Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230610

Performance analysis of using multimodal embedding and word embedding transferred to sentiment classification

Zhuo Zou

Materials science and engineering, Hunan University of Technology, ZhuZhou, 412007, China

15020140132@xs.hnit.edu.cn

Abstract. Multimodal machine learning is one of artificial intelligence's most important research topics. Contrastive Language-Image Pretraining (CLIP) is one of the applications of multimodal machine Learning and is well applied to computer vision. However, there is a research gap in applying CLIP in natural language processing. Therefore, based on IMDB, this paper applies the multimodal features of CLIP and three other pre-trained word vectors, Glove, Word2vec, and BERT, to compare their effects on sentiment classification of natural language processing, to test the performance of CLIP multimodal feature tuning in natural language processing. The results show that the multimodal feature of CLIP does not produce a significant effect on sentiment classification, and other multimodal features gain better effects. The highest accuracy is produced by BERT, and the Word embedding of CLIP is the lowest of the four accuracies of word embedding. At the same time, glove and word2vec are relatively close. The reason may be that the pre-trained CLIP model learns SOTA image representations from pictures and their descriptions, which is unsuitable for sentiment classification tasks. The specific reason remains untested.

Keywords: CLIP, Multimodal Machine Learning, Sentiment Classification, BERT.

1. Introduction

The Pre-training word vectors include Word2vec (sparse word vectors based on window prediction), glove (Matrix Factorization and context for word embedding learning), Bert (BERT based on encoder of transformer and trained by two pre-training tasks including MLM (Masked language model) and Next Sentence Prediction Model (NSP) [1-3]. They were widely used in natural language processing tasks including sentiment analysis and Text similarity.

At the same time, Multimodal machine learning (Hereinafter referred to as the MML) is recently the hot research direction, and that study semantics, audio, image, video. MML can hand with and understand information of Multiple source mode about by the way of machine learning. Multimodal machine learning gains tremendous progress in computer vision, such as VirTex, ICMLM, ConVIRT, and CLIP [4]. More Specifically Speaking, Mahajan et al. and his team trained the models on large-scale data sets of images, while ConVIRT and other two model trained on small data sets of images [5]. Alec Radford and his team collected a dataset of 400 million (image, text) from the Internet and used it to train the model. When Model have trained, enabling transfer to the downstream tasks by the method that are Using visual concepts learning Natural Language feature and the zero-shot. Alec Radford and his team finshed different experiment exceeding 30 about computer vision and proved the

DOI: 10.54254/2755-2721/5/20230610

performance about the model [1]. However, it is blank to use multimodal features for the task of the Natural language process, So it may be consequential that Using multimodal features introducing computer vision features apply to the Natural language process [5]. Fine-tuning is the most common transfer learning method when deep learning models are used. It specifically refers to obtaining a pre-trained model on the originating tasks and then further training it on the target task, reducing the demand for target label data and improving the model's performance.

The word vector technology has demonstrated strong performance and transfer ability on many tasks. Mathematical embedding of each word in a space of one dimension into a vector space which is low dimensional. The approach which generate mapping relationships include the neural networks, dimensionality reduction about word co-occur matrices probabilistic models , interpretable knowledge base methods, and explicit representation of terms in the context of word occurrences .The Results of the task including sentiment analysis and syntactic analysis are increased by word embeddings or phrase embeddings, when embeddings is the underlying input of the tasks .For further comparing the performance about various word vectors including the multimodal feature of CLIP , this paper selects sentiment classification task as the research object, By studying the impact of different word vectors applied in sentiment analysis tasks on model accuracy. By studying the results of word embedding layer applied to sentiment analysis task between different word vectors, and then comparing the differences between word vectors. It mainly refers to the application of word vectors that have been trained by predecessors on large-scale pre-training corpus. The selection of a good contrast word vector is also a very important factor for the experiment, and the selection of an appropriate training model is also very important for the experiment.

