Boundary Of Meaning

Mattia Vinciguerra

Department of Computer Science, University of Milan.

Abstract

This study investigates how literal and figurative language are represented in the embedding space of transformer-based language models, such as BERT and RoBERTa. In particular, this project aims to explore the separability of literal and figurative expressions in the latent space and to determine whether a semantic boundary can be identified. Latent representations are obtained using contextualized embedding techniques, and their analysis is performed using clustering and classification methods. Qualitative results are obtained using dimensionality reduction techniques, while quantitative results come from measurable outcomes that help to evaluate and compare different models and contextual embedding types.

GitHub repository

1 Introduction

Human language operates on multiple levels of meaning, ranging from literal descriptions to figurative expressions such as metaphors and idioms. Literal language communicates meaning directly, with words retaining their conventional semantic roles (e.g. *"The cat sleeps on the carpet"*), while figurative language relies on indirect meaning, exploiting cognitive mechanisms like analogy and irony to convey richer or more abstract concepts (e.g. *"He is a volcano of ideas"*). Understanding figurative language is a non-trivial task for both humans and artificial systems, as it requires integrating lexical semantics with contextual and world knowledge ([1], [2]).

The detection and interpretation of figurative language have become a significant research area in Natural Language Processing (NLP). Modern transformer-based language models, such as BERT and RoBERTa, learn contextualized representations of text through large-scale pre-training. These models have demonstrated remarkable

success in a wide range of NLP tasks, but their ability to differentiate literal and figurative meaning remains an open challenge, as the two forms of meaning often overlap in surface structure, while differing in semantic depth and conceptual mapping ([3]).

2 Research questions

The intersection between computational representation learning and figurative meaning raises fundamental questions.

- How do transformer-based language models internally represent literal and figurative expressions?
- Are contextualized embeddings linearly separable?
- Does fine-tuning enhance the separation?
- Can token embeddings capture the differences?
- Can multimodal models achieve better separation?

This study also presents a comparison between different models, contextualized embedding types, dimensionality reduction techniques, and classification methods.

Answering these questions has both theoretical importance for understanding the semantics of LLMs and practical relevance for applications such as metaphor detection, sentiment analysis, and figurative language generation.

3 Methodology

This section describes the methodological framework adopted to investigate how literal and figurative expressions are represented in the embedding spaces of transformer-based language models.

3.1 Models

In this study, three models from the HuggingFace library are used.

BERT (Bidirectional Encoder Representations from Transformers), introduced by Devlin et al. (2019) [4], is a transformer-based language model that revolutionized NLP by using a bidirectional transformer encoder architecture that aggregates both the left and right contexts. It produces contextualized embeddings that capture both syntactic and semantic relationships within a sentence. BERT's architecture allows representation of polysemous words and complex linguistic phenomena, which is particularly relevant when modeling figurative expressions.

RoBERTa (A Robustly Optimized BERT Pre-training Approach), proposed by Liu et al. (2019) [5], is an optimized variant of BERT trained on larger datasets with longer training schedules. These modifications improve the ability to capture semantic relationships and lead to better performance on a wide range of NLP benchmarks.

CLIP (Contrastive Language-Image Pre-Training), developed by Radford et al. (2021) [6], extends transformer-based learning to the multimodal domain by jointly training a text encoder (a transformer-based language model) and an image encoder (a vision transformer) using a contrastive loss. It learns to align textual and visual representations in a shared embedding space. In this work, CLIP is used to determine

whether the introduction of visual information enhances the distinction between literal and figurative expressions.

3.2 Contextualized Embeddings

Transformer-based language models such as BERT and RoBERTa process text by dividing input sentences into tokens (smaller units of text) and converting them into token embeddings: fixed-dimensional vectors that capture the token's identity. After tokenization, the transformers aggregate contextual information taking into account the entire sequence, capturing semantic and syntactic meanings. Contextual embeddings are obtained from the transformer hidden states and can be extracted in different ways. In this project are used three embedding strategies, providing different perspectives on the separability of literal and figurative expressions.

CLS token embedding: a special token (CLS) is added at the beginning of each sequence, its last layer embedding is a summary representation of the sequence.

Average token embedding: average of all token embeddings from the last hidden state.

