SemEval-2025 Task 1

AdMIRe: Advancing Multimodal Idiomaticity Representation

Viraj Surana
Department of Data Science and Artificial Intelligence
22BDS065
Dharwad, Karnataka

***Abstract*—** Effective and accurate representation of non-compositional language is crucial to avoid interpretation errors being propagated to downstream tasks. To evaluate to what extent recent advances in language modelling have improved their ability to identify and interpret non-compositional language and to encourage advances in this area, this task presents the challenge of idiomaticity representation using multimodal data. This task consists of the following subtasks:
(A) identifying which of several images best represents an idiomatic expression as it is used in a given sentence, and (B) selecting the best completion for a 3-image sequence representing the meaning of a given expression. The data consists of text sentences involving idiomatic expressions and images depicting these expressions. This is a follow-up to SemEval 2022 Task 2 which focused on text, but with substantial advances in foundational language models, it is time for more challenging tasks that target semantic understanding in multiple modalities; in this case, static and temporal visual depictions

Introduction (About Problem Statement)

Overview

Idioms are multi-word expressions (MWEs) with meanings that often can't be derived from the individual words composing them, posing a challenge for NLP models. For example, "eager beaver" doesn't describe an actual animal but a keen person. This ambiguity between literal and idiomatic meanings makes idioms a valuable test for assessing how well NLP models capture meaning. Since idiom understanding often involves real-world, multisensory interactions, we aim to build on the SemEval-2022 Task 2 by exploring multimodal models that incorporate both visual and textual inputs to improve idiom representation. Effective idiom understanding can benefit tasks like sentiment analysis, machine translation, and natural language understanding, where poor interpretations may lead to significant errors, as when an idiom was mistranslated, calling the Eurovision 2018 winner a "real cow" instead of a "real darling." This research seeks to advance the NLP field in idiomaticity representation through enhanced multimodal approaches.

SUBTASK A: STATIC IMAGES

In Subtask A, participants will be presented with a set of 5 images and a context sentence in which a particular potentially idiomatic nominal compound (NC) appears. The goal is to rank the images according to how well they represent the sense in which the NC is used in the given context sentence. In order to reduce potential barriers to participation, we also provide a variation of the task in which the images are replaced with text captions describing their content. Two settings are therefore available for the subtask; one in which only the text is available, and one which uses the images.

Methodology

A. Model Building

Model building is the process of creating a mathematical or computational representation of a system or phenomenon to analyze, predict, or understand its behavior. In data science and machine learning, model building involves selecting algorithms, training the model on data, tuning parameters, and validating performance to make accurate predictions or uncover patterns. This process is iterative, often requiring adjustments based on model performance metrics and new data, and is central to developing solutions in various fields, from business forecasting to scientific research.

1) Transfer Learning

Transfer learning is a machine learning technique where a model developed for a particular task is reused as the foundation for a model on a second, related task. This approach leverages knowledge gained from solving one problem to improve the performance of another task, particularly when the latter has limited data. In transfer learning, a pre-trained model, typically developed using a large dataset and extensive computational resources, is fine-tuned or adapted to address a specific task by retraining on a smaller dataset.

The main advantage of transfer learning is its ability to reduce the computational cost and data requirements for training models. Instead of training a new model from scratch, which can be time-consuming and data-intensive, transfer learning allows the reuse of already-learned features from models trained on related data.

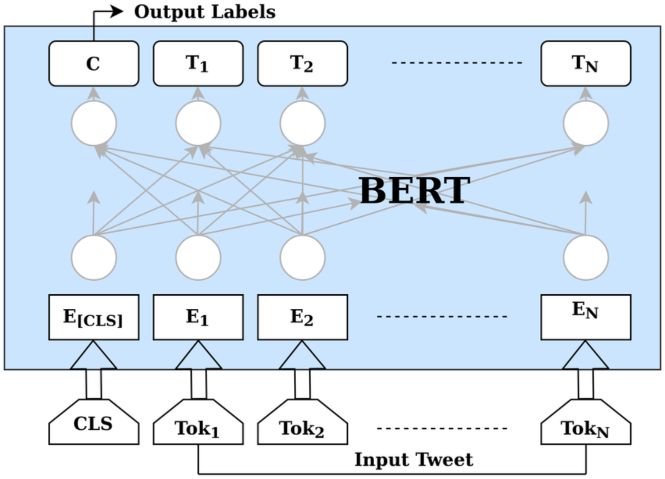
This technique is especially useful in applications such as natural language processing (NLP) and computer vision, where pre-trained models like BERT, GPT, and ResNet-50 can be fine-tuned for various specialized tasks. To summarize, transfer learning accelerates the development of accurate models for related tasks by building on pre-existing knowledge, enabling effective solutions even with limited training data.