This paper thinks less model parameter is better for the experiment to compare different word embedding, which is used as word vector. The paper selects K-Nearest Neighbors and a fully connected and single-layered network as classification algorithms because Their parameters can be trained less than other machine learning models. This paper based on IMDB used text features including pre-trained CLIP multimodal features, Word2vec, Glove, and BERT for word embedding and applied K-Nearest Neighbors and a fully connected and single-layered network to IMDB for sentiment classification.

At the same time, few-shot learning is used for FCC, and Zero-shot learning is used for KNN. This paper proposes the following hypotheses regarding the effect of different word embedding layers and multimodal features for sentiment analysis results. 1) CLIP pre-training possess multimodal features, and pre-training CLIP uses a Transformer as a text encoder, so the accuracy of CLIP text features should be better than that of Glove and Word2vec in the process of the experiment. 2) Given the strong performance of BERT based on the encoder of Transformer. The result of Bert and CLIP should be close and slightly worse than CLIP. 3) The result of Glove and Word2vec should be close but is not as good as BERT. 4) The fully connected and single-layer network effect is better than that of KNN. Finally, comparing different accuracy of the model is used to evaluate the effect of CLIP multimodal features in natural language processing.

The test results demonstrate BERT has the best performance for two control experiments, including the fully connected layer and KNN layer, which reach 87.64% and 67.05%, respectively. This is mainly due to the strong semantic capturing ability of BERT, although the training time of BERT is the longest, which takes about a few hours. The second best is Word2vec, which is close to Glove, but much better than BERT trained. The most unexpected, and the worst effect, is that the text feature of CLIP is only 67.70% and 47.54% respectively, which is the worst effect. The author guessed that this was mainly because of the poor effect caused by the corpus of CLIP model training.

2. Methods

2.1. Classification model

This paper studies the effect of multimodal vector transfer in NLP. The dependent variable of this experiment is the accuracy of the model, which is derived from two groups of data and is a continuous variable. To control the influence of parameters of the model on the results as much as possible, the K-

DOI: 10.54254/2755-2721/5/20230610

nearest neighbor algorithm and a fully connected and single-layered network with fewer model parameters are used in this study.

K-nearest neighbor (KNN) is a basic classification and regression algorithm, a supervised learning method. Suppose A task is given a training dataset in which the instance class is specified. Using the KNN, a new instance is predicted by majority voting according to its nearest neighbors' instance category [6,7].

Fully connected network. The matrix-vector product linear transformation (weighted and biased):$z = w^T x + b$.is a core operation about the fully connected network. The calculation is equal to linear regression calculation, assigning weights to each input vector x to calculate a result vector z. It is essentially that a space map to another about a linear convert . It plays the role of classifier in the network.

2.2. The word vector

Word2Vec. After Mikolov introduced the concept of Word vectors in his 2013 paper "Efficient Estimation of Word Representation in Vector Space," the field of NLP has suddenly entered the world of embeddings. Such as Sentence2Vec, Doc2Vec, and Everything2Vec. Word embedding is hypothesis which is "the sense of a word can be expressed by context of word "about language model. Word vector trained by Word2vec is ow-dimensional and intensive, when it compare with the traditional word vector called one-hot that is the high-dimensional and thin. Word2vec is the process of using a one-layer neural network (CBOW) to take a sparse one-hot word vector map and call it an n-dimensional (n is usually hundreds) dense vector [1].

CBOW. When model trained by using a text corpus and a determinate length window, the context is used to predict the target vocabulary. Although CBOW or Skip-gram are great, both methods are based on local the context window method. They do not availably use the global information of vocabulary is co-occurrence statistics. Jeffrey Pennington et al. created a new way called Glove in 2014,to solve the disadvantage that is global matrix factorization and local context window. This method learns word vectors based on the statistical information of all word co-occur. It combines the merits about statistical information, and the context window way improves the effect [2].

Glove combines global word co-occurrence of statistical information and the advantage of context of wors window way, can be said to be the two mainstream methods is a kind of comprehensive, but compared to the global matrix decomposition method, because the Glove does not need to compute the number of co-occurrence of 0 vocabularies, therefore, can significantly reduce the amount of calculation and data storage space.

