Mini Project 1 Part 2 Report

Contribution:

@Huapeng Zhou: Implemnt the fine-tuning part, and do some experiments.

@Junfeng Zhou: Implement both ranking and fine-tuning parts, get better MAP score with triple InputExample. Write the Ranking Document Report.

Ranking Document Report

Comparison of Encoding Methods

> Q: Compare GloVe embeddings vs. Sentence Transformer embeddings.
> A:
>
> - The GloVe embeddings are pre-trained on a general corpus for word embeddings, which are more suitable for word similarity tasks.
> - The Sentence Transformer embeddings are trained for sentence embeddings, which are better to capture the semantic meaning of sentences.

> Q: Which method ranked documents better?
> A: The Mean Average Precision (MAP) of the Sentence Transformer embeddings is 0.4586, which is higher than the MAP of the GloVe embeddings, 0.0509.

> Q: Did the top-ranked documents make sense?
> Yes, the 1st ranked document to the query "Breast Cancer Cells Feed on Cholesterol" is, "While many factors are involved in the etiology of cancer, it has been clearly established that diet significantly impacts one's risk for this disease. More recently, specific food components have been identified which are uniquely beneficial in mitigating the risk of specific cancer subtypes. Plant sterols are well known for their effects on blood cholesterol levels, however research into their potential role in mitigating cancer risk remains in its infancy. As outlined in this review, the cholesterol modulating actions of plant sterols may overlap with their anti-cancer actions. Breast cancer is the most common malignancy affecting women and there remains a need for effective adjuvant therapies for this disease, for which plant sterols may play a distinctive role."
> The document is relevant to the query and provides information about the relationship between cholesterol and breast cancer.

> Q: How does cosine similarity behave with different embeddings?
> A: The cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. The Sentence Transformer embeddings have a higher cosine similarity with the query than the GloVe embeddings.

Observations on Cosine Similarity & Ranking

> Q: Did the ranking appear meaningful?
> A: The ranking of the documents is meaningful. The top-ranked documents are relevant to the query.
> For example, the 1st ranked document to the query "Breast Cancer Cells Feed on Cholesterol" is relevant to the query, which is about the relationship between cholesterol and breast cancer.

> Q: Were there cases where documents that should be highly ranked were not?
> A: No, the top-ranked documents are relevant to the query.

> Q: What are possible explanations for incorrect rankings?
> A: Although the top10 documents are relevant to the query, the possible explanations for incorrect rankings are:
>
> - Sentence Transformer embeddings are not fine-tuned on the dataset.
> - The cosine similarity may be not the best distance metric for ranking documents.
> - The documents are not preprocessed correctly (e.g., removing stopwords).

Possible Improvements

> Q: What can be done to improve document ranking?
> A: 1. Fine-tune the Sentence Transformer embeddings on the dataset. 2. Preprocess the documents (e.g., removing stopwords) before ranking.

> Q: Would a different distance metric (e.g., Euclidean, Manhattan) help?
> A: Yes, a different distance metric may help improve ranking. For example, the Euclidean distance metric may be better for ranking documents.

> Q: Would preprocessing the queries or documents (e.g., removing stopwords) improve ranking?
> A: Yes, preprocessing the queries or documents may improve ranking. For example, removing stopwords can reduce noise in the documents and help the model focus on the important words.

Fine-Tuning Report

2.1 Comparison of Different Training Strategies

> Q: [anchor, positive] vs [anchor, positive, negative]. How did the model behave differently?
> A: The SBERT model can learn better representations with the help of negative samples and get higher MAP score in [anchor, positive, negative] approach.

2.2 Impact on MAP Score

> Q: Did fine-tuning improve or hurt the Mean Average Precision (MAP) score?
> A: Fine-tuning decreased the MAP score in [anchor, positive] approach, the MAP score is **0.**4537, and decreased more in [anchor, positive, negative] approach with **triple loss**, the MAP score is **0.3618**. However, there is a marginal improvement in [anchor, positive, negative] approach with **negative loss**, the MAP score is **0.4611**.

> Q: If MAP decreased, why might that be?
> A: The MAP score decreased with the positive sample only approach partly because the model was not able to differentiate between positive and negative samples.

> Q: Is fine-tuning always necessary for retrieval models?
> A: No, fine-tuning is not always necessary for retrieval models. The necessity depends on several factors:
>
> 1. Performance Requirements: If the pre-trained model meets the required performance metrics, fine-tuning may be unnecessary.
> 2. Dataset Size: With very small datasets, fine-tuning might lead to overfitting rather than improvement.
>
> The decision to fine-tune should be based on empirical evaluation of the pre-trained model's performance on your specific use case and requirements.

2.3 Observations on Training Loss & Learning Rate

> Q: Did the loss converge?
> A: Yes, the loss converged in 4 epochs.

> Q: Was the learning rate too high or too low?
> A: The learning rate was appropriate, default as 2e-5.

> Q: How did freezing/unfreezing layers impact training?
> A: Freezing/unfreezing layers impacted training. When some layers were frozen, the loss converged faster. When the layers were unfrozen, the loss converged slower.

2.4 Future Improvements

> Q: Would training with more negatives help?
> A: Yes, training with more negatives might help improve performance, because the model can learn the difference between positive and negative samples better.

> Q: Would changing the loss function (e.g., using Softmax Loss) improve performance?
> A: Yes, changing the loss function might help improve performance.

> Q: Could increasing the number of epochs lead to a better model?
> A: Yes, increasing the number of epochs might lead to a better model. In the [anchor, positive] approach, when the number of epochs is 1, the MAP score is 0.4456 and model is not converged yet. When the number of epochs is 4, the MAP score increases to 0.4537.