Enhanced protein function prediction by fusion embedding based on protein language models

Yang Wang *

School of Computer Science, Anhui University of Technology, Anhui, China, 243099

* Corresponding Author Email: yasuo626.sci@gmail.com

Abstract. Natural language models can accomplish non-natural language tasks such as protein prediction, but the actual prediction effect is low and occupies large computational resources. In this paper, a fusion embedding model is proposed to improve the prediction effect of the model and reduce the computational cost of the model by fusing information of different dimensions. The paper is validated by the downstream task of protein function prediction, which provides a reference for solving practical tasks using fusion embedding methods.

Keywords: Protein Function; Language Models; Fusion Embeddings.

1. Introduction

In recent years, large language models have achieved amazing results in the field of natural language processing, and various large models have excelled in language tasks as well as cross-domain tasks. There are many researchers in the field of bioinformatics who have applied them to predictions related to RNA and proteins. These organisms are structurally similar to natural language and can be used to accomplish prediction tasks using natural language models.

In the field of protein prediction, there are already many protein language models dedicated to protein prediction. These include AlphaFold [1] and Protrans [2]. Their various prediction results in the protein domain based on large language models using deep learning methods outperform the original mathematical or machine learning models to some extent.

Although there are many language models that can achieve a certain level of accuracy on protein prediction tasks, the results of these models do not vary much and it is difficult to improve the performance further. Moreover, the computational resources required for natural language modeling tasks are large. In the paper by protrans [2], they use more than thousands of computational resources. The huge computational consumption makes it difficult to specialize these models for specific tasks. Secondly, the generality of the model tasks is poor. Models are limited to predictions using tasks that are consistent with their training tasks.

In this paper, a fusion multi-embedding model is proposed, which makes full use of the large model output feature embedding and fusion methods, and can predict different downstream tasks quickly and effectively through multi-embedding features. Compared with the prediction of individual embeddings, the fused embedding model in this paper shows a significant improvement in performance. And the supported feature embeddings can also be of different dimensions.

In this paper, several protein function prediction tasks are performed on different task dimensions using the fusion embedding model, and the fusion model performs best on all multidimensional tasks compared with other models.

In conclusion, this paper provides a new approach to improve model performance and reduce training costs using large model fusion embedding.

2. Related Work

This section of the paper will present methods for protein function prediction, and related work in protein modeling, fusion embedding.

2.1. GO label

Gene ontology (GO) is a dynamically controlled vocabulary, first proposed by M Ashburner [3], that can be used to label biological functions, where a GO tag corresponds to the concept of a biological function, thus transforming abstract functions into ontologies with unique numbers for representation. There are containment and inclusion relationships among the ontologies, which ultimately constitute a complex network of ontology relationships.

2.2. Protein Language Model

Protein itself is a finite-length sequence of 20 amino acids arbitrarily combined, and the sequence of a protein affects the structure of the protein, while the structure of the protein determines the biological function of the protein. Early protein structure and function prediction mainly used traditional physical models, or simple neural network or sequence models, with low prediction efficiency and accuracy. The later convolutional neural network algorithm [4] improved the prediction accuracy of the model to some extent, but the results were still poor.

After the transformer [5] was proposed, more and more large-scale models based on the transformer were introduced, among which Bert, T5, and XLnet [6-7] were applied more in downstream tasks with better results. These models are more sensitive to sequence data, while proteins are also sequence information objects, and natural language models are also applicable to protein languages.

When LLMs were introduced soon after, they were applied on a large scale to protein prediction in the protein domain [8-9] and related applications [10]. These models reached the state-of-the-art in protein structure prediction. It made progress in protein prediction.

Two years later, AlphaFold developed by DeepMind successfully utilized deep learning to achieve unprecedented performance in protein structure prediction, but protein structure prediction still takes days and requires sufficient computational resources. In addition, AlphaFold is a specialized model for structure prediction and cannot be used directly for function prediction.

2.3. Fusing Embedding

The sequence of proteins inherently exists in the special information of proteins, and this paper hopes to be able to directly predict the biological function of proteins by the sequence information of proteins in a simpler way, so as to simplify the process of predicting the function of proteins. Therefore, this paper inevitably thinks of utilizing the embedding of the existing model for fusion to improve the accuracy.