Layer-wise embedding: aggregation of all token embeddings from some hidden states. This approach captures linguistic information at multiple levels of abstraction, from shallow lexical features to deeper semantic representations.

3.3 Dimensionality Reduction

In order to visualize and analyze high-dimensional embeddings, it is necessary to project them into a lower-dimensional space while preserving structural information. In this work are adopted three dimensionality reduction techniques that allow to explore patterns, clusters, and separability between literal and figurative expressions in the embedding space.

PCA (Principal Component Analysis) is a linear technique that projects data in the directions of maximum variance, providing a global view.

t-SNE (t-distributed Stochastic Neighbor Embedding) is a nonlinear method that preserves local neighborhood structures.

UMAP (Uniform Manifold Approximation and Projection) is a manifold learning technique that preserves local and global structure.

3.4 Clustering and Classification

To analyze the separability between literal and figurative expressions in the embedding space, are applied three clustering and classification methods.

K-Means Clustering is an unsupervised clustering algorithm that partitions data into k distinct clusters, identifying natural groups in the data.

Logistic Regression is a supervised classification algorithm that predicts the probability that an observation belongs to a specific class.

Support Vector Machine (SVM) is a supervised classification algorithm that finds the optimal hyperplane that separates classes in the data.

4 Results and Discussion

This section presents the empirical evaluation of the proposed methodology with the relative qualitative and quantitative results.

4.1 Datasets

In this experiment, three datasets from the HuggingFace library are used to evaluate the separability of literal and figurative expressions in textual and multimodal settings.

- **VUAMC**: textual corpus containing literal and metaphorical sentences.
- **V-FLUTE**: multimodal dataset containing figurative image–claim pairs. It constitutes the figurative part of the multimodal dataset.
- **Flickr8k**: multimodal dataset consisting of images and the corresponding literal captions. It constitutes the literal part of the multimodal dataset.

4.2 Evaluation Metrics

To quantitatively assess the performance of the clustering and classification methods, several standard metrics are used.

- **Adjusted Rand Index**: evaluates clustering quality by comparing predicted clusters with ground truth.
- **Accuracy**: measures the proportion of correctly predicted instances over the total number of instances.
- **Precision**: calculates the proportion of true positive (figurative) predictions among all instances predicted as figurative.
- **Recall**: measures the proportion of true figurative among all actual figurative instances.
- **F1-score**: harmonic mean of precision and recall.

4.3 Experimental Methodology

The experiments are designed to provide both qualitative visualizations, which are dimensionality reduction plots, and quantitative evaluations, which are clustering and classification performances.

These results provide an overview on the figurative meanings understanding of transformer-based language models and also give insights into the effectiveness of different models and embedding types in capturing semantic distinctions. They also allow for a comparison among different dimensionality reduction techniques and different classification methods.

4.3.1 Embeddings Visualization

First, with BERT and RoBERTa, are extracted the contextualized embeddings of sentences in the textual dataset VUAMC. In particular, three embedding types are adopted: CLS embedding, average token embedding, and layer-wise embedding. All these embeddings are reduced to two dimensions using three techniques of dimensionality reduction: PCA, t-SNE, and UMAP. Their plots provide a simple overview of the

distribution of literal and figurative expressions in the embedding space. Fig. 1 shows that PCA is the most expressive technique in this case.

Figs. 2 and 3 shows the BERT and RoBERTa embeddings distribution. It is qualitatively observed that the two models behave in a similar way and that layer-wise embeddings better capture the separation between literal and figurative instances.

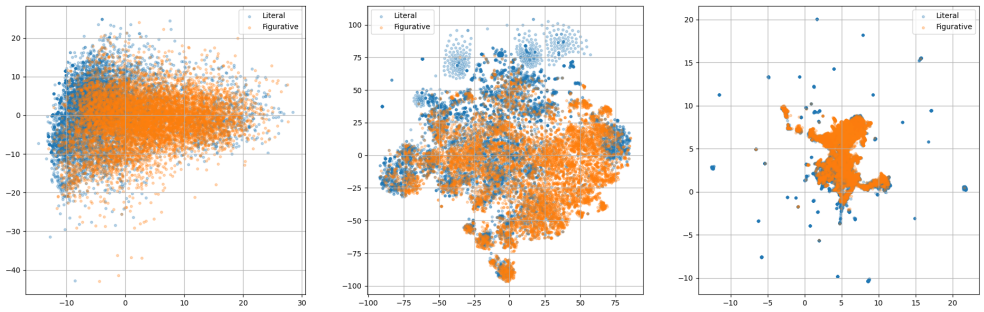


Fig. 1: PCA, t-SNE and UMAP on BERT CLS embeddings.

