

Multimodal Data Mining Based on Self Attention Feature Alignment

Binxi Gao*

Department of Computer, Beijing Institute of Technology, Beijing, China

* Corresponding Author Email: 1120212160@bit.edu.cn

Abstract. This study explores the application of multimodal data mining in text and image vector processing, aiming to improve the depth and breadth of data analysis by integrating information from different data types. We use the FairFace dataset combined with the CLIP model encoding layer to obtain text and image vectors, and use the K-Means clustering algorithm to achieve vector dimensionality reduction. Subsequently, we introduced the bipartite graph matching algorithm to achieve maximum matching between text vectors and image vectors, and calculated the contrastive learning loss and similarity loss. The entire process covers steps such as data preparation, feature extraction, vector dimensionality reduction, matching algorithms, and loss assessment, constructing a complete text and image matching task process. Our research contributions include using K-Means clustering algorithm to achieve vector dimensionality reduction, as well as introducing bipartite graph matching algorithm and calculating two types of losses in text and image vector matching, further improving matching quality.

Keywords: Multimodal, Data Mining, Feature Alignment, Attention.

1. Introduction

1.1. Research significance of the topic

Multimodal data refers to data composed of two or more modalities, which can be text, images, audio, video, etc. Multimodal data mining refers to the process of extracting knowledge and information from various types of data. The significance of multimodal data mining lies in integrating different types of data (such as text, images, audio, etc.) to achieve more in-depth and comprehensive data analysis. This comprehensive analytical ability can not only reveal the complex patterns and correlations behind the data, but also improve the accuracy and wide applicability of data understanding. For example, combining image and text data for sentiment analysis can provide a more accurate understanding of the user's emotional state, thereby improving user experience design. In addition, multimodal data mining has expanded the application fields of data analysis, such as medical diagnosis, intelligent transportation systems, etc., providing more accurate and innovative solutions for various industries. Most importantly, this approach promotes interdisciplinary cooperation and development, advances cutting-edge research in data science and engineering technology, and contributes to social progress.

1.2. Previous methods and existing defects

Compared to multimodal data mining, unimodal data mining has some obvious shortcomings. Firstly, unimodal data mining can only extract information from a single type of data and cannot fully utilize the complementary and interactive effects between different data types. This leads to limitations in the analysis results and incomplete information, which may overlook important patterns or correlations in the data.

Secondly, unimodal data mining often struggles to provide in-depth and comprehensive analysis when facing complex and diverse real-world problems. For example, relying solely on text data for sentiment analysis may not capture the emotional content contained in image or audio data, limiting a comprehensive understanding of the user's true feelings.

In addition, unimodal data mining often fails to provide effective solutions when dealing with interdisciplinary problems or tasks that require the integration of multiple data types. The advantage

of multimodal data mining lies in its ability to integrate different data types, provide more accurate and comprehensive perspectives through comprehensive analysis, and promote progress and innovation in data science and related fields.

1.3. Our method and improvements achieved

We use the FairFace dataset and combine it with the CLIP[1] model to obtain the vectors of text and image. The usual approach is to first use the CLIP model to encode the text and image separately, and obtain their feature vectors. Next, we use the K-Means clustering algorithm to achieve vector dimensionality reduction. Next, we use the bipartite graph matching algorithm to obtain the maximum match between the text vector and the image vector, which is used to find the best matching between the text vector and the image vector. Finally, to evaluate the quality and performance of the matching, we calculate the loss function - contrastive learning loss and similarity loss (MSE).

The entire process covers data preparation (using the FairFace dataset), feature extraction (CLIP model), vector dimensionality reduction (K-Means clustering), matching algorithm (bipartite graph matching), and final evaluation (loss calculation). These steps together constitute a complete process for text and image matching tasks.

2. Related works

2.1. Multimodal data mining

Multimodal learning is an important branch of machine learning. Currently, mainstream multimodal large models are mostly based on Transformers. Transformer is mainly used for natural language processing (NLP) tasks, especially for sequence-to-sequence problems such as text generation, completely changing the traditional RNN and LSTM sequence modeling paradigms. In the early days, [5] was the earliest literature in the field of multimodal emotion recognition to adopt cross modal fusion technology. It proposed a MuIT model based on the Transformer framework for cross modal interaction, designed Transformer oriented paired cross modal attention, and laid the foundation for the modal fusion part of subsequent articles; Another article, [6] delves into the phenomenon of multimodal large models experiencing content that does not match the image information in their responses (hereinafter referred to as "hallucinations"), and proposes an alignment framework called RLHF-V to reduce "hallucinations" at the data and algorithm levels; Li et al. proposed a new visual language framework, BLIP [7] which flexibly shifts towards visual language understanding and generation tasks. By tracking the title and effectively utilizing noisy network data, the title can generate composite subtitles and filter out noise. And has achieved state-of-the-art results in a wide range of visual language tasks.