At present, there is no research on fusion through the output embedding of language models, and the main fusion embedding research is mostly for multimodal and text embedding representation. For example, Huang, S.-C used the fusion embedding of text and image to improve modal or overall prediction [11], Hirasawa, T used multimodal embedding to improve machine translation [12], Lange, L they proposed FAME [13], which achieves the performance enhancement by fusing the additional information of different dimensions and textual information in a weighted manner.

In this paper, a fusion method dedicated to language model embedding is proposed to improve the functional prediction of proteins.

2.4. Multi-Embedding Classification Model

In this section, this paper introduces a multi-label classification model for a single embedding in a downstream task as shown in Figure 1, and a classification model using multiple embeddings as shown in Figure 2. Among the multi-embedding classification models, this paper constructs a transformer-based feature compression network and a fusion embedding network implemented in this paper itself. For the current downstream task of language model the commonly used network is a direct classification using a fully connected neural network that uses a sigmoid activation function to obtain the confidence level for each label.

2.4.1. Transformer Fusion Embedding

In this paper, other fusion embedding methods are investigated and the transformer performs very well in fusion embedding in domains such as text images. Therefore, therefore, this paper uses converters for embedding fusion and compression at the beginning, uses converters for sequence feature fusion without position information, and classifs the output through the last layer.

2.4.2. Fusional Multi Embed Model

Due to the differences in the structure of the language model and the pre-training task, the meaning of the embedding features obtained from different models for the same sequence data varies in different dimensions, but these features do contain the original information of the sequence data. The embeddings of different models are stacked to obtain a two-dimensional feature matrix, where different positions in the matrix correspond to partial information of the sequence object.

To be able to obtain more accurate and complete feature information from the feature matrix, this paper uses an image-like feature extraction method to extract the local information combined into a new feature matrix by an operator kernel, and adds regularization constraints. Specifically, assuming that the feature dimension of input in is (d,s), the dimension of output out is determined by the size and number of kernels set in this paper. When a total of m kernels of size (r,c) are extracted from the input to the output, each line of the output is a feature vector extracted by a single kernel. The final out can be expressed as shown in Equation 1.

$$out_{(m,n)} = norm(\sum_{i=0}^{r}\sum_{j=0}^{c} Kernel_{m(i,j)} \times in_{(i,n+j)} + bias_m). \quad (1)$$

Different features are obtained using different operator kernels, and the embedding is finally compressed into a one-dimensional vector after multiple extractions, and the classification is performed by a fully connected network. Of course, this paper can also reduce the dimensionality of the features by controlling the step size.

Unlike the convolutional neural network, this paper does not pool the matrix, because unlike the image matrix which requires feature extraction at different levels, the sequence feature matrix is essentially a fusion of high-dimensional information through the operator kernels, and pooling will only make the embedding information lost.

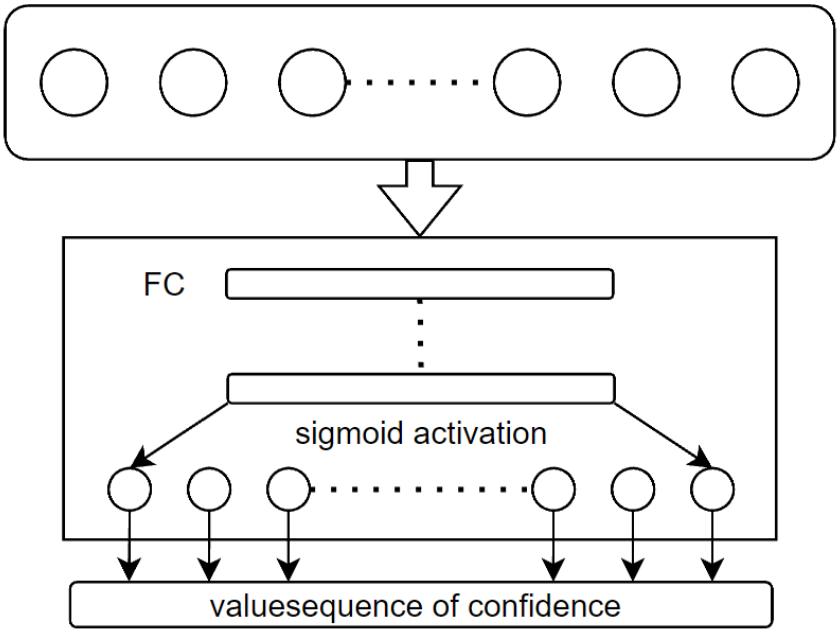


Figure 1. Single embedding model.

