Multimodal Fusion of Text and Image Data using Cross-Attention Mechanism

Anumula Chaitanya Sai
anumulacs@iitbhilai.ac.in

Sunkara Vamsi Krishna
sunkaravamsi@iitbhilai.ac.in

Gandu Sai Prabhath
gandusai@iitbhilai.ac.in

Rekha
kondurinaga@iitbhilai.ac.in

Abstract

Human cognition integrates diverse sensory inputs like sight, smell, and touch, whereas computational models often rely on a single data type. Multimodal approaches aim to bridge this gap by leveraging and analyzing multiple data types simultaneously, enabling more human-like analytical capabilities. However, effectively combining modalities such as image and text data remains a challenge due to their fundamental differences: images contain positional and semantic information, while text primarily conveys semantic information.

Traditional fusion methods, such as dimensional alignment and concatenation, often result in the loss of positional information inherent to images, limiting the depth of analysis. Addressing this requires techniques that preserve both semantic and positional features for more robust and flexible multimodal integration.

This study evaluates two primary fusion methods: cross product mapping (e.g., CLIP-based methods), and cross-attention mechanisms, we are not considering concatenation because experiments show it compromise to learn text to image positional information relation. In contrast, cross-attention mechanisms apply attention to both text and image features, enabling detailed integration by aligning text data with each pixel of the image.

Through qualitative and quantitative evaluations, this research highlights cross-attention as a superior approach for preserving the unique characteristics of both modalities, resulting in more fluent and effective multimodal fusion.

1. Introduction and Problem Motivation

Motivation: Human cognition seamlessly integrates diverse sensory inputs, enabling nuanced analysis of multi-modal data. In contrast, computational models typically rely on single-modal data or struggle to effectively combine distinct modalities, such as text and image data. A major challenge in multi-modal systems lies in preserving the unique characteristics of each modality during fusion, particularly the positional information inherent in image data. Common approaches, such as concatenation and mapping methods like CLIP, effectively merge modalities but often lose critical positional details. This limitation hinders accurate alignment between textual descriptions and image features, restricting the applicability of multi-modal systems in tasks requiring fine-grained image-text interactions. Addressing this gap is crucial for advancing the flexibility and fluency of multi-modal systems in real-world applications.

Objective: The primary objective of this research is to develop and evaluate methods for multi-modal data fusion that preserve both semantic and positional information in text and image data. Specifically, the study explores and compares cross product, and cross-attention mechanisms to determine their effectiveness in achieving accurate text-image alignment. By leveraging the cross-attention mechanism, this research aims to enable models to learn relationships between text features and image pixels, maintaining positional information. The ultimate goal is to enhance the capability of multi-modal systems to process and generate contextually coherent outputs, evaluated through image generation and performance metrics.

2. Experimentation

Model Architecture

Text Encoding with BERT

The text description is passed through a pre-trained BERT model to generate embeddings. The output of BERT for each input text is a vector of size 768:

$$\mathbf{t} = \text{BERT}(\text{Text Description}) \in \mathbb{R}^{768}$$

This vector $\mathbf{t}$ represents the text embedding and serves as the input query for the attention mechanism.

Image Encoding with ResNet50 and Adapter Module

The image is passed through ResNet50 to obtain a high-dimensional embedding. ResNet50 is a pre-trained convolutional network that generates embeddings of size d_{image}. Since the dimensions of the image embeddings differ from the text embeddings, we use Adapter Modules to project the image embeddings into the same 768-dimensional space as the text embeddings.

$$\mathbf{i} = \text{ResNet50}(\text{Image}) \in \mathbb{R}^{d_{\text{image}}}$$

Adapter Modules then project $\mathbf{i}$ into the same space as $\mathbf{t}$:

$$\mathbf{i}_{\text{adapted}} = \text{Adapter}(\mathbf{i}) \in \mathbb{R}^{768}$$

Cross Attention and Fusion Strategies

We experiment with three distinct strategies for combining the text and image embeddings.

Case 1: Cross-Multiply Embeddings

In this case, we directly compute the element-wise multiplication (cross product) between the text and image embeddings:

$$\mathbf{C} = \mathbf{t} \times \mathbf{i}_{\text{adapted}} \in \mathbb{R}^{768}$$

This results in a correlation matrix $\mathbf{C}$, which captures the interaction between the two embeddings. We then use this matrix to reconstruct the object mask.

Case 2: Simple Cross-Multiply Embeddings

In this variant, we skip the attention mechanism and directly multiply the embeddings without any processing, similar to Case 1:

$$\mathbf{C} = \mathbf{t} \times \mathbf{i}_{\text{adapted}} \in \mathbb{R}^{768}$$

This results in a correlation matrix, which is then used for mask reconstruction.