BERT is the Encoder of a bidirectional Transformer. BERT proposes a new pre-training thought, called the masked language model (MLM), and overcomes the unidirectional limitation. The inspiration of MLM come from the Cloze task. MLM randomly generate some masked tokens that is the model input, and the goal is that predict the original lexical ID based just on the masked words [3]. Unlike the left-to-right language pre-training model such as LSTM and RNN, the MLM allows representations to mix together with the left and right sides of the context to pre-training Transformer. Except for a masking language model, the authors introduce a "Next sentence prediction task that works with MLMS to pretrain representations of text pairs. BERT is compared with the older sequence model including RNN, LSTM and GRU, BERT can be trained concurrently, extract the semantic features of words in all texts meanwhile, and extract the features at multiple levels to reflect the semantic of texts more comprehensively. It is that it can get the semantic of word by the context about sentence to avoid ambiguity to compare with Word2Vec [1,8-10].

CLIP experimented with large-scale data (four hundred million pairs data about image-text) and models. The core of the model approach is to train visual models through contrastive learning using multimodal features from supervised natural language signals. It also tests different Vision models (from ResNet to Vision Transformer) and Text models (from CBOW to Text). The authors have studied the successful application of cloze and autoregressive pre-training tasks in NLP in the field of CV. CLIP uses NLP features to train visual models and is used to transfer to multiple datasets with the support of large models and large-scale data by text zero-shot style. The performance of this method is comparable to that of supervised learning. CLIP is a specific model that can reach the effect of a

DOI: 10.54254/2755-2721/5/20230610

supervised task that requires labeled data without requiring large-scale task-specific annotations using text zero-shot. Therefore, it simplifies our data processing, model training, and downstream inference tasks and can be applied to many computer vision and multimodal fields.

3. Experimental results and analysis

3.1. Data description

The labelled dataset includes Fifty thousand IMDB movie reviews exclusive to sentiment analysis. The sentiment include positive and negative about model reviews, a sentiment score of 0 mean that an IMDB rating exceed five, while a sentiment score of 1 is a rating under seven . There are less than 30 reviews about the same movie. The twenty-five thousand movie reviews labelled training set and the twenty-five thousand movie review test set is not in the same movie as now.

3.2. Experiment design

The dataset is equally divided into two sets, including training and testing in the first set of experiments. The training set comprises the 25,000 labeled reviews the model generalizes on the word vector. The experiment evaluates the word vectors using the nearest neighbor classifier on the cross-validated training set and the fully connected network for text categorization tasks in all cases.

The dataset is divided into training, validation, and test sets in the second set of experiments. The training set comprises the 25,000 labeled reviews on which the model generalizes the word vector, the validation set comprises 17,500 data points, and the test set is 7500. The same fully connected network is used in all cases to get the model's performance and thus evaluate how good the vector is.

The explained variable in this study is the accuracy of sentiment analysis on IMDB. The study uses the result of sentiment analysis in IMDB. To keep the control variables, the parameters and methods remained unchanged in the same dataset of controlled trials except for different word embedding layers. The key independent variable in this study is the pre-training and initialized word vector of the word embedding layer. To compare the effect of word vectors of the multi-modal pre-training model called CLIP with multiple sets of pre-trained word vectors, three pre-trained text features of Word2vec, Golve and BERT are selected. Word2vec vectors are 300-dim pretrained on Google News. Glove vectors are 300-dim pre-trained on Wikipedia by Stanford. BERT is the base revision, its word vector is a 768-dimensional word vector, and CLIP is Vit-B /32. It is a model with a 512-dimensional word vector [1-4].

3.3. Results and analysis

Table 1. Word2vec, Glove, Bert, CLIP applied in KNN (Zero Shot learning).

Embedding	Glove	Word2vec	BERT	CLIP
Accuracy	51.99%	62.06%	68.05%	47.54%

Table 1 illustrates the experimental result of the word vector of multimodal features and other three kinds applied to the sentiment classification tasks by KNN in IMDB. Among group 1, the powerful BERT consistently outperformed Wor2vec and Glove, and Word2vec performed better than Glove in this experiment (see Table 1). Hypothesis 1 (Accuracy of CLIP text features probably exceed Glove and word2vec) was not confirmed.

