Cross attention for Text and Image Multimodal data fusion

Stanford CS224N Custom Project

Dongyeong Kim
Department of Computer Science
Stanford University
dkim@stanford.edu

Abstract

The distinction between human cognition and single-model processing lies in humans utilizing diverse data from the external world such as smell, taste, sight, touch, and hearing, whereas models typically rely on a single type of data. To address this limitation, developments have been made to leverage multiple data types or architectures capable of receiving and analyzing various data forms, known as multimodal approaches. Multimodal systems have the potential to enable more human-like analysis, offering flexibility and fluency with diverse data; however, challenges remain in effectively combining different data types. For instance, sight (image) and text data are fundamentally different, with distinct embeddings, dimensions, and data ranges. A common methodology involves encoding these disparate data types into a single dimension and fusing them through concatenation. However, this approach faces the inherent limitation that images contain both positional and semantic information, whereas text primarily contains semantic information. Consequently, encoding image data into a single dimension aligned with text data results in the loss of positional information.

If during the data fusion process, semantic and position-related information is preserved, our multimodal approach can achieve more fluent and flexible outputs with the integration of two different data types. Therefore, this paper explores methods to enhance multimodal image and text data fusion, aiming to evaluate the results through generated images and performance scores.

In my research, three primary methods for multimodal data fusion have been identified: concatenation, mapping (CLIP method), and cross-attention mechanisms. My experiments indicate that while concatenation and mapping are effective in combining the two different data types, they often result in the loss of positional information inherent in image data. However, the cross-attention mechanism, which applies attention to both text and image features, allows for the combination of data by analyzing text data alongside each pixel of the image data.

1 Introduction

The main approaches for data fusion in multimodal systems are primarily divided into two categories: combining and feature-to-feature analysis. Both methods elucidate how features relate to each other, enabling models to effectively handle image-to-text or text-to-image transformations. However, for practical multimodal applications that interact with both image and text commands, it is crucial that the image and text commands correspond accurately, especially in relation to image data, since AI interacts with the world through images.

The CLIP methods [1], employed by OpenAI, and the concatenation data combining approach are widely used for image and text analysis. These methodologies have demonstrated their ability to explain the interconnections between images and text. Nevertheless, a significant issue arises: while

these methods perform well in image-to-text or text-to-image conversion, they struggle with text alignment for each image position. When using concatenation, the simple combination of data can learn overall data correlations during the decoding layers. The CLIP method, which utilizes the outer product, learns the relationship between features of image and text data. However, the limitation lies in the fact that the outer product captures only feature data from images, not positional data. Consequently, while image encoders can capture various semantic features through Convolutional Neural Networks (CNN), they lack positional information. To address this, I propose adding patches and positional embeddings to images before applying the outer product with image and text data.

Another method to align without positional embedding and patching is the cross-attention mechanism. Unlike the outer product approach, which performs feature-to-feature product between text and image without modifying the text encoder's data and reducing image data via average pooling (thus losing positional information), the cross-attention mechanism uses attention to directly learn the relationships between text features and image pixels. By employing attention mechanisms on image and text data, models can learn the relationships between text features and each pixel of the image encoder, thus preserving positional information [2].

2 Related work

2.1 Multimodal Methodology

In the paper "Foundations and Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions" [3], the authors provide a comprehensive overview of multimodal machine learning. The primary principles of multimodal learning are Representation, Alignment, Fusion, Translation, and Co-learning. The challenges highlighted in this paper include data inequalities, modality interaction, and scaling. These challenges arise due to differences in data composition and analysis methods across various data types, such as images and text. Image data encompasses spatial information, colors, positions, and semantics, while text data includes semantic, positional, and grammatical information. The encoding processes also differ; text is typically analyzed based on word relationships, whereas image analysis involves feature extraction. Consequently, when combining these distinct data types, issues of data inequality and modality interaction emerge due to the differing interpretation and encoding methodologies.

2.2 CLIP (Contrastive Language-Image Pretraining)

Several notable studies have addressed the fusion of different data types. For example, the CLIP (Contrastive Language-Image Pretraining) method [4] maps text and image features into a shared embedded space, enabling models to analyze feature-by-feature relationships between text and images. This approach facilitates image-to-text and text-to-image interactions by learning the interconnections between images and corresponding text. However, CLIP employs pooled features of images derived from Convolutional Neural Networks (CNN), which retain semantic information but lack positional data corresponding to each semantic feature.

