

Question answering with DistilBERT

Introduction – Extractive question answering

Extractive question answering (QA) is a task where a model identifies and extracts a specific segment of text from a given context to answer a question. Unlike generative models that create new text, extractive QA focuses on pinpointing the exact phrase or sentence from a predefined passage. This approach is particularly effective for tasks like retrieving facts, summarizing key points, or navigating dense documents. While the method has its limitations, one of its biggest advantages is that it requires much less computational power than generative tasks.

Model

As a model we use DistilBERT, a lightweight version of the BERT model, fine-tuned for the task of question answering. DistilBERT is designed to retain 97% of BERT's performance while being 40% smaller and 60% faster ¹

Base Model

- DistilBERT: Pretrained on large text corpora, the model leverages self-attention mechanisms to encode contextual information in text.
- Fine-tuned on the SQuAD (Stanford Question Answering Dataset) to specialize in answering questions based on a given context.

Parameter Efficiency

- Uses Parameter-Efficient Fine-Tuning (PEFT) with LoRA (Low-Rank Adaptation) for efficient training and adaptation:
- Adds trainable low-rank matrices to the pre-trained model's weights, reducing the need for full model retraining.
- LoRA parameters, such as `lora_alpha`, `r`, and `lora_dropout`, optimize the model for the task while maintaining computational efficiency.

Tokenization

- The model uses the AutoTokenizer from Hugging Face's Transformers library to preprocess questions and contexts into input tokens.
- Tokenized inputs are split into manageable chunks (max 384 tokens per chunk) to handle long contexts efficiently.

Quantization

Incorporates Bits and Bytes Quantization (4-bit) to reduce memory usage and inference time, enabling deployment on resource-constrained hardware.

Dataset

The MRQA (Machine Reading for Question Answering) dataset was developed as part of the MRQA 2019 Shared Task to evaluate the generalization capabilities of reading comprehension systems. This dataset unifies 18 question-answering (QA) datasets into a standardized extractive QA format, allowing models to test their ability to answer questions by identifying text spans in a given context.

The dataset consists of three splits, each containing six of the original datasets. The training split contains datasets usually made up of Wikipedia articles and general trivia questions. The two remaining datasets are for model development and evaluation and contain questions from different domains than the training split, mostly from biology textbooks in order to evaluate the generalizing performance of the trained models.

Training and evaluation

Preprocessing

Training

One major constraint for the training was that we did not have unlimited access to hardware powerful enough for the task. This meant that we had to rely on parameter-efficient training methods to reduce computational time. We implemented both quantization and low-rank adaption. We opted for a 4-bit quantization, which improved training speed by a lot. We have experimented with multiple lora ranks, and ended up

Evaluation

Quantifying a model's performance on extractive question answering is not a straightforward task. We chose three metrics to evaluate the model's performance, the exact match, fl-score and the bleu score.

Inference pipeline, ML as a service

We built an inference pipeline for the model that automatically fetches context for the questions through a google search and the wikipedia API. It works by creating a search term from the question through part-of-speech tagging each word in the original question using the nltk library and removing words and punctuation marks that fall into categories deemed irrelevant for a google search term. The final search term is then created by appending the remaining words with 'wikipedia' so that we retrieve the wikipedia articles in the top results.

Following this, the titles of the top 10 results are checked with the wikipedia API and the summaries are returned for the valid wikipedia articles. These summaries will be used as contexts for the question answering task. The summaries are then tokenized along with the questions. Since the model has a maximum capacity for taking in 512 token long contexts, but was finetuned with contexts no longer than 384 tokens, contexts that exceed that length are broken down into smaller chunks and fed to the model individually. A softmax function is then applied to the outputs, and the softmax score for the start and end indices is added together, effectively creating a confidence score (ranging from 0 to 2). The answer with the highest confidence score is returned as the final answer. This is done for all the retrieved valid wikipedia articles, and the final answers are sorted based on their confidence scores and returned in a descending order.

The project includes a basic reference implementation of Machine Learning as a Service. The service enables users to input questions and receive answers. The answers can be rated, and the ratings can be used for further tuning of the model. The ratings are saved along with the question, answer, and the full context, and saved individually in JSON files. In a real-life scenario, the ratings could be saved in a database and used for tuning the model.

The whole solution runs in a docker container. The API was written in Python using FastAPI. The front-end was written in Svelte.

Use cases

- Customer support
 - Automated response based on FAQ
- Knowledge bases
 - Trained on documents that are not publicly available a similar solution can provide answers on confidential material