2) BERT (Bidirectional Encoder Representations from Transformers) Model

BERT (Bidirectional Encoder Representations from Transformers) is a powerful language model developed by Google in 2018, transforming natural language processing. Unlike prior models that read text in one direction, BERT reads bidirectionally, enabling it to understand the full context of each word based on surrounding words. This bidirectional approach allows BERT to excel in context-sensitive tasks like question answering and sentiment analysis.

Built on Transformer architecture, BERT uses attention mechanisms to weigh word relevance, achieving high performance on NLP benchmarks. Pre-trained on extensive text, BERT learns through masked language modeling (predicting masked words in a sentence) and next sentence prediction (determining logical sentence order). BERT's versatility and success paved the way for subsequent transformer-based models like GPT and T5, further advancing NLP capabilities.

We used "BERT model" for encoding the text present in the captions of the images. We extracted contextual embeddings for sentences with idioms to capture nuanced meanings. BERT's output provides a dense vector of 768 dimensions per token, averaged to create sentence-level embeddings.



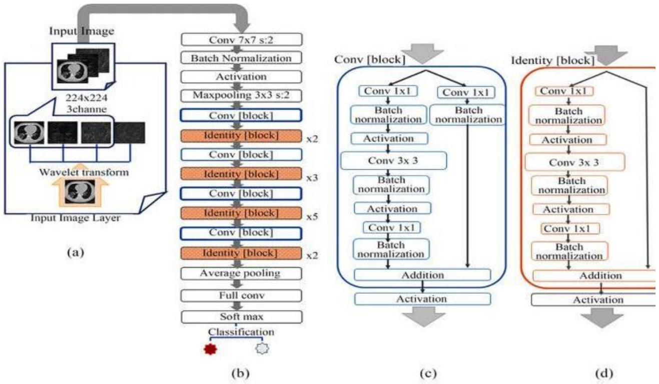
BERT model architecture

3) RESNET-50

ResNet-50, short for Residual Network with 50 layers, is a deep convolutional neural network designed to address the problem of vanishing gradients in deep networks, which often impedes effective learning. Introduced by researchers at Microsoft, ResNet-50 is built on a concept known as residual learning, where shortcut connections are introduced to bypass certain layers in the network. This allows the model to "skip" a few layers during training, making it easier to propagate information and gradients backward through the network. ResNet-50 specifically consists of 50 layers, combining both convolutional layers and identity blocks that leverage these residual connections, which helps prevent degradation of performance as the network depth increases.

ResNet-50's functionality has been widely adopted for various applications beyond image classification, including object detection, segmentation, and even transfer learning, where pretrained weights on a large dataset like ImageNet are fine-tuned for specific tasks with smaller datasets. The model's efficiency in learning deep representations has made it a popular choice in computer vision, enabling advancements in fields like medical imaging, facial recognition, and autonomous driving.

We used "RESNET-50" for encoding the 5 images present for each data point the sample data. Since it is pretrained on ImageNet, it transforms images into 2048 dimensional feature vectors, capturing visual features relevant to idiomatic representation.



RESNET-50 model architecture

4) Loss Functions

Loss functions are essential components in machine learning and deep learning algorithms, acting as a measure of the difference between the predicted values and the true values. They quantify how well or poorly a model is performing during training, guiding the optimization process to minimize the error through techniques like gradient descent. By minimizing the loss function, the model learns to make better predictions. There are various types of loss functions suited for different types of machine learning tasks, such as regression, classification, or ranking.

Mean Squared Error (MSE): It is a commonly used loss function for regression tasks. It calculates the average squared difference between the predicted values and the actual target values.

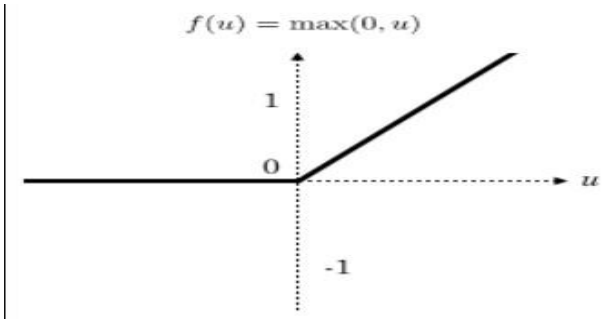
The formula for MSE is given by:

$$\mathrm{MSE} = \overset{\text{Mean}}{\boxed{\frac{1}{n}\sum_{i=1}^{n}}}\Big(\overset{\text{Error}}{\boxed{Y_i - \hat{Y}_i}}\Big)^{\overset{\text{Squared}}{\boxed{2}}}$$

where, Yi represents the actual value, Yi^ is the predicted value, and n is the number of data points. MSE penalizes larger errors more than smaller ones due to the squaring of differences, which encourages the model to reduce large deviations.

5) Activation Function

Activation functions introduce non-linearity in neural networks, enabling them to learn complex patterns. Without them, the network would behave like a linear model, limiting its ability to capture intricate data relationships. One popular activation function is ReLU (Rectified Linear Unit), defined as f(x)=max(0,x) f(x) = \max(0, x)f(x)=max(0,x), which outputs the input for values ≥0 and zero for negative inputs. ReLU is computationally efficient, mitigates the vanishing gradient problem, and is effective in sparse data scenarios.