Case 3: Self-Attention + Cross-Attention

For this case, we first apply self-attention to the individual embeddings (text and image). Self-attention allows the model to focus on different parts of each embedding based on the context.

$$\mathbf{t}_{\text{self}} = \text{SelfAttention}(\mathbf{t})$$
$$\mathbf{i}_{\text{self}} = \text{SelfAttention}(\mathbf{i}_{\text{adapted}})$$

We then apply cross-attention between the text embedding (as query) and the image embedding (as key-value pair). The output of this cross-attention mechanism is used to reconstruct the object mask:

$$\mathbf{R} = \text{CrossAttention}(\mathbf{t}_{\text{self}}, \mathbf{i}_{\text{self}}) \in \mathbb{R}^{768}$$

The result of the cross-attention mechanism is used for mask reconstruction.

Mask Reconstruction

For all three cases, the final step is to generate the object mask using a series of transposed convolution layers (ConvTranspose). These layers map the final embedding into a binary mask image. The final reconstruction is given by:

$$\mathbf{M} = \text{ConvTranspose}(\mathbf{C} \text{ or } \mathbf{R}) \in \mathbb{R}^{H \times W}$$

Where $\mathbf{M}$ is the predicted binary mask of the object. The mask is of size $H \times W$, where H and W are the height and width of the output mask image.

Loss Function

The model is trained using **Binary Cross-Entropy (BCE)** loss, which measures the difference between the predicted mask and the ground truth mask. BCE is suitable for binary classification tasks, where the target mask values are either 0 (background) or 1 (object). The BCE loss function is defined as:

$$\mathcal{L}\text{BCE} = -\frac{1}{N} \sum i = 1^N \left(y_i \log(\hat{y_i}) + (1 - y_i) \log(1 - \hat{y_i})\right)$$

Where:

- y_i is the ground truth mask value at pixel i,
- $\hat{y_i}$ is the predicted mask value at pixel i,
- N is the total number of pixels in the mask image.

This loss function is minimized during training to optimize the model's parameters and improve the accuracy of the predicted binary masks

Loss Functions

In addition to Binary Cross-Entropy (BCE) loss, we explore the use of two other popular loss functions for evaluating the model's performance in predicting object masks: *BCE + IoU loss* and *BCE + Dice loss*. These hybrid losses combine the BCE loss with the Intersection over Union (IoU) and Dice coefficients, respectively, to address issues related to class imbalance and improve the accuracy of the mask predictions.

BCE + IoU Loss

The *BCE + IoU loss* combines the Binary Cross-Entropy loss with the Intersection over Union (IoU) metric. While BCE is effective in pixel-wise classification, it may not perform well when there is a significant class imbalance (e.g., most pixels are background). The IoU loss helps by penalizing the model based on the overlap between the predicted mask and the ground truth mask.

The BCE + IoU loss function is defined as:

$$\mathcal{L}\text{BCE + IoU} = \mathcal{L}\text{BCE} + \lambda \mathcal{L}_{\text{IoU}}$$

Where: - $\mathcal{L}_{\text{BCE}}$ is the Binary Cross-Entropy loss, - $\mathcal{L}_{\text{IoU}}$ is the Intersection over Union (IoU) loss, - λ is a hyperparameter to control the weight of the IoU term.

The IoU loss is calculated as:

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{\text{Intersection}(y, \hat{y})}{\text{Union}(y, \hat{y})}$$

Where: - y is the ground truth mask, - $\hat{y}$ is the predicted mask, - The intersection is the number of pixels where both masks are 1 (i.e., the overlap), - The union is the total number of pixels where either of the masks is 1.

BCE + Dice Loss

The *BCE + Dice loss* combines Binary Cross-Entropy loss with the Dice coefficient, a metric that is especially effective when dealing with imbalanced classes. The Dice coefficient measures the similarity between two sets, and it is often used in segmentation tasks to evaluate the overlap between the predicted and ground truth masks.

The BCE + Dice loss function is defined as:

$$\mathcal{L}\text{BCE + Dice} = \mathcal{L}\text{BCE} + \lambda \mathcal{L}_{\text{Dice}}$$

Where: - $\mathcal{L}_{\text{BCE}}$ is the Binary Cross-Entropy loss, - $\mathcal{L}_{\text{Dice}}$ is the Dice loss, - λ is a hyperparameter to control the weight of the Dice term.

The Dice loss is computed as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \cdot \text{Intersection}(y, \hat{y})}{|\hat{y}| + |y|}$$

Where: - Intersection$(y, \hat{y})$ is the number of pixels where both the predicted mask and ground truth mask are 1, - $|\hat{y}|$ is the number of pixels in the predicted mask, - $|y|$ is the number of pixels in the ground truth mask.

The Dice coefficient ranges from 0 (no overlap) to 1 (perfect overlap). Therefore, the Dice loss is minimized when the Dice coefficient is maximized, leading to better segmentation accuracy.