2.2. Multimodal data feature extraction

Extracting a good feature representation from raw data has always been an important issue of concern in machine learning. [8] uses a Transformer network with feature reconstruction to fill in missing data; Another article, [9] proposes the MISA framework, aiming to use domain adaptive methods to construct schemes suitable for multimodal feature space interaction, learn the commonalities between different modalities and narrow modal gaps, effectively enhancing the effectiveness of modal fusion; [10] focuses on using contrastive learning to perform multimodal entity alignment tasks. There are a total of five modalities, namely graph structure modality, relationship modality, attribute modality, text modality, and image modality. Adaptive multi task learning weights are used to obtain joint representation for the five modalities, and then loss is calculated. [11] achieves deep modal feature extraction (meta modal mixing) by learning dynamic cross modal weights.

3. Method

3.1. Overview

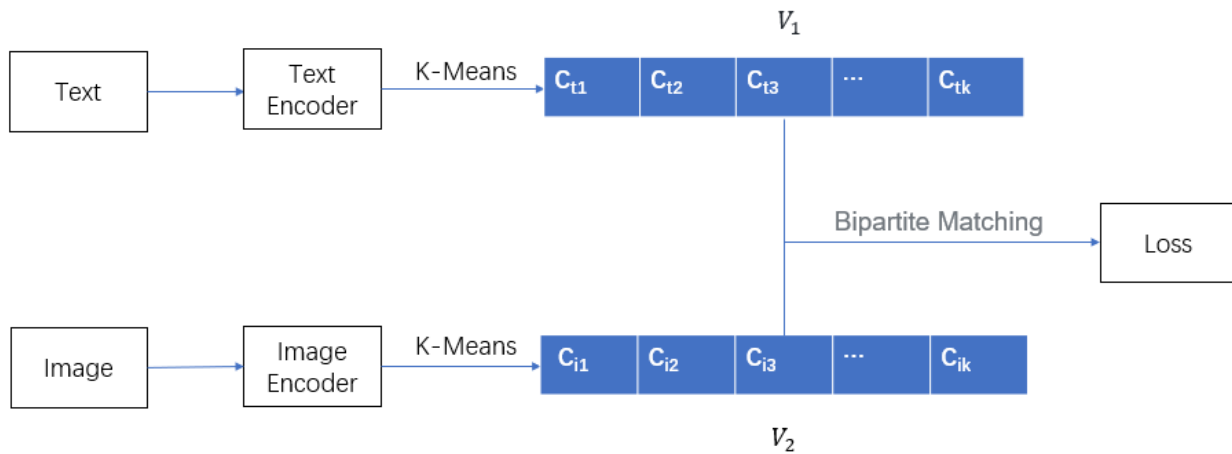


Fig. 1 Overview of the proposed method. Using the CLIP model's text encoder and image encoder to obtain high-dimensional vectors, then using K-means clustering to obtain the text vector V and image vector V' , using bipartite graph matching algorithm to obtain the overall optimal matching relationship between them and calculating the contrastive learning loss and similarity loss.

The core idea of using the CLIP model is to use a large amount of paired data between images and text for pre training, in order to learn the alignment relationship between images and text. The CLIP model has two modalities, one is the text modality and the other is the visual modality, consisting of two main parts:

- (1) Text Encoder: Used to convert text into a low dimensional vector representation - Embedding.
- (2) Image Encoder: Used to convert images into similar vector representations - Embedding.

The multi-dimensional vectors of Text Encoder and Image Encoder are clustered using K-Means to obtain the text vector V and image vector V' as shown in Figure 1. Then, the bipartite graph matching algorithm is used to obtain the overall optimal matching relationship between them, and the contrastive learning loss and similarity loss are calculated. The overall flowchart is shown in Figure 1.

3.2. Self attention feature alignment

Self-Attention, also known as internal attention mechanism, is an attention mechanism that associates different positions of a single sequence to calculate the representation of the same sequence. The basic formula is as follows:

$$Attention(Q, K, V) = softmax \left(\frac{QK^T}{\sqrt{d}} \right) V \quad (1)$$

It allows models to dynamically focus on input information at different positions when processing sequence data, rather than fixed operations on the entire sequence like traditional recurrent neural networks or convolutional neural networks. Through self attention mechanism, the model can calculate the importance weight of each input position for all other positions, and then apply these weights to the input representation, effectively capturing long-distance dependencies.