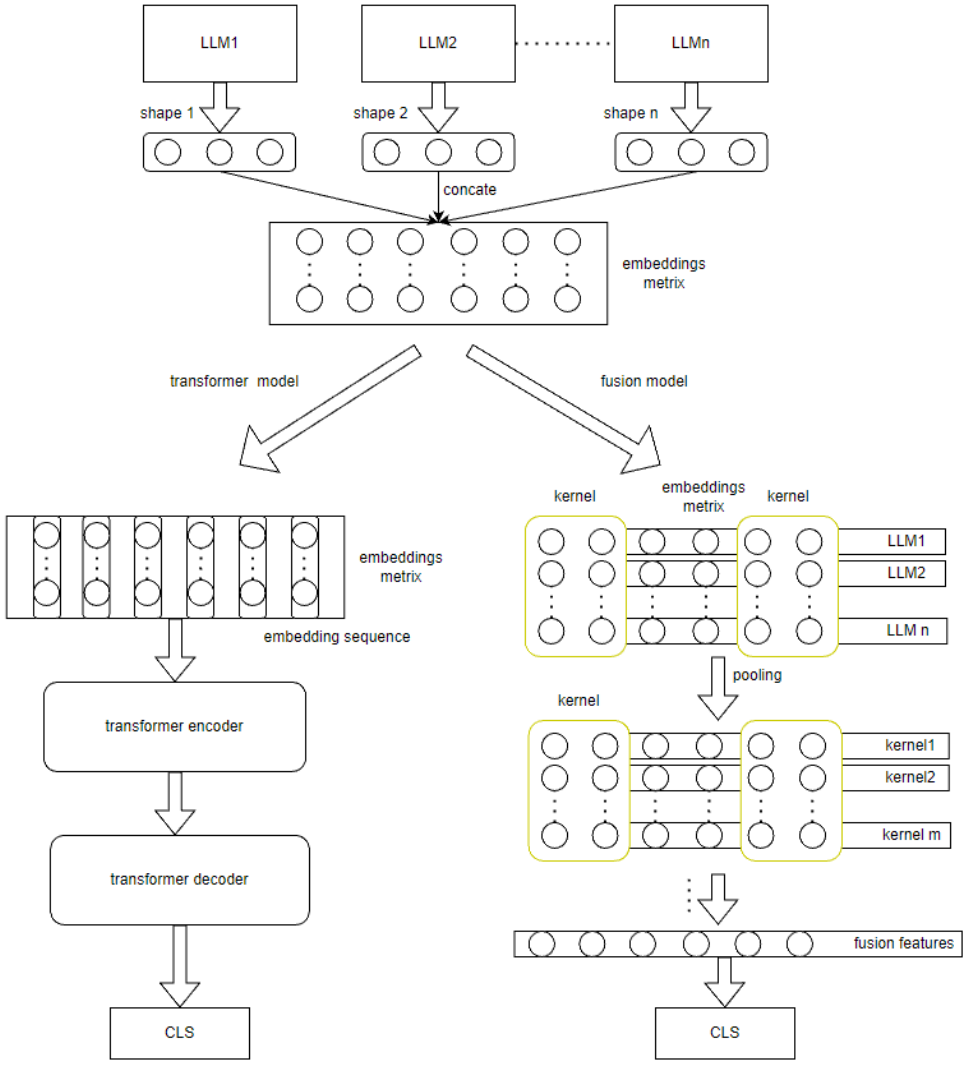


Figure 2. Multi-embedding model.

3. Experiment

3.1. Data Source

The dataset in this paper is taken from UNiprot, and a total of 140,000 proteins are obtained, including protein IDs, sequences and GO genes corresponding to biological functions. The information of the dataset can be referred to Table 1.

Table 1. Basic information about the dataset.

