शब्दकोष

Enhancing Idiomaticity Representations in LLM's

Team 25
Amey Choudhary
Aryan Gupta
Sarthak Bansal

Motivation

- Effective representation of idioms is crucial for applications such as sentiment analysis, machine translation, and natural language understanding.
  - For instance, poor automatic translation of an idiom once caused the Israeli Prime Minister to describe the winner of Eurovision 2018 as a ”real cow” instead of a ”real darling.”

- Idioms are conceptual constructs requiring **interactions** between entities in a sentence and **real-world knowledge** to grasp their meaning
  - While large language models like GPT-3 achieve substantial success in general NLP tasks, their performance in idiom comprehension remains subpar, achieving only 50.7% accuracy in understanding idiomatic expressions

Literature Survey Review

- Idiomatic expressions are non-compositional, meaning their figurative meanings cannot be inferred from individual words, making them challenging to represent in traditional models.

- Approaches to address idiomaticity include :
  - Combining compositional components with adaptive weights .
  - Creating phrase embeddings that capture both compositional and idiomatic meanings .
  - Using idiom-specific tokens .

Literature Survey Review

- Intrinsic evaluations measure how well model embeddings align with human judgments, using datasets like AStitchInLanguageModels and semantic similarity tasks.

- Extrinsic evaluations assess the impact of idiomatic representations on downstream tasks such as machine translation, sentiment analysis, and conversational systems.

- SemEval-2022 Task 2B provides a multilingual evaluation framework that uses Semantic Textual Similarity (STS) scores to evaluate the semantic understanding of idioms across languages.

- Advanced methods, such as idiomaticity-specific loss functions and contrastive learning frameworks, further improve the robustness of idiomatic representations by optimizing models to distinguish between literal and idiomatic meanings.

Task Description

- The task involves ranking five different captions based on how well they represent the meaning of the nominal compound in a given query sentence.
  - Nominal compound can be used in both literal and idiomatic sense
  - For e.g. query sentence ”Kapil ate a bad apple and felt disgusted” contains literal meaning of the compound word (”bad apple”) while the query sentence ”The students in this class are good except for a few bad apples” uses the compound word (”bad apples”) in an idiomatic sense.

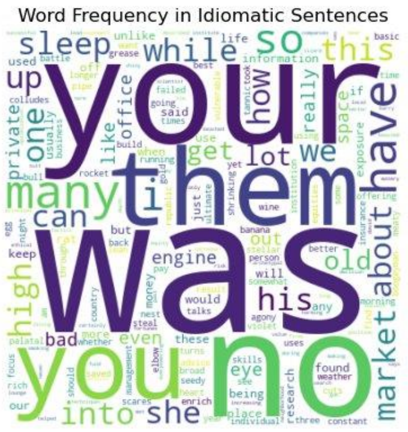
- This project focuses on models that process textual information to interpret idiomatic expressions effectively. We evaluate the comprehension abilities of these models and explore approaches to enhance their performance.

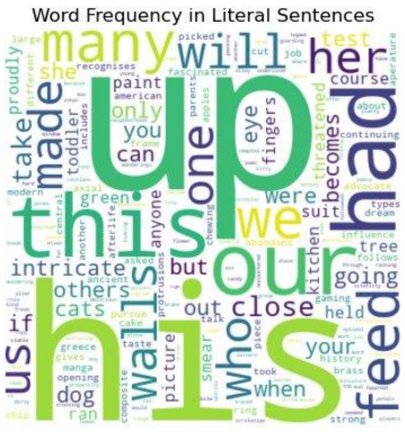
Dataset Exploration

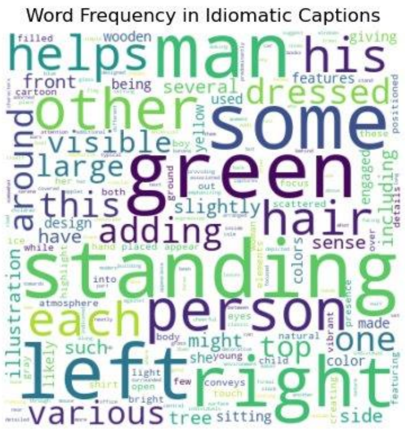
- The dataset contains 70 samples, each with a query and five captions.
- The captions are of :
  - A synonym for the idiomatic meaning of the NC.
  - A synonym for the literal meaning of the NC.
  - Something related to the idiomatic meaning, but not synonymous with it.
  - Something related to the literal meaning, but not synonymous with it.
  - A ‘distractor’, which belongs to the same category as the compound (e.g. an object or activity) but is unrelated to both the literal and idiomatic meanings.