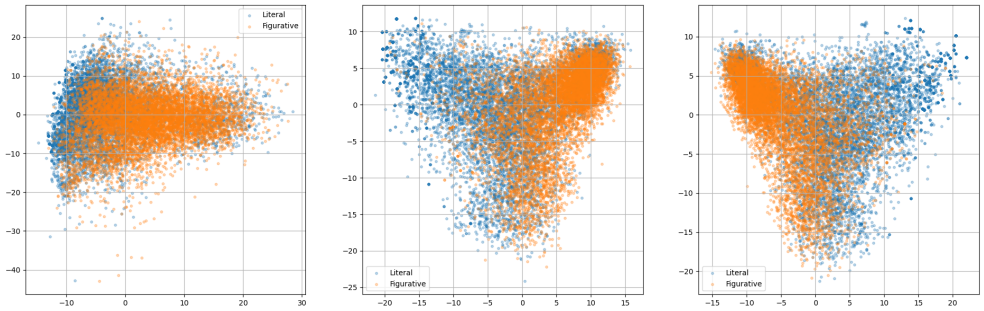


Fig. 2: PCA on BERT CLS, average and layer-wise embeddings.

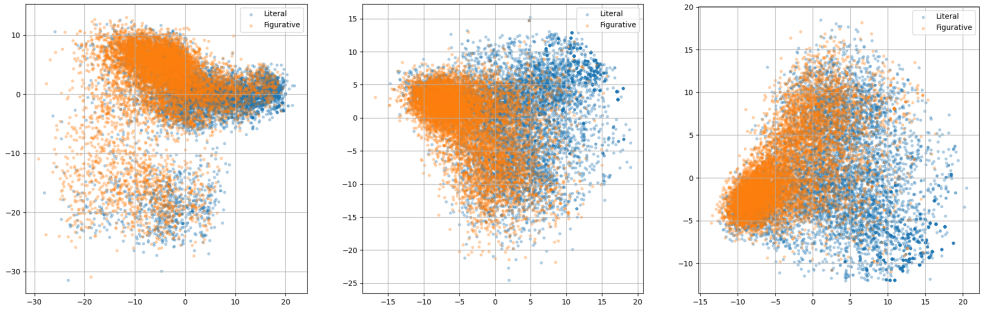
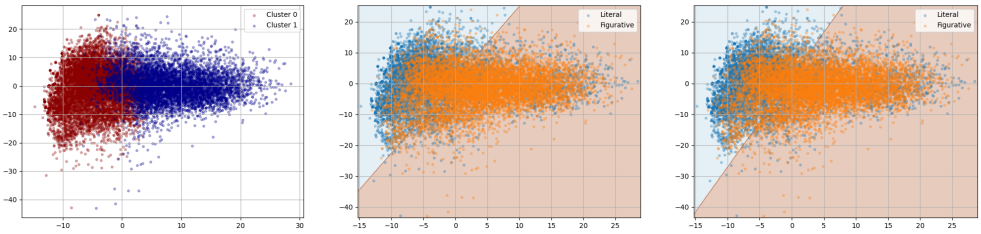


Fig. 3: PCA on RoBERTa CLS, average and layer-wise embeddings.

4.3.2 Clustering and Classification

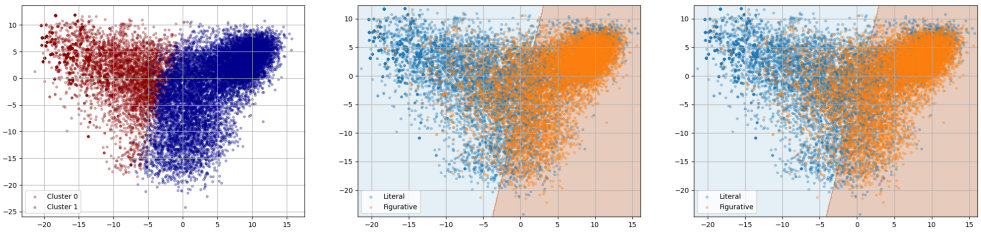
The separability of literal and figurative embeddings is quantitatively assessed using clustering and classification methods. K-Means is applied to group embeddings, while Logistic Regression and Linear SVM are used for linear probing. Are provided both visual plots and measurable outcomes (ARI, accuracy, precision, recall and F1).