2.3 Cross Attention Mechanism

One potential enhancement to the CLIP methodology is the addition of patch embeddings and positional embeddings to each patch. A more effective improvement involves applying a cross-attention mechanism between image and text. The basic concept of cross-attention is derived from the self-attention mechanism used in transformers, which utilizes keys, queries, and values. In self-attention, all three components come from the same data. In cross-attention, however, keys and values are derived from the image data, while queries are sourced from the text. This mechanism enables a one-to-one comparison between text and image features, establishing how each text feature relates to image features at specific positions. This approach has been explored in [5] for text and image interconnection similar to CLIP. In this paper, this approach will be considered for data fusion.

This approach is unique in my view, as traditional multimodal learning typically involves interchangeable analysis, such as text-to-image or image-to-text. However, this method focuses on utilizing both data types simultaneously during the learning process. Other approaches, like ImageBERT [6], involve several preprocessing steps, including semantic image patching and position embedding. By using whole image and text data without preprocessing, it is possible to demonstrate which

methodology excels in complexity analysis and relationship understanding. Furthermore, I believe addressing data fusion is crucial because it enables models to handle more complex data and enhances their ability to analyze the world fluently. While text or image data alone can provide valuable information, combining text and image data offers better generalization for world models. In real life, humans rely on multiple senses, such as voice, sight, sounds, and smells. Learning from combined data is essential for AI models to develop better world models. As large language models (LLMs) and large vision models (LVMs) already assist people in many areas, improving multimodal data fusion can enhance the integration of these diverse models, fostering greater flexibility and innovation in AI applications.

3 Approach

3.1 Baseline

The baseline will be CLIP methodology in data fusion. The CLIP invented from OpenAI is the most famous in image to text interaction and widely used for multimodal for image and text. The baseline will be compared with cross-attention data fusion, and concatenation data fusion. Here is the basic equations for image encoder and text encoder which is shared with other methodologies.

First, the image encoder is based on retrained resnet50 [7] which is composed with redidual CNN layers and normalization. **Image Encoder:** ResNet50, **Text Encoder:** BERT

Let $I \in \mathbb{R}^{H \times W \times C}$ be an image, and $T \in \mathbb{R}^{L}$ be a text sequence.

$$E_I = ResNet50(I) \in R^d \tag{1}$$

$$E_T = BERT(T) \in R^d \tag{2}$$

After encoding them in same dimension and make a dot product.

$$M = E_I \otimes E_T \in R^{d \times d} \tag{3}$$

Where $\otimes$ denotes the outer product. And then apply the 8 layered CNN decoder to match target image and see that the text and image is aligned with containing positional and semantic data.

$$O = CNNDecoder(M) \tag{4}$$

Where O is the output of the CNN decoder applied to the outer product M

3.2 Proposed method

To compare with baseline, the methodology will be used in this paper is concatenation and cross attention mechanism. The other parts for image encoder, text encoder, and CNN Decoder are same but inside of architecture, the only data fusion parts will be changed. For the concatenation, the E_I and E_T is connected and make d+d dimension as a result.

$$O_{cat} = concat(E_I, E_T) \in R^{d+d} \tag{5}$$

The cross-attention will be applied in same equation of self-attention, but different value for each key, value, and query - $\mathbf{K_{image}}$, $\mathbf{V_{imag}}$, $\mathbf{Q_{text}}$ The attention score A is calculated:

$$\mathbf{A} = softmax\left(\frac{\mathbf{Q_{text}}\mathbf{K_{image}}^T}{\sqrt{d_k}}\right) \in \mathbb{R}^{n \times n} \tag{6}$$

The d_k is the dimension of Key. The combined vector is calculated with Attention score and Value as:

$$\mathbf{O_{C-Attention}} = \mathbf{A}\mathbf{V_{image}} \in \mathbb{R}^{n \times d_V} \tag{7}$$

The figure1 shows architecture of cross attention and Value and Key are related with image, and Query is related to text, so the attention score can calculate the image and text relationship, and apply it into image data.