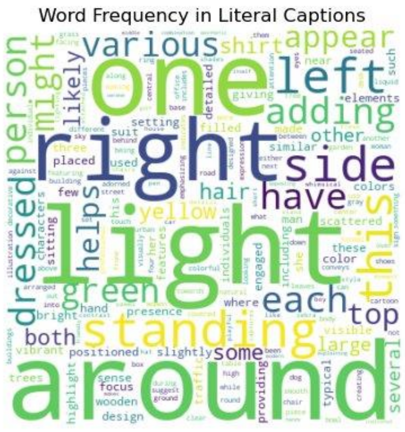
Dataset Exploration

- Here are a few notable insights :
  - Idiomatic sentences contain more unique words.
  - The length of sentences for both idiomatic and literal sentences tends to be almost the same.


Word Frequency in Idiomatic Sentences


Word Frequency in Literal Sentences


Word Frequency in Idiomatic Captions


Word Frequency in Literal Captions

Baseline Approach

- Our baseline model is Sentence BERT (*sentence-transformers/all-MiniLM-L6-v2*), pre-trained on semantic similarity tasks, which provides robust dense sentence embeddings.
- We generated dense vector representations for the query sentence and its corresponding captions, then computed the cosine similarity between each caption and the query sentence.
- Based on the similarity scores, we ranked the captions.

- However , as Sentence BERT is not trained to capture interactions between words in non-compositional idioms , it purely ranks captions based on the general knowledge of English .

Methodology

- We divide our task into 3 major components:
  - **Data Augmentation** : To highlight the representation of nominal compound in both the captions and query sentence.
  - **Loss Function** : Defines the training objective to optimize the model's ability to differentiate between captions effectively based on their idiomatic similarity to a query sentence, ensuring a focus on idiomatic alignment.
  - **Fine-Tuning Models**: Refines BERT's pre-trained knowledge to specialize in understanding and ranking idiomatic expressions by adapting its internal representations through task-specific training.