Despite issues like the "dying ReLU" problem, ReLU and its variants (e.g., Leaky ReLU) remain fundamental in deep learning architectures.

6) Optimizer

Adam Optimizer

The Adam (short for Adaptive Moment Estimation) optimizer is a widely used algorithm in deep learning due to its efficiency and adaptability. It combines the advantages of two other popular optimizers: AdaGrad and RMSProp. Adam computes adaptive learning rates for each parameter by considering both the first moment (mean) and the second moment (uncentered variance) of the gradients. Specifically, it maintains two moving averages for each parameter: one for the gradients and one for the squared gradients. These moving averages are used to adaptively scale the learning rate for each parameter, allowing the optimizer to take larger steps in areas where the gradients are small and smaller steps where the gradients are large. This makes Adam particularly useful for problems involving large datasets or complex models, as it often converges faster and requires less memory than other optimizers. Moreover, Adam is less sensitive to hyperparameters such as the learning rate and is capable of handling sparse gradients, making it a robust choice for a wide range of machine learning tasks**.**

Workflow of Model

The model workflow involves the fusion of textual and visual data to predict the idiomaticity ranking of idioms based on sentence contexts and accompanying images. The following key stages outline the model's processing flow:

i. Data Loading and Preprocessing*:*

The model ingests sentence data and image sets linked to idiomatic expressions. Each idiomatic phrase in the dataset is represented by a sentence and a set of images.

Text is tokenized using BERT’s tokenizer, ensuring uniform length across sequences by truncating or padding to a maximum length of 128 tokens.

Images undergo transformations, including resizing to 224x224 pixels, normalization to ResNet standards, and conversion to tensor format.

ii. Multimodal Feature Extraction:

Textual Embeddings: BERT is utilized to generate contextual embeddings for the sentence. The final hidden state's mean is taken to represent the text features, which are then projected to a lower-dimensional embedding of 128 dimensions.

Visual Embeddings: Each image passes through a pre-trained ResNet-50 model with a modified output layer, producing a 2048-dimensional feature vector. This vector is subsequently projected into a 128-dimensional space to align with text embeddings.

iii. Feature Fusion and Ranking Prediction:

Text and image embeddings are combined by replicating the text feature vector across all image features within a batch. This produces paired text-image embeddings, capturing the multimodal interactions for each idiom.

These paired embeddings are concatenated and passed through a fully connected neural network to produce a ranking score for each image relative to its idiomatic expression in context.

iv. Training and Optimization:

The model is trained to minimize the Mean Squared Error (MSE) loss between predicted rankings and expected order values derived from the dataset.

Using the Adam optimizer, the model parameters are iteratively updated based on backpropagation of the computed loss, aiming to align the model's ranking predictions closely with target idiom rankings.

v. Evaluation:

During evaluation, the model generates predicted rankings for test samples. The predicted rankings are adjusted to align with a user-friendly format and are assessed for accuracy in predicting idiom ranks based on the context provided.

vi. Final Output:

After training and evaluation, the model's weights are saved for potential reuse. The predictions are returned in rank-ordered lists for each test instance, providing an interpretable ordering of images in alignment with idiomatic meaning.

This workflow enables a seamless integration of textual and visual modalities, leveraging BERT for text processing and ResNet-50 for visual feature extraction, which together facilitate nuanced ranking predictions for idiomaticity tasks.

Result

In evaluating the multimodal idiomaticity ranking models, Mean Reciprocal Rank (MRR) was employed as the primary metric, reflecting the models' effectiveness in correctly ranking idiomatic expressions based on contextual relevance.

Three variations of the model were trained and evaluated:

- **Model 1** achieved an MRR score of **0.2940**, establishing a baseline performance with limited ranking accuracy.
- **Model 2** improved substantially, reaching an MRR score of **0.5167**, indicating stronger alignment with target idiom rankings.
- **Model 3** yielded the highest MRR score of **0.5321**, highlighting further enhancement in the model's ranking precision and multimodal understanding.

The progressive improvement across models suggests that adjustments in the model's architecture and training dynamics were effective in enhancing the model's ability to interpret and rank idioms accurately.

Conclusion

The results demonstrate the potential of multimodal learning for idiomaticity ranking, with Model 3 showing the highest accuracy. Future work could further refine these results through hyperparameter optimization and more extensive multimodal integration

Code Github

Reference

[1] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[3] H. Benbrahim, J. El Asri, and H. Qjidaa, "Multimodal sentiment analysis: A survey," *Procedia Comput. Sci.*, vol. 192, pp. 3285–3295, Jan. 2021, doi:10.1016/j.procs.2021.09.092.

[4] A. Vaswani, N. Shazeer, N. Parmar, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*. [Online]. Available: https://arxiv.org/abs/1312.6114