Figs. 4, 5 and 6 show the BERT results and present better performances with layer-wise embeddings (this also holds for RoBERTa). This implies that layer-wise embeddings better capture the separation with figurative embeddings forming a quite defined cluster, while literal remaining more sparse. It is also observed that clustering struggles with the separation, whereas supervised algorithms achieve better results. Lastly, there are no large differences between BERT (Fig. 6) and RoBERTa (Fig. 7).



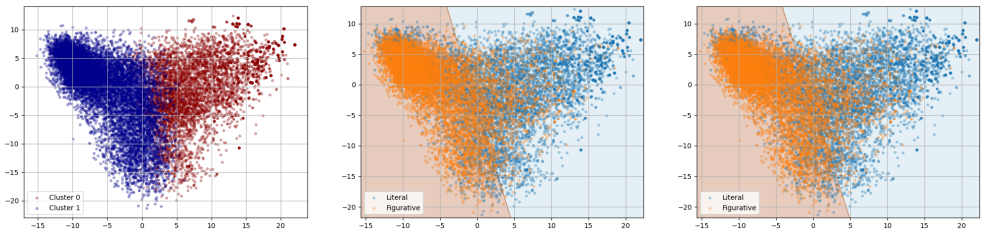
Method	ARI	Accuracy	Precision	Recall	F1
K-Means	0.156				
LogReg		0.803	0.813	0.794	0.803
SVM		0.803	0.814	0.791	0.802

Fig. 4: K-Means, Logistic Regression and SVM results on BERT CLS embeddings.



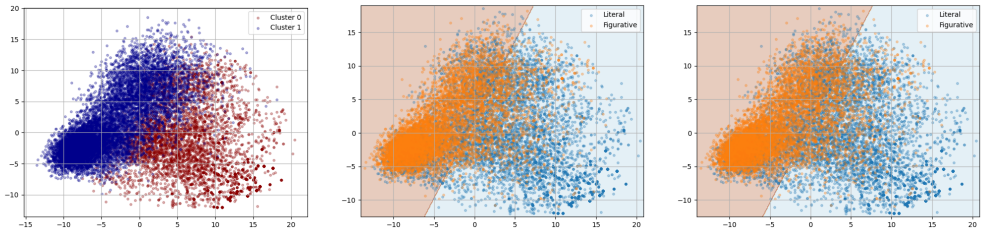
Method	ARI	Accuracy	Precision	Recall	F1
K-Means	0.213				
LogReg		0.825	0.829	0.823	0.826
SVM		0.819	0.82	0.821	0.821

Fig. 5: K-Means, Logistic Regression and SVM results on BERT average embeddings.



Method	ARI	Accuracy	Precision	Recall	F1
K-Means	0.257				
LogReg		0.833	0.835	0.836	0.835
SVM		0.831	0.829	0.839	0.834

Fig. 6: K-Means, LogReg and SVM results on BERT layer-wise embeddings.



Method	ARI	Accuracy	Precision	Recall	F1
K-Means	0.245				
LogReg		0.833	0.833	0.838	0.835
SVM		0.829	0.825	0.841	0.833

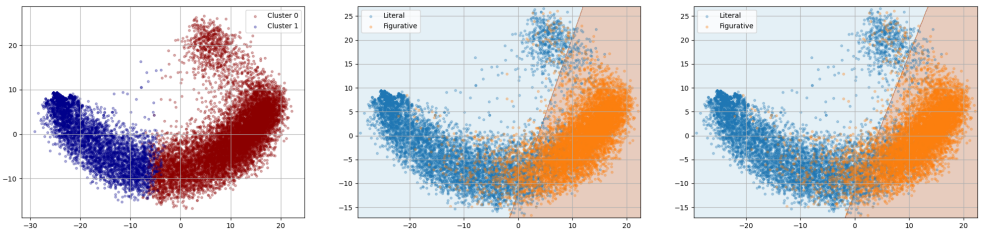
Fig. 7: K-Means, LogReg and SVM results on RoBERTa layer-wise embeddings.

