using cosine similarity, the BERT model performs extractive question answering to generate a span-based answer. To simulate a conversational setting, we incorporated chat memory by appending the last two (question, answer) pairs into the retrieval query. The system was evaluated using a 100-example subset of the SQuAD v1.1 validation set.

Evaluation:

We used 2 evaluation metrics:

1. **Exact Match:**
   The percentage of predictions that exactly match any one of the ground truth answers .
2. **F1-Score:**
   Measures the overlap (precision and recall) between your predicted answer and the ground truth answer, giving partial credit for partially correct answers.

Exact Match (EM)	**F1 score**
84.0%	**87.73%**

2. FLAN-T5:

In this experiment, we applied transfer learning on FLAN-T5-base model on a subset of the SQuAD dataset. The goal was to adapt the pre-trained model to the task of extractive question answering while minimizing computational overhead.

Transfer Learning:

To implement transfer learning, we used **selective layer freezing**, where only a small part of the model is trainable:

- **Frozen components**:
    - All encoder layers except the **last block**
    - All decoder layers except the **last block**
- **Trainable components**:
    - **LM head** : responsible for generating the output sequence
    - **Last encoder block**: allows limited representation adaptation
        - Allows the model to adapt its higher-level understanding of the input to the task.
        - Earlier encoder layers are responsible for more general language patterns (e.g., syntax, structure), while the last layer captures task-specific semantics.
    - **Last decoder block**: enables task-specific decoding adjustments
        - Adapts the **final stage of generation** to match the task's target outputs.
        - This layer has a strong influence on how the output is constructed—what type of phrasing, formatting, or keywords are used.

Evaluation:

Exact Match (EM)	**F1 score**
79.0%	**82.55%**

Visualization:

This is a comparison for Exact Match and F1 Score for all models that we tried. We chose BERT and FLAN-T5-base (transfer learning) as they achieved the highest Exact Match and F1 Score.