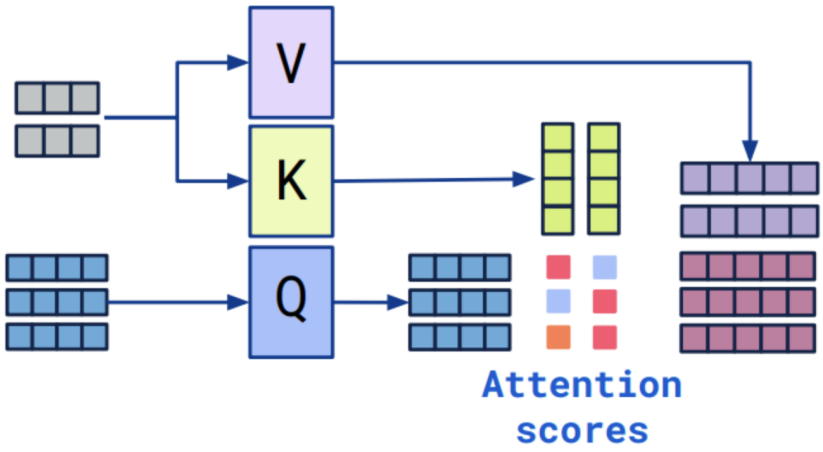

Figure 1: Cross Attention Architecture

3.3 Experiments

3.3.1 dataset

The dataset is used in MagicBrush provided by HuggingFace API. The data set contain source image, instruction text, and target image. The target image is modified source image based on the instruction text. I think this dataset is adequate for text and image combined data and how that data can effect on image modification. ML loop is transformed with normalized 512x512 image size. The dataset is mainly composed with 3 parts, source image, mask image, and target image. The mask image is where need to be modified with text instruction, and target image is where the modified image is added in the mask area. The figure2 shows one example of MagicBrush Dataset with instruction "leave only the scissor on the cup". The mask image in the example will be used as white(True)

Figure 2: Example of MagicBrush Dataset

in masking area, and all other area will be colored in black(False). This dataset is not originally focusing for catching multimodal can catch the positional and semantic information for different dataset, because with masking, the positional data is already targeted manually, but instead of using the target image as labeling data, this paper will use mask image as labeling data that the model can make the mask with source image and text instruction. Therefore, the similarity of mask data can prove that whether the model can classify which points of image is semantically, and positionally related to text data.

3.3.2 Evaluation Method and experiment details

The evaluation will be done in evaluation dataset which is 20% of proportional to train data, that the train data is 20% of whole data and evaluation data will be 4% of whole data outside of train data. The loss function is BCELoss(Binary Cross Entropy Loss) which is widely used for binary

classification loss function.

$$BCELoss(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N}(y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)) \tag{8}$$

The model's output is between [0,1] after sigmoid function, and masking image is also divided into two value 1(True) for white, and 0(False) for black. Learning rate is fixed with 0.002 without learning rate scheduler to see model's learning speed. The reason for using just portion of whole dataset, and only using mask image is because of memory and time. The target image is based on image generation. The original experiment set was decoder as ESRGAN which is pretrained GAN(generative adverisal neural network) for image generation, but there was memory problem in my experiment environment. In addition, even the model architecture with masking image, using whole dataset need a day to see only 10 epoches.

The model architecture is as I mentioned in Approach section, the ResNet50 and Bert model are used in all cases(cat, CLIP, and cross-attention), and 8 layers of CNN decoder is used to compare different data fusion methodology. The 8 layer of CNN decoder is not exactly same, because based on the data fusion methodology the data dimension is changed, but data dimension is modified with 1 or 2 small layers as I can.

Model training loop is composed with 50 epoches with 1000 source image and masking image pairs, and evaluation score is tested with 200 source image and masking image pairs. The training time took about 5 hours for each models with GPU parallel computing in cloud computing environment.