MagicBrush dataset:

It is a large-scale, manually annotated resource designed for instruction-guided image editing and spatial understanding tasks. It consists of approximately 10,000 annotated samples in the form of triplets-Source Images, Instructions, Masks.

Source Images: The visual inputs representing the initial state.

Instructions: Textual descriptions specifying regions or features of interest within the source image.

Masks: Binary images where white areas denote regions corresponding to the instruction, and black areas represent untouched regions.

Dataset Details:

Training Set: 8,807 samples.

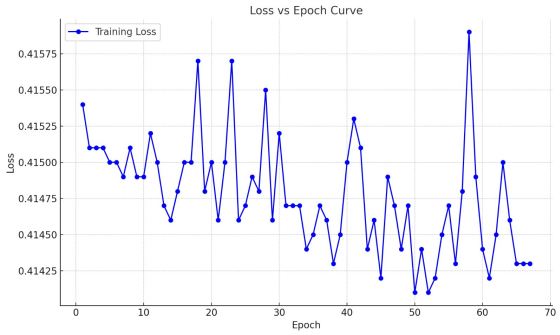
Validation Set: 528 samples.

Test Set: Reserved for evaluation purposes.

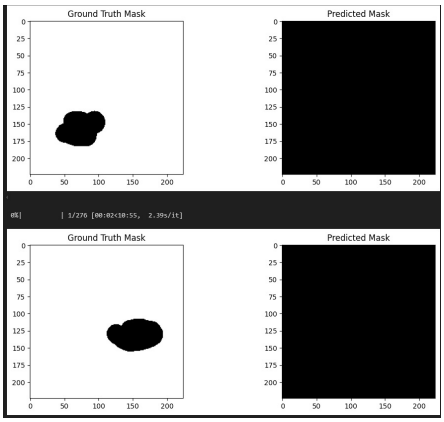
In our project, we exclusively use the source images and instructions as inputs to train a model to predict the corresponding mask images. This approach helps the model learn spatial reasoning guided by natural language.

3. Results

Our Baseline CLIP results are these:

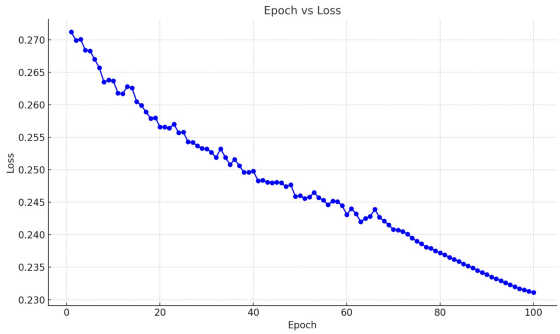


(a) CLIP

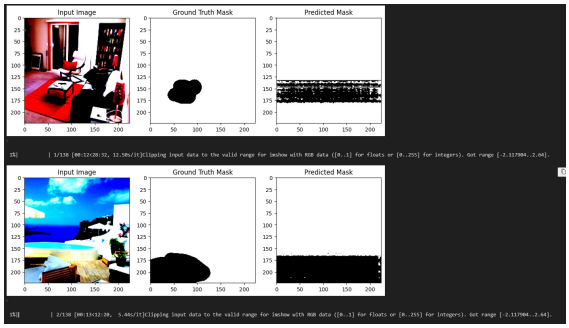


(b) CLIP

The results of cross-product approach:

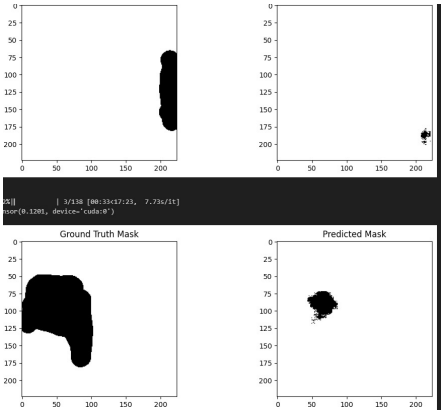


(a) Average IoU: 0.1991



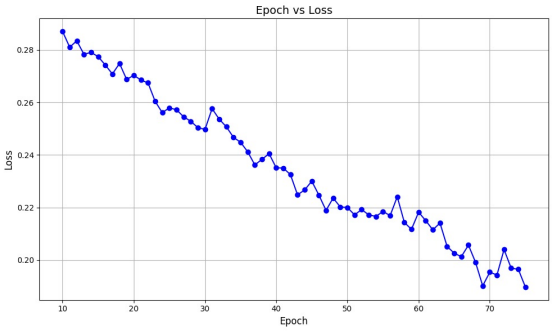
(b) Cross product

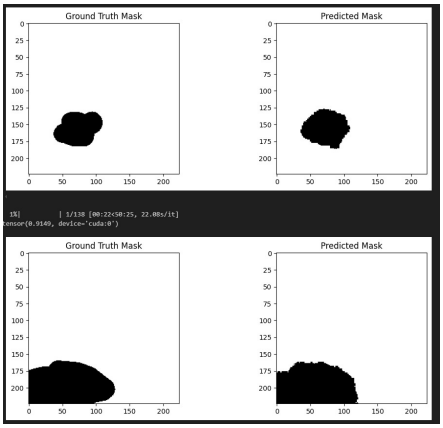
The results with Multi-Head Cross Attention using BCE+IOU(Jaccard distance) Loss :