The reason why Self Attention can perform feature alignment is due to its unique calculation method and mechanism. In Self Attention, each input position updates its own representation by calculating its weight with all other positions. This weight is calculated through linear transformations of queries, keys, and values, and is normalized by Soft-Max to ensure that the sum of all weights is 1. By dynamically assigning weights through association based calculations, weight allocation, and feature integration, the model effectively integrates and utilizes feature information from different locations, enabling it to better capture long-range dependencies and global structures in input data.

3.3. Bipartite graph matching and loss calculation

A bipartite graph refers to a graph $G = (V, E)$, where the vertex set V can be divided into two disjoint subsets $V = V_1 \cup V_2$, and each edge in the edge set E is connected to a vertex belonging to V_1 and a vertex belonging to V_2 . Our goal is to find a maximum match in this bipartite graph, which is a set of edges such that each edge connects a vertex belonging to V_1 and a vertex belonging to V_2 , and any two edges do not share the same vertex.

To achieve this goal, we consider the following specific steps and methods. Firstly, use the feature extraction method of the CLIP model to convert text data into text vector V and image data into image vector V' . These vector representations are typically high-dimensional, with each dimension corresponding to an aspect or attribute in the feature space. Secondly, input the text vector V and image vector V' into the K-Means clustering algorithm respectively, and divide them into K categories (or clusters). Each category represents a set of similar vectors, which can be seen as a set of vertices. Next, consider the clusters $C_{t1}, C_{t2}, \dots, C_{tk}$ of the text vector obtained by K-Means clustering as one vertex set V_1 of the bipartite graph, and the clusters $C_{i1}, C_{i2}, \dots, C_{ik}$ of the image vector as another vertex set V_2 , where k is the number of clusters. Then, in the established bipartite graph, find a set of edges E such that each edge connects a vertex from V_1 and a vertex from V_2 . The goal of this matching is to select as many edges as possible, and any two edges do not share the same vertex. Finally, use the appropriate graph theory algorithm - the Hungarian algorithm - to solve the bipartite graph maximum matching problem. This algorithm can effectively find the maximum match and ensure that each vertex is connected to at most one matching edge.

Above, we can maximize the dimensionality matching between text vectors and image vectors, effectively exploring and utilizing their similarity and complementarity in the feature space. This method can be applied not only to text and images, but also to other types of data vector matching problems, such as feature matching between audio and video.

4. Experiments

4.1. Dataset Introduction

The FairFace dataset is a dataset used for facial recognition and analysis, which aims to cover facial images of various races, genders, and ages to support the development of fair and accurate facial recognition technologies, solve the problem of imbalanced factors such as race and gender in traditional datasets, and improve the universality and fairness of existing technologies among different groups.

4.2. Implementation details

We conducted a series of model training, including 5 ResNets and 3 Vision Transformers. ResNets includes the classic ResNet-50 and ResNet-101, as well as three models extended in EfficientNet style (RN50x4, RN50x16, and RN50x64), which have computing resources 4 times, 16 times, and 64 times that of ResNet-50, respectively. Vision Transformers include ViT-B/32, ViT-B/16, and larger ViT-L/14. All models were trained on 32 epochs using Adam optimizer and decoupled weight decay regularization, with learning rates scheduled through cosine decay. The initial hyperparameters are set through grid search, random search, and manual adjustment based on the ResNet-50 model. We use a large mini batch (32768) and mixed precision training to accelerate training and save memory, while utilizing gradient checkpoints, semi precision Adam statistics, and semi precision random rounding of text encoder weights to further optimize memory usage. The largest model, such as RN50x64, was trained on 592 V100 GPUs for 18 days, while the largest Vision Transformer was trained on 256 V100 GPUs for 12 days. The ViT-L/14 model also pre trained an epoch with a higher 336-pixel resolution to improve performance, similar to FixRes (Touvron et al., 2019).

4.3. Evaluating Indicator

Accuracy is one of the common indicators used to evaluate the performance of classification models, which represents the proportion of correctly predicted samples to the total number of samples. In multi classification problems, the calculation of accuracy is to consider a certain category, while all other categories are considered negative. In this way, in multi classification scenarios, the accuracy value of each category can be obtained. A high accuracy indicates that the model has a strong ability to predict correctly overall, but in some cases, accuracy is not the only criterion for evaluating model performance, especially when the number of samples in different categories varies greatly, there may be imbalanced dataset problems.