Table 2. Application of word2vec, Glove, Bert, CLIP in FCC (Few Shot learning).

Embedding	Glove	Word2vec	BERT	CLIP
Accuracy	86.73%	87.18%	87.64%	63.70%

Table 2 shows the experimental result of the word vector of multimodal features and the other three kinds applied to the sentiment classification tasks by FCC in IMDB. In group 2, with the fully

DOI: 10.54254/2755-2721/5/20230610

connected network, BERT is the best in this experiment, and the research results again confirmed the power of BERT. The effect of the word embedding layer with CLIP multimodal feature words is lower than that of other pre-training word vectors (see Table 2). Again hypothesis 2 is not confirmed. Besides, Hypothesis 3 and Hypothesis 3 were confirmed.

Different word embedding layers lead to the results, which may be due to the following reasons. Firstly, BERT is based on an encoder of Transformer with attention mechanism of powerful feature capture ability and individual training tasks including MLM (Masked language model) and Next Sentence Prediction Model (NSP), which captures the semantics of the text well and shows a good result. Secondly, Word2vec and glove also show the ability of traditional static word vectors. However, the experiment does not transfer CLIP ability in the CV to that is in the tasks about natural language processing, which may be related to the training data of CLIP and the structure of the model itself. The pre-training CLIP uses many image-text pairs, which are static descriptions of static images instead of images with expressions and emotional text pairs. Secondly, the CLIP model structure does not capture the semantics of the text well but just matches the picture and the text.

It should be illustrated that there are limitations in this paper because the data in this study are all from the IMDB dataset. Firstly, the corpus of the embedded pre-trained model is not uniform. Secondly, the hardware Settings, the time of model training, and the learning efficiency are insufficient. Future Work

Subsequent consideration will be given to applying multimodal features to more natural language processing tasks. CLIP is a picture-text matching task that may achieve better results in named entity recognition.

4. Conclusion

In this work, the research has shown that the multimodal features of CLIP do not achieve a pleasing effect, so much so that its accuracy is lower than other pre-training word vectors on application to the embedding layer. The findings were unexpected, probably because of the following. On the one hand, the pre-training model's CLIP data describes images rather than emotional text. On the other hand, most images of the data are static instead of emotional vision. Thus, models do not acquire skills in capturing the emotion of the text. Finally, the study does not consider data for pre-trained multimodal machine learning models. However, it is excited about the future of multimodal features. It plans to apply them to other tasks about Natural language processing. As illustrated in this article, the CLIP model needs to be trained on images with expressions, descriptions of expressions, and text with emotions for the IMDB. The above-related factors can be further refined to facilitate the in-depth study of this issue. Pre-trained word embeddings become a necessary component of neural network architectures for Natural Language Processing tasks. In the next work, this paper plan to use multimodal features for more tasks for NLP, such as named entity recognition.

References

[1] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[4] Desai, K. and Johnson, J. Virtex: Learning visual rep-resentations from textual annotations. arXiv preprint arXiv:2006.06666, 2020.

[5] Alec Radford, Jong Wook Kim et al. Learning Transferable Visual Models From Natural Language Supervision arXiv:2103.00020 ,2021.Association for Computational Linguistics.

[6] Tenindra Abeywickrama, Muhammad Aamir Cheema, David Taniar: k-Nearest Neighbors on Road Networks: A Journey in Experimentation and In-Memory Implementation. CoRR abs/1601.01549 ,2016

DOI: 10.54254/2755-2721/5/20230610

[7] Vaswani, Ashish, et al. Attention is all you need. Advances in neural information processing systems 30 (2017).
[8] Elman, J. L. . Finding Structure in Time.Cognitive Science 14.2(1990):179-211.
[9] Hochreiter, S. , and J. Schmidhuber . Long Short-Term Memory. Neural Computation 9.8(1997):1735-1780.
[10] Matthew E. Peters, Waleed Ammar, Chandra Bhagavatula, Russell Power.Semi-supervised sequence tagging with bidirectional language models. arXiv:1705.00108,2017