3.3.3 Results

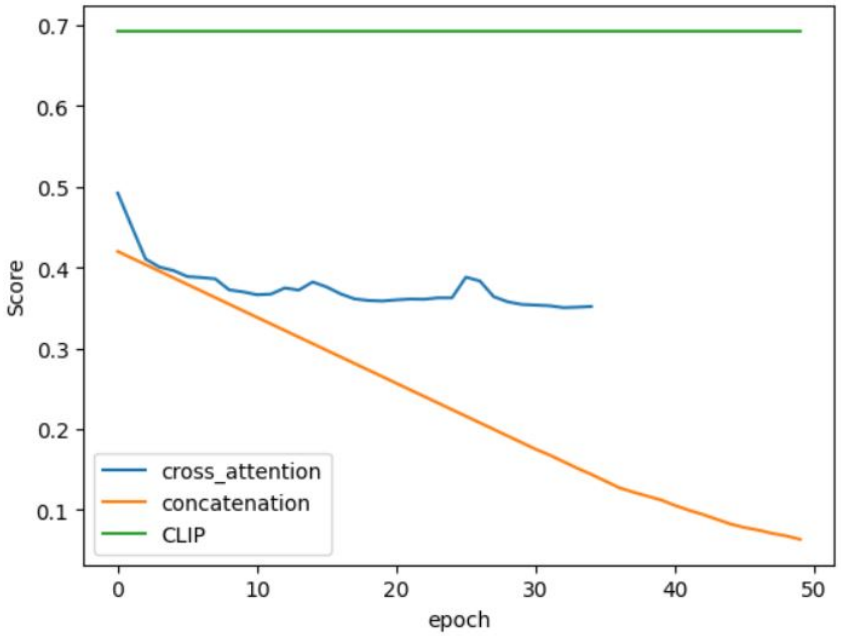

Figure 3: Loss score change with 50 epochs

Figure 3 illustrates the change in loss scores. The three cases were trained for 50 epochs, and both the CLIP and cross-attention mechanisms experienced early stopping at about 40 epochs due to no improvement on the training set. This result was unexpected; compared to concatenation, the cross-attention mechanism, which compares each text feature with image pixels, was expected to handle more complex data. The training loop showed that the cross-attention score plateaued at 0.36. This could be due to various reasons, including memory and time limitations, as I scaled down the image data and reduced the training dataset size. Interestingly, the concatenation methodology performed better than anticipated. The concatenation method directly connects text and image features without any intermediate fusion process, yet the model steadily learned to align with the training data. The

primary challenge was the dataset size and image scaling due to computational constraints. I started with the smallest model size and gradually increased it, but the running time remained long (about 5 to 6 hours per model).

Table 1: Evaluation Scores for Different Methods

Data Fusion Method	Evaluation Score
Concatenation	0.7093
CLIP	0.6929
Cross_attention	0.4459

The evaluation scores in Table 1 align with my expectations that the cross-attention method would match its training performance, but the concatenation method's performance did not align. In terms of data fusion, cross-attention demonstrated better regularization, whereas the concatenation method exhibited about 35 times higher loss.

Figure 4 presents input examples and the corresponding model predictions for the three cases. The masking data consists of values 0 and 255, where the model's '1' (True) is 255, and '0' (False) is 0, with a binary classification threshold set at 70%. The overall quality of the test examples is poor. Despite the cross-attention method achieving the best evaluation score, the score remains low, indicating that the model complexity is insufficient to accurately perform image masking based on textual instructions.

3.3.4 Analysis

Based on the accuracy changes observed in the training loop (Figure 3), the decoding layer was insufficient to upsample and analyze the results from the cross-attention data fusion. Unlike the simple data structure of concatenation, cross-attention involves more complex data due to the interconnections between each piece of data. All three cases used the same convolutional layers; thus, simple data fusion could fit well in the training loop with low-sized convolutional layers. However, for more complex data fusion, larger convolutional layers might be necessary. Therefore, tuning the model architecture is essential to verify this.

Despite this, the evaluation scores (Table 1) show that the training and evaluation scores are aligned for the cross-attention method. This indicates that the issue with the concatenation data fusion method's training score is not inherent to the methodology itself but rather related to the encoding and decoding process. Additionally, the cross-attention method shows balanced training and evaluation scores, suggesting meaningful improvements in data fusion. In comparison, the CLIP method disrupts the model architecture entirely, indicating that cross-attention provides better data analysis for image and text fusion. However, the low scores suggest that additional architectural changes, especially in the decoder, are necessary. Figure 4's image-text experiment further indicates that the models are still inadequate in classifying the target correctly. All three methodologies used pretrained text and image encoders, so it is unlikely that the encoders cannot provide complex analysis. However, the decoders were not pretrained, implying that improving the decoders could enhance the analysis of fused data.