4.3.3 Fine-Tuning

Experiments were also carried out with fine-tuned versions of the models to assess whether task-specific adaptation improves separability in the latent space. BERT and RoBERTa are fine-tuned with a classification task on the same dataset (VUAMC), and the relative clustering and classification results are reported in Figs. 8 and 9. It is observed that they provide slightly improved performances and clearer plots with respect to pre-trained versions.

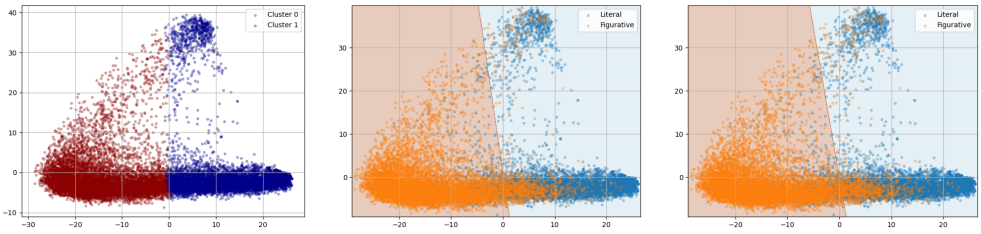
4.3.4 Top Words Analysis

To gain insights into internal representations, it is performed an analysis of the most present words in both literal and figurative meanings. We consider their token embeddings enriched with the context, from the last transformer layer of BERT. This allows investigating whether they are meaningful alone, without their integration in contextual embeddings.



Method	ARI	Accuracy	Precision	Recall	F1
K-Means	0.428				
LogReg		0.863	0.861	0.87	0.865
SVM		0.863	0.865	0.863	0.864

Fig. 8: K-Means, LogReg and SVM results on fine-tuned BERT layer-wise.



Method	ARI	Accuracy	Precision	Recall	F1
K-Means	0.607				
LogReg		0.873	0.88	0.867	0.873
SVM		0.869	0.877	0.863	0.87

Fig. 9: K-Means, LogReg and SVM results on fine-tuned RoBERTa layer-wise.

Some words show a good separation such as *"Found"* in Fig. 10, while some other words present a casual distribution such as *"Back"* in Fig. 11. This underlines the importance of contextual embeddings, that provide a sentence summary information.

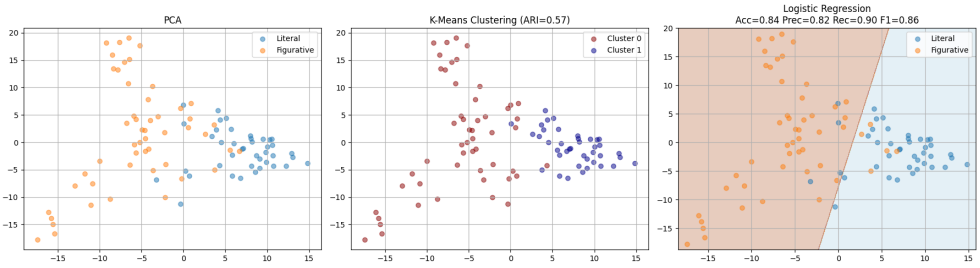


Fig. 10: Token embeddings of *"Found"*.

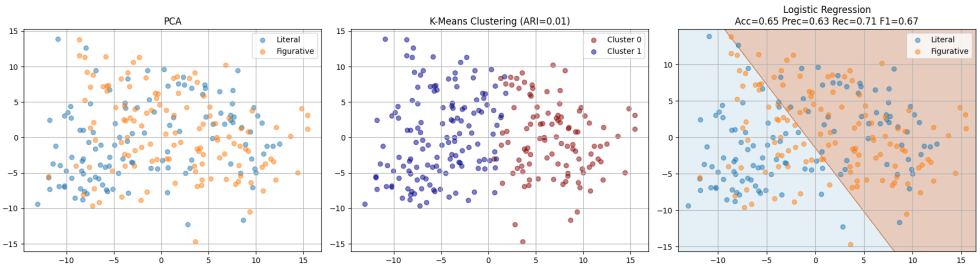
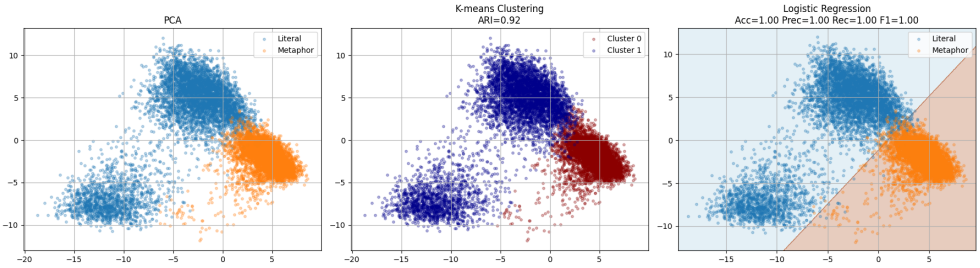


