Balla Aarathisree

aarathisree.1535@gmail.com

Precog Recruitment Task

Task 2: Paper Reading Report

09 – 07 -2024

1 Summary

The research paper addresses BERTSCORE, an automatic evaluation metric for text generation tasks. The model aims at text generation tasks like machine learning translation and image captioning. It holds pre-trained BERT contextual embeddings to compute sentence similarity by comparing the cosine similarity of individual tokens in the candidate and reference sentences. The paper includes evaluation of BERTSCORE on mixed datasets and tasks:

- It manipulates WMT18 metric evaluation dataset, comparing the BERTSCORE with existing metrics like BLEU, METEOR, and RUSE. It emphasizes strong correlation with human judgments and superior model selection performance.
- It evaluates BERTSCORE on COCO 2015 Captioning Challenge dataset, comparing it with task-specific and task-agnostic metrics, including SPICE. The results have left behind the existing metrics, demonstrating its relevancy for the task.
- The BERTSCORE's robustness was tested against challenging examples using PAWS dataset.

Instead of relying on pre-trained Transformer model(as it was not released then), they used publicly available pre-trained models from fairseq library.

2 Strengths of the Paper

BERTSCORE can identify semantically equivalent summaries by accommodating the essential variability in code summarization. It captures the underlying semantic meaning of code, taking the relationships between code elements into account rather than just restricting to the keywords.

Some of the strengths of the paper are as follows:

- ➢ The concept of contextualized embeddings, can better identify semantic equivalence then strictly matching the string.
- ➢ BERTSCORE addresses two major pitfalls in *n*-gram-based metrics. First, failure to robustly match paraphrases. Second, *n*-gram model fails to capture distant

dependencies and ordering information. In BERTSCORE, embeddings are trained to be less sensitive to semantic-critical ordering changes.

- The BERTSCORE was experimented on 363 systems by correlating the score with related metrics to accessible to human judgment.
- Regardless of using a large pre-trained model, BERTSCORE has managed to compute at faster rate, making it appropriate for practical applications during development and testing.
- Conquering the traditional metrics, BERTSCORE aims to choose best performing models in machine translation tasks.
- BERTSCORE is outlined to access different NLP tasks and languages.

3 Weaknesses of the Paper

As nothing is perfect in this world! This paper too highlights few weaknesses that can be technically addressed.

- In few cases, BERTSCORE misleads the evaluation results as it struggles to identify factual errors in candidate sentences.
- It alerts the users to look into the task and language with at most consideration as different BERTSCORE configurations show diversity in performance.
- The paper lacks the idea of potential ethical implications of using BERTSCORE. For instance, there may be situations of exploitation by bad actors to spread false information.
- Due to biases, BERTSCORE may struggle with complex code structures and sensitive code topics,.

4 Suggestions for Improvements

There is always a better scope for improvements. Following are the list of suggestions to make the model more valuable for researchers and practitioners:

- As mentioned in the weaknesses, an initiative to come across the potential ethical implications of using BERTSCORE.
- Choosing better models like RoBERTa, DistilBERT, XLNet, ALBERT, T5, GPT-3.5/GPT-4 that are robust large scale transformers and wiser in performance to utilize best BERTSCORE configuration for specific tasks and languages thereby improving user experience and reliable evaluation.

- By incorporating external knowledge sources i.e knowledge graphs (e.g. Wikidata, DBpedia), hybrid models, fact-checking modules, pre-trained models with external data and cross-referencing systems, a further research should discover ways to improve BERTSCORE's ability to detect factual errors in generated text.