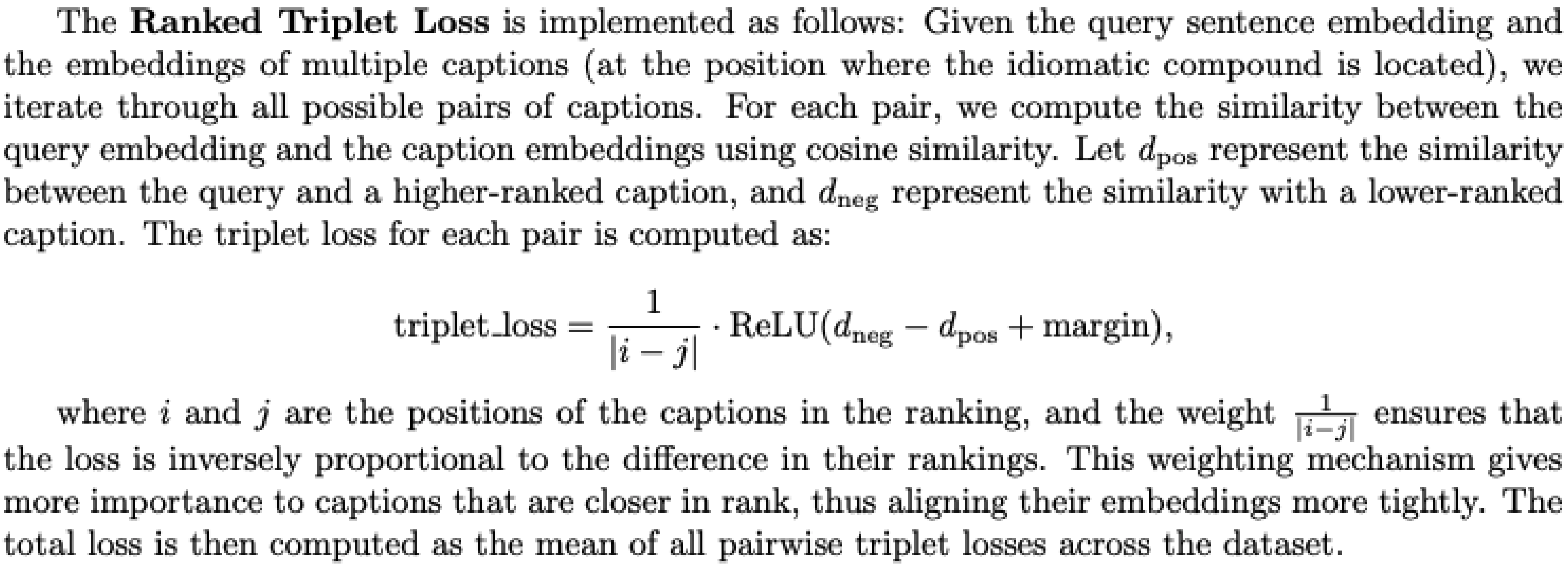
Data Augmentation

- We aim for the embeddings to focus on the position where the idiom is used rather than the entire sentence.
- To enforce this within the loss function, we additionally highlight the idioms by placing [SEP] tokens around the nominal compound.
- This explicit marking allows the model to differentiate the idiom from the rest of the sentence, thereby improving its ability to interpret the idiomatic usage in context.

Loss Function

- To enhance the model's understanding of idiomatic usage, we designed the **Ranked Triplet Loss**, a convex objective function that aligns model-generated embeddings with ground truth rankings, ensuring a clear distinction between higher- and lower-ranked captions.
- Instead of using embeddings from the entire sentence, we extract embeddings specifically from the idiomatic compound's position, allowing the model to focus on the context and meaning of the idiom for improved semantic interpretation.

Loss Function



The **Ranked Triplet Loss** is implemented as follows: Given the query sentence embedding and the embeddings of multiple captions (at the position where the idiomatic compound is located), we iterate through all possible pairs of captions. For each pair, we compute the similarity between the query embedding and the caption embeddings using cosine similarity. Let d_{pos} represent the similarity between the query and a higher-ranked caption, and d_{neg} represent the similarity with a lower-ranked caption. The triplet loss for each pair is computed as:

$$\text{triplet\_loss} = \frac{1}{|i-j|} \cdot \text{ReLU}(d_{\text{neg}} - d_{\text{pos}} + \text{margin}),$$

where i and j are the positions of the captions in the ranking, and the weight $\frac{1}{|i-j|}$ ensures that the loss is inversely proportional to the difference in their rankings. This weighting mechanism gives more importance to captions that are closer in rank, thus aligning their embeddings more tightly. The total loss is then computed as the mean of all pairwise triplet losses across the dataset.

- Therefore, the embeddings are optimized to reflect rank-based distances. Captions with higher rank are closer to query sentence, than lower ranked captions.

Fine-Tuning Models

BERT is powerful for generating context – aware embeddings, but full fine-tuning can be impractical due to its complexity. To adapt BERT to our task of interpreting idiomatic expressions, we employ the following fine-tuning approaches:

1. LoRA

Low-Rank Adaptation (LoRA) fine-tunes BERT efficiently by adding trainable low-rank matrices into transformer layers, adapting the model with minimal computational overhead. This method is particularly effective when capturing **attention updates** is essential , which can be useful for our task.

Fine Tuning Models

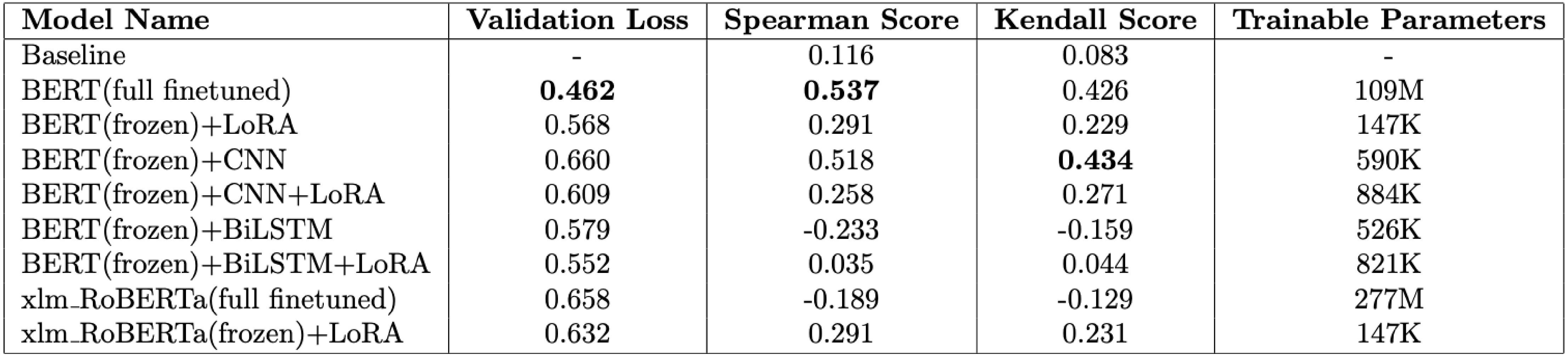
2. CNN
Convolutional Neural Networks (CNNs) are used to capture local dependencies and n-gram patterns within BERT-generated embeddings. By applying convolutional filters, CNNs **identify critical features for interpreting idiomatic expressions**, focusing on subtle text patterns essential for meaning extraction.