Name	Number	Total number of labels	Average protein length	Maximum length
uniprot	43248	142246	553.6	35375

3.2. GO Label Processing

The main goal of this paper is to predict the biological functions possessed by the corresponding proteins from the embedded information given by the language model. In the previous section it was mentioned that the functions of proteins are composed of a complex network of connections in which each node belongs to a function. Therefore, the ultimate goal of this paper is to dichotomize the nodes in the network, i.e., to determine whether each protein function is owned by that protein.

However, this approach is currently not desirable because the current GO database has more than 50k GO ontologies in total, meaning that this paper needs to use a smaller model for dichotomous classification of 50k tags.

In order to solve the problems of excessive number of labels and invalid labels, this paper selects the optimal combination of labels as a predictive labeling strategy by a comprehensive evaluation of the frequency and weight of the labels.

This is done by converting the number of GO ontologies appearing in all protein samples in the dataset counted in this paper into frequencies and performing a deep search of the GO network starting from the root node, and when the node is located at a deeper level, it will be given a greater weight to indicate a more accurate function. Finally, the importance evaluation factor I is used as the evaluation screening index of the label.

$$I_{id} = \frac{count(id)}{\sum_{id} count(id)} \times w_{id} \quad (2)$$

Table 2. Comprehensive Evaluation Tag.

only frequency		weighted	
Label Name	Weights	Label Name	Weights
GO:0008150	0.000000	GO:0008152	1.598544
GO:0009987	0.589193	GO:0032501	1.655270
GO:0005575	0.000000	GO:0032502	1.684844
GO:0005622	0.366945	GO:0050896	1.568071
GO:0043226	0.584346	GO:0065007	1.153288
GO:0043227	0.134798	GO:0016020	1.824773
GO:0043229	0.017165	GO:0031974	2.539921
GO:0110165	0.025471	GO:0032991	2.479665
GO:0003674	0.000000	GO:0071944	2.157039
GO:0005488	0.454654	GO:0003824	1.634664

By comparing the top 10 labels selected through the frequency and weighting methods in Table 2, it can be concluded that there will be many invalid labels selected only through the frequency method, while the labels selected through the evaluation factors have higher overall evaluation. Finally, the top 300, 500, 1000 labels with the highest importance are selected as the label sets for the three different dimensional tasks.

3.3. Get Embedding

Since this paper separates the inference of the protein language model from the downstream fusion model training task, the embedding needs to be acquired in advance in this paper.

Table 3. Model embedding information.

Model	Parameters	Embedding Dimension
T5	3B	[1,1024]
bert	420M	[1,1024]
bertbfd	420M	[1,1024]
EMS	420M	[1,1080]
albert	224M	[30,1024]

Based on the currently published protein models, the five models given in Table 3 are used in this paper to obtain feature embeddings. The details of the models can be shown by Table 3. These embeddings will be used as the input of the fusion models in this paper to perform downstream task prediction by fusing the embedding models.

3.4. Experimental Design

In the previous paper, embedding data of multiple models and labels of multidimensional tasks have been obtained. In order to verify the enhancement effect of the fusion model, this paper lets the single embedding model and the fusion transformer model as well as the fusion model fusing multiple

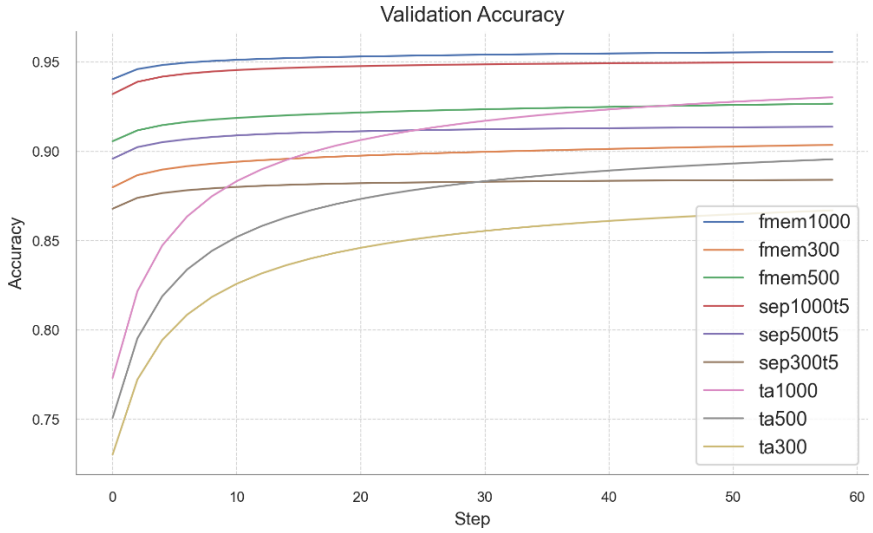
embedding models to predict 300, 500, and 1000 tasks respectively to compare the prediction effect of the models.