4.4. Experimental results and analysis

Table 1. Percent Accuracy on gender classification of images by FairFace race category

Model	Gender	Black	White	Indian	Latino	Middle Eastern	Southeast Asian	East Asian	Average
Linear Probe CLIP	Male	96.9	96.4	98.7	96.5	98.9	96.2	96.9	97.2
Zero-Shot CLIP	Male	96.3	96.4	97.7	97.2	98.3	95.5	96.8	96.9
Linear Probe Instagram	Male	92.5	94.8	96.2	93.1	96.0	92.7	93.4	94.1
Ours	Male	92.3	95.9	97.2	96.6	96.0	95.3	95.5	94.2

Racial comparison: In most models, Indian and Middle Eastern populations have higher accuracy, while East Asian and Southeast Asian populations have lower accuracy. This indicates that the model may have certain advantages in identifying image features of Indian and Middle Eastern ancestry, while relatively poor in identifying East Asian and Southeast Asian ancestry.

Comparison between models: The Linear Probe CLIP model performs consistently and with high accuracy across most races and genders. The Zero Shot CLIP model performs better in Latino and East Asian populations, but slightly decreases in other races. The Linear Probe Instagram model has relatively low accuracy in various races and genders, and overall performance is not as good as other models. The Ours model performs well among Indian, Latinx, and East Asian populations, but performs moderately among other races. The overall classification situation is shown in Table 1.

Overall, different models have varying levels of accuracy for different races and genders, and currently no model performs best across all races and genders. This indicates that the performance of the model may be influenced by the sample distribution in the dataset, as well as the characteristics and training methods of the model itself.

When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

5. Conclusion

We employ the CLIP model to extract semantically meaningful features from textual descriptions and corresponding images within the FairFace dataset, converting them into high-dimensional vector representations. Subsequently, to enhance computational efficiency and reduce the dimensionality of these feature vectors, we utilize K-Means clustering. This clustering method partitions the original high-dimensional vectors into clusters represented by lower-dimensional centroids, thereby facilitating more efficient subsequent operations. Following by dimensionality reduction via K-

Means, we proceed with bipartite matching between the text and image vectors to achieve optimal alignment. The bipartite matching process aims to maximize the correspondence between textual descriptions and images by leveraging similarity metrics computed from the vector representations. This optimization is crucial for ensuring that each textual description is paired with the most relevant image counterpart.

Moreover, beyond K-Means, we explore alternative clustering algorithms such as DBSCAN and Gaussian mixture models on the initial CLIP-derived vectors. This exploration allows us to evaluate the efficacy of different clustering strategies based on the specific characteristics and distribution of the dataset, aiming to refine the representation of feature vectors for improved downstream tasks. In optimizing the bipartite matching algorithm, particularly enhancing upon the Hungarian algorithm, we focus on improving computational efficiency and accuracy. The Hungarian algorithm is well-suited for maximizing bipartite graph matchings, and enhancements in its implementation can significantly elevate the performance and scalability of our matching process.

Overall, this methodology integrates advanced techniques from machine learning and optimization, tailored to effectively handle the complexities and scale of modern datasets like FairFace, ultimately aiming to achieve robust and accurate associations between textual and visual data elements.

References

- [1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International conference on machine learning. PMLR, 2021: 8748-8763.
- [2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
- [3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
- [4] Kärkkäinen K, Joo J. Fairface: Face attribute dataset for balanced race, gender, and age[J]. arxiv preprint arxiv:1908.04913, 2019.
- [5] Tsai Y H H, Bai S, Liang P P, et al. Multimodal transformer for unaligned multimodal language sequences[C]//Proceedings of the conference. Association for computational linguistics. Meeting. NIH Public Access, 2019, 2019: 6558.
- [6] Yu T, Yao Y, Zhang H, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 13807-13816.
- [7] Li J, Li D, Xiong C, et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation [J]. 2022.DOI:10.48550/arXiv.2201.12086.
- [8] Sun L, Lian Z, Liu B, et al. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis[J]. IEEE Transactions on Affective Computing, 2023, 15(1): 309-325.
- [9] Hazarika D, Zimmermann R, Poria S. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis[C]//Proceedings of the 28th ACM international conference on multimedia. 2020: 1122-1131.
- [10] Lin Z, Zhang Z, Wang M, et al. Multi-modal contrastive representation learning for entity alignment[J]. arxiv preprint arxiv:2209.00891, 2022.
- [11] Chen Z, Chen J, Zhang W, et al. Meaformer: Multi-modal entity alignment transformer for meta modality hybrid[C]//Proceedings of the 31st ACM International Conference on Multimedia. 2023: 3317-3327.