The dataset focuses on both semantic and positional data fusion, as the main model training aims at instruction-based image classification. In this regard, the cross-attention methodology demonstrates better semantic and positional fusion, with aligned training and evaluation scores and the best performance in the evaluation set. While the concatenation method achieved the best score in the training loop, the evaluation scores suggest overfitting.

3.3.5 Conclusion

From this project, the evaluation accuracy score of cross-attention suggests that, despite the training score being suboptimal, the alignment between training and evaluation scores indicates the model's capability to apply image and text data fusion with regularization. In contrast, the CLIP methodology failed to perform effectively, and the concatenation method exhibited overfitting. These results imply that cross-attention can provide more complex data fusion and better regularization in terms of text and image fusion, considering both positional and semantic data.

	Input Instruction	Experiment test
Concatenation	show the cows tongue	
	Make the man smile.	
	Put some fries on the plate.	
CLIP	show the cows tongue	
	Make the man smile.	
	Put some fries on the plate.	
Cross_attention	show the cows tongue	
	Make the man smile.	
	Put some fries on the plate.	

Figure 4: input examples and output

However, a fundamental limitation of this investigation is that, despite cross-attention achieving the best evaluation score, the classified images remain poor in quality. Additionally, the dataset used was not inherently designed for multimodal purposes but rather to generate images with human-made masking data. Therefore, while the image classification quality is affected by the model architecture and the original dataset's purpose, the evaluation scores offer valuable insights for methodology comparison.

Throughout this project, I gained an understanding of how multimodal data fusion works and its inherent complexities. Initially, my focus was on identifying which data fusion method could best capture semantic and positional data. However, as the project progressed, it became clear that the data sizes from pretrained encoding results are fundamentally different. The image encoder's output,

though smaller after tuning, was still about 30 times larger than that of the text encoder, introducing an additional perspective on data imbalance when using cross-attention. Despite this imbalance, cross-attention proved applicable, though the potential risks associated with data imbalance should be considered.

As noted, I believe the cross-attention methodology has the potential for application in various areas, necessitating further future work. The main limitations for the models in this project were training time and computational resource constraints. Future work could involve utilizing the entire dataset and incorporating much larger CNN or GAN decoders to clarify that the fused data includes complex information from both text and images. By using only label data without masking images, the model could potentially generate modified images more effectively.

References

[1] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator, 2015.

[2] R. Gnana Praveen and Jahangir Alam. Recursive joint cross-modal attention for multimodal fusion in dimensional emotion recognition, 2024.

[3] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions, 2023.

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[5] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching, 2020.

[6] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

4 Ethical Consideration

A significant ethical issue in the fusion of text and image data is the potential for bias in the correlation between images and text. For example, the dataset used for this fusion might associate certain races or genders with specific texts, but the corresponding images may not accurately reflect those attributes. This can lead to critical misclassifications in tasks such as race identification, resulting in unfair and biased outcomes. To address this issue, it is essential to use a balanced and unbiased dataset. Additionally, thorough evaluation methodologies should be employed to detect and mitigate any biases that may arise during model training and deployment.

Privacy is another critical ethical issue when compiling datasets for text and image correlation. To capture a wide range of textual commands, a vast array of images is required. While many public and positively connotated images can be used, some text commands might relate to negative or private contexts, necessitating careful handling of such data. For instance, if a text command includes a specific name and the dataset contains images revealing that name, it could lead to privacy violations. To mitigate this, AI alignment techniques can be applied. During pre-training, datasets should be modified or preprocessed to generalize specific terms or private information, such as replacing a person's name with more general terms like "man" or "human." In the post-training phase, additional training loops, human feedback, or evaluation processes should be implemented to ensure that the alignment of text and images does not result in privacy breaches, even if the input data contains sensitive information.

By addressing these ethical considerations through careful dataset construction and robust evaluation methodologies, we can reduce the risks of bias and privacy violations in multimodal data fusion systems.