3.5. Experimental Results

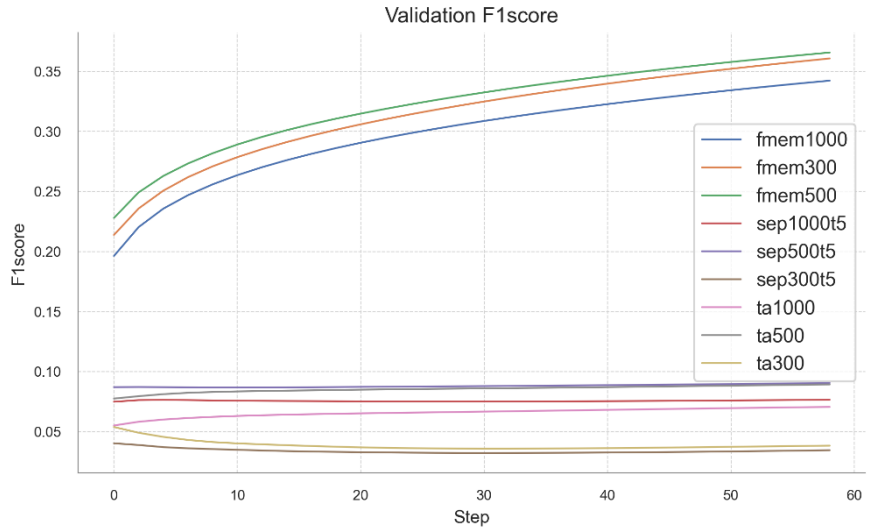
In this paper, the accuracy of the models was tested on three tasks for three models, and the final test set results were obtained as shown in Table 4, and the training and testing process is shown in Figure 3.

Table 4. Test results.

metric	task 300			task 500			task 1000		
	SEP	TM	FMEM	SEP	TM	FMEM	SEP	TM	FMEM
accuracy	0.883	0.856	**0.900**	0.912	0.884	**0.923**	0.948	0.918	**0.954**
f1	0.032	0.035	**0.328**	0.088	0.086	**0.335**	0.075	0.067	**0.311**



(a)



(b)

Figure 3. Training process evaluation results.

With Figures 3(a) and (b), it can be observed that the FMEM model has excellent evaluation performance in all three task dimensions, where the accuracy and F1score are optimal in the task dimensions of 300 and 500.. The F1score in the 300 task dimension far exceeds the performance of the single embedding model sep and the transformer model.

It can be seen that the performance of the transformer model decreases severely, but the gap becomes smaller as the number of training steps increases, and the final performance differs from that of the single embedding model by about 2%. Throughout the testing process, the performance of the multiple-embedding model outperforms that of the single-embedding model.