3. Bi-LSTM
Bidirectional Long Short-Term Memory (BiLSTM) networks are incorporated to model sequential and contextual dependencies by processing input in both directions. This bidirectional approach allows the model to understand idiomatic usage more effectively, **as it considers the full sentence context and sequential relationships between words**.

Evaluation Metrics

- **Spearman Score**: The Spearman rank correlation coefficient (ρ) is a non-parametric metric that measures the strength and direction of the association between two ranked variables, assessing the overall alignment of ranks using a monotonic function. It considers all pairs and evaluates global rank consistency, with a range from −1 (perfect negative correlation) to +1 (perfect positive correlation).
- **Kendall Score**: The Kendall rank correlation coefficient (τ) evaluates the similarity in the orderings of data by counting concordant and discordant pairs. It highlights local rank relationships and is more sensitive to changes in adjacent or nearby item ranks, with a range from −1 to +1.
- **Comparative Analysis of metrics**: Spearman's ρ focuses on global rank consistency and is less sensitive to small changes in rank order, making it suitable for assessing overall alignment. In contrast, Kendall's τ emphasizes local rank relationships and captures discrepancies in the ordering of adjacent items, providing a finer-grained analysis. While both metrics measure rank correlation, Spearman's ρ is better for general consistency, while Kendall's τ is more responsive to local changes in rank.