Fig. 11: Token embeddings of *"Back"*.

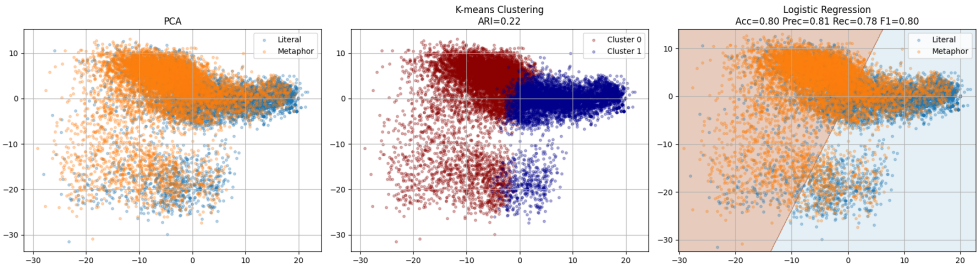
4.3.5 Multimodal Analysis

Finally, a multimodal analysis is performed using CLIP, to evaluate whether visual context enhances the separation of literal and figurative embeddings. For each text-image pair in the combined dataset (V-FLUTE + Flickr8k), the mean between textual embedding and visual embedding is computed. Fig. 12 shows the embeddings distribution and the clustering and classification results. It is obtained a perfect classification and also a very high ARI. The same analysis on the textual part of the dataset (Fig. 13) provides significantly worse results. This confirms that the introduction of visual information can greatly improve the model understanding of figurative expressions.



K-Means (ARI)	Accuracy	Precision	Recall	F1
0.92	1	1	1	1

Fig. 12: Clustering and classification results on CLIP embeddings.



K-Means (ARI)	Accuracy	Precision	Recall	F1
0.22	0.80	0.81	0.78	0.80

Fig. 13: Clustering and classification results on textual part.

5 Conclusion

In conclusion, literal and figurative expressions in the embedding space of BERT and RoBERTa present a sort of separation. Pre-trained versions obtain a good classification with about 83% accuracy and a poor clustering with about 0.25 ARI. Fine-tuned versions obtain little improvements in classification reaching 87% and quite better clustering with an ARI that increases to 0.43 for BERT and 0.60 for RoBERTa.

Contextualized embeddings at sentence level are fundamental to capture figurative meanings, in fact the token embedding analysis does not give meaningful results. Finally, the multimodal analysis with CLIP underlines that visual information is important and can significantly improve the separability of literal and figurative embeddings, reaching 100% of accuracy (+20% with respect to text analysis) and also an optimal clustering (0.92 ARI versus 0.22 in text analysis).

References

[1] Lin, Y., Liu, J., Gao, Y., Wang, A., Su, J.: A Dual-Perspective Metaphor Detection Framework Using Large Language Models (2024). https://arxiv.org/abs/2412.17332

[2] Yang, C., Li, Z., Liu, Z., Huang, Q.: Deep Learning-Based Knowledge Injection for Metaphor Detection: A Comprehensive Review (2024). https://arxiv.org/abs/2308.04306

[3] Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., Lee, J.: MelBERT: Metaphor Detection via Contextualized Late Interaction using Metaphorical Identification Theories (2021). https://arxiv.org/abs/2104.13615

[4] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2019). https://arxiv.org/abs/1810.04805

[5] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (2019). https://arxiv.org/abs/1907.11692

[6] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2021). https://arxiv.org/abs/2103.00020