3.6. Predictive Analysis

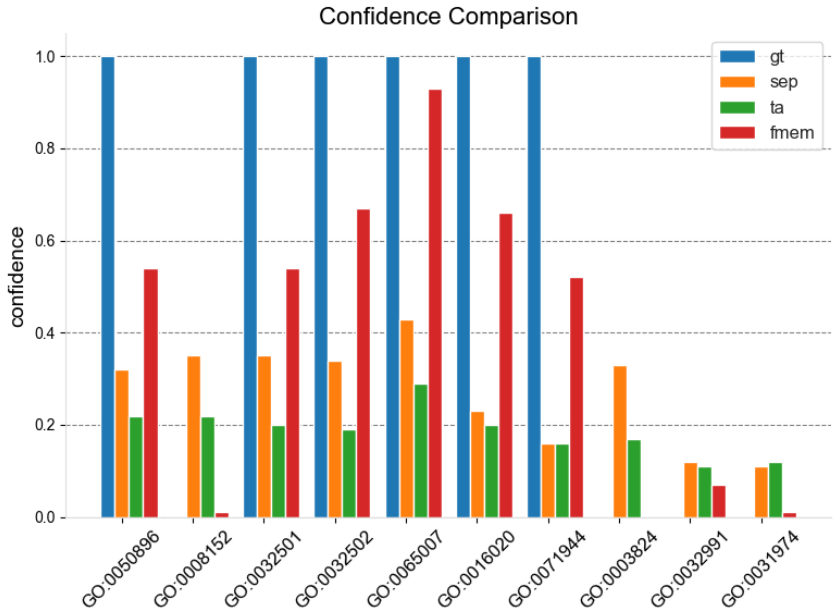


Figure 4. Comparison of Prediction Confidence.

Table 5. model predicted loss.

model	mse	total
SEP	0.3202	3.2018
TA	0.3858	3.8580
FMEM	0.0888	0.8881

To further compare the confidence prediction effect of the fused embedding model and the single embedding model on the samples, this paper compares the difference between the prediction confidence of the first 10 labels of multiple models and the true labels. With the confidence differences in Figure 4 and the average and total losses of the models in Table 5, it is easy to see that the confidence given by the fused embedding model is closer to the ground truths, with improved accuracy and stability.

3.7. Model Reasoning Evaluation

Compared to fine-tuning the language model and separating the inference training, using fusion embedding not only improves the accuracy of the model, but also is incomparable in terms of model training speed.

Table 6. Model Inference Time.

model	size	batch	cost time
LLM	<400M	4	2s
SEP	<1M	128	5ms
TA	<2M	128	13.24ms
FM	<1M	128	5.88ms

According to the training speed comparison between fine-tuning training and fusion-embedding training for the language model in Table 6, the speed improvement of fine-tuning training to separation training can reach more than 100 times, and the supported batch size is much larger than that of the fine-tuning model. Fusion embedding inference can train the model quickly and adjust the model parameters in time.

4. Conclusions

In this paper, a fusion embedding model is proposed, which aims to extend the applicability of downstream tasks by fusing the embedding of pre-trained models with downstream tasks, improve prediction performance and robustness, and provide better generalization results than single embedding models and fine-tuned models. Using the fused embedding approach can significantly save resources and increase computational speed.

In addition, this paper finds that although fusion embedding can improve prediction performance and robustness, there is an upper limit to its improvement. Therefore, in practical applications, this paper needs to select the most suitable models and methods according to different downstream tasks and data sets.

References

[1] Jumper, J. M. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021).

[2] Elnaggar, A. et al. ProtTrans: Toward Understanding the Language of Life through Self-Supervised Learning. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 7112–7127 (2022).

[3] Ashburner, M. et al. Gene Ontology: tool for the unification of biology. Nature Genetics 25, 25–29 (2000).

[4] Wang, S., Peng, J., Ma, J. & Xu, J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Scientific Reports 6, (2016).

[5] Vaswani, A. et al. Attention is all you need. arXiv (Cornell University) vol. 30 5998–6008 (Cornell University, 2017).

[6] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (Devlin et al., NAACL 2019).

[7] Raffel, C. et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21, 1–67 (2020).

[8] Gligorijević, V., Barot, M. & Bonneau, R. deepNF: deep network fusion for protein function prediction. Bioinformatics 34, 3873–3881 (2018).

[9] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N. & Linial, M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38, 2102–2110 (2022).

[10] Dallago, C., Schütze, K., Heinzinger, M., Olenyi, T., Littmann, M., Lu, A. X., Yang, K. K., Min, S., Yoon, S., Morton, J. T., & Rost, B. (2021). Learned Embeddings from Deep Learning to Visualize and Predict Protein Sets. Current protocols, 1(5), e113.

[11] Huang, S.-C., Pareek, A., Seyyedi, S., Rubin, D. L. & Lungren, M. P. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. Npj Digital Medicine 3, (2020).

[12] Hirasawa, T., Yamagishi, H., Matsumura, Y. & Komachi, M. Multimodal Machine Translation with Embedding Prediction. (2019).

[13] Lange, L., Adel, H., Strötgen, J. & Klakow, D. FAME: Feature-Based Adversarial Meta-Embeddings for Robust Input Representations. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (2021).