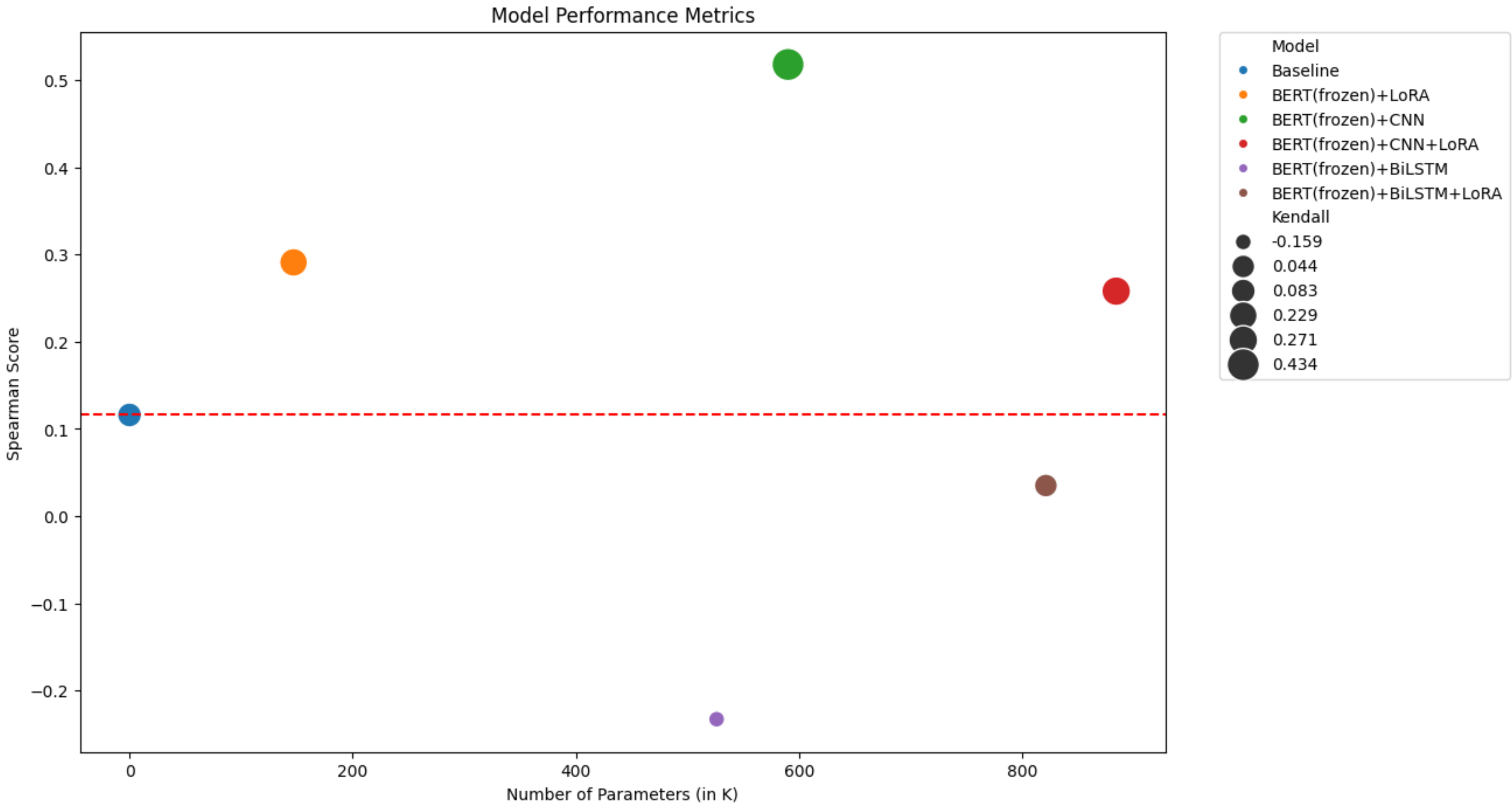
Results



Model Name	Validation Loss	Spearman Score	Kendall Score	Trainable Parameters
Baseline	-	0.116	0.083	-
BERT(full finetuned)	**0.462**	**0.537**	0.426	109M
BERT(frozen)+LoRA	0.568	0.291	0.229	147K
BERT(frozen)+CNN	0.660	0.518	**0.434**	590K
BERT(frozen)+CNN+LoRA	0.609	0.258	0.271	884K
BERT(frozen)+BiLSTM	0.579	-0.233	-0.159	526K
BERT(frozen)+BiLSTM+LoRA	0.552	0.035	0.044	821K
xlm_RoBERTa(full finetuned)	0.658	-0.189	-0.129	277M
xlm_RoBERTa(frozen)+LoRA	0.632	0.291	0.231	147K

Table 1: Model Performance Metrics : Best results are highlighted in **bold**

Comparative Analysis



Discussion

- The fully fine-tuned BERT model exhibited the best performance, indicating that BERT's contextual embeddings are highly effective in capturing idiomatic interpretability. However, the significant computational and memory overhead associated with fine-tuning all 109 million parameters makes it impractical for real-world applications where efficiency is a priority.

- The BERT (Frozen) + CNN configuration demonstrated competitive performance, outperforming most other fine-tuning approaches while maintaining a much smaller number of trainable parameters. This higlights that capturing subtle text patterns and capturing intricate features adds improvement to idiomatic representations.

Discussion

- In contrast, the BERT (Frozen) + BiLSTM configuration underperformed, even scoring lower than the baseline. This result can be attributed to the sequential modeling of BiLSTM introducing unnecessary noise into the task. This suggests that idiomatic usage often relies on precise contextual information rather than long-range dependencies

- Surprisingly, LoRA did not significantly improve the performance of the base model in most configurations, including BERT and CNN. However, it showed some improvement when paired with BiLSTM and XLM-RoBERTa, though these results were still relatively poor compared to other approaches. This suggests that updating attention mechanisms alone may play a limited role in capturing idiomatic information. While LoRA does not produce entirely negative results, its impact on creating agreeable rankings is minimal, indicating that idiomatic interpretation relies more heavily on the base embeddings and local contextual patterns than on fine-grained attention updates.

Conclusion

- In this work, we focus on developing models that can effectively capture idiomatic usage in a sentence. To this extent, we designed a unique loss function, tried simple data augmentation, and experimented with various fine-tuning techniques found in the literature.

- We observe that capturing fine-grained "**contextual interactions"** is most important for this task, which BERT is able to achieve effectively when fully finetuned.

- However, lighter alternatives like CNNs exist which appear as strong practical contenders.

Future Works

- Expand the model to be multi-lingual as per the availability of data.
  - Use multi-lingual bert as base and finetune .

- Add multi–modality given the images.
  - Use CLIP features to represent the data in same embedding space .

- Enhance idiomatic representations further by training on subtask-B of semeval – 2025 .

Other Ablations

- We performed the following additional ablations:

Model Name	Validation Loss	Spearman Score	Kendall Score	#Trainable Params
RoBERTa (full finetuned)	0.658	-0.189	-0.129	277M
BERT+LoRA (uniformly weighted loss)	0.804	0.041	-0.0209	147K
BERT+CNN (multilayer)	0.591	0.518	0.434	4M