(a) BCE+IOU with a result of Average IoU: 0.2604

The results with Cross Attention with BCE+DICE Loss:





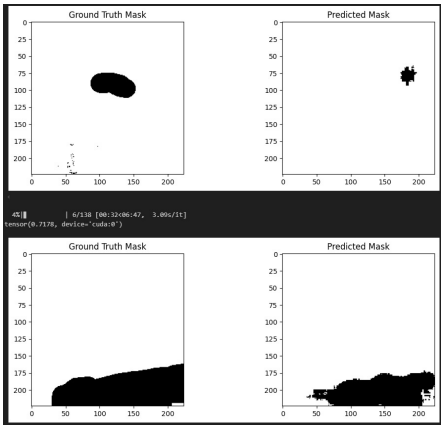
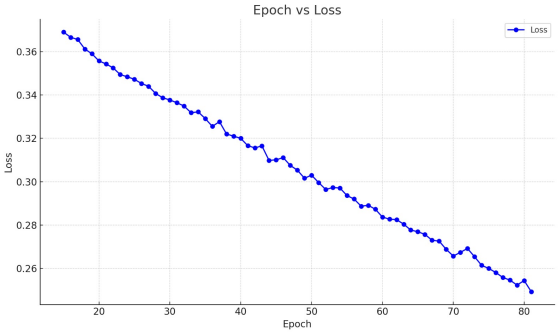
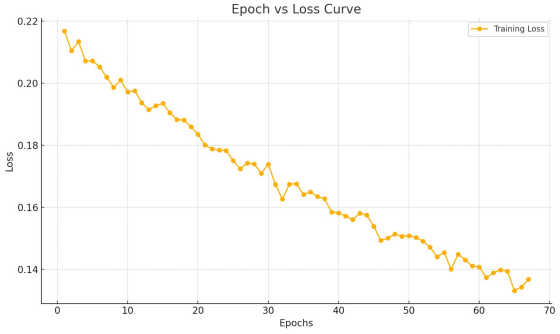


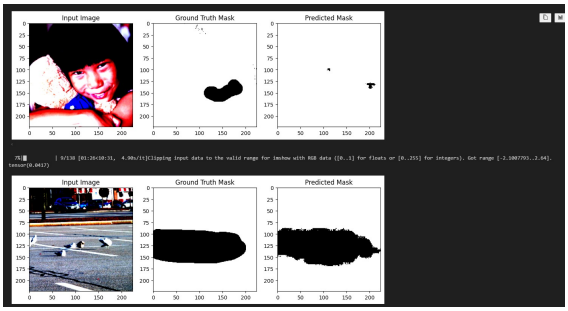
Рис. 5: DICE + BCE with a result of Average IoU: 0.6186



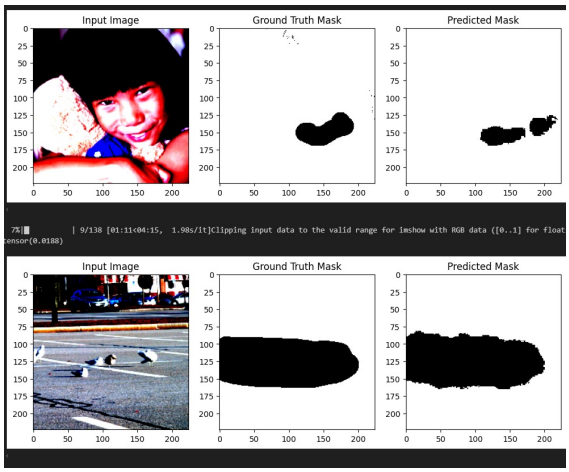
(a) CROSS ATTENTION + BCE with 100 epochs results in Average IoU: 0.9166



(b) CROSS ATTENTION + BCE with 200 epochs results in Average IoU: 0.9617



(a) BCE with 100 epochs



(b) BCE with 200 epochs

The results on using Multi-head Cross Attention with BCE Loss:
Observations:

- The Clip model performs poorer and according to the experimentation results, it seems to not encode the semantics in the image encoding, which lead to poorer understanding of relation between text and image
- Cross Product is not a good approach to understand the relationship between different modalities, but multi-head cross attention gave very good results on test data, the best IOU was **0.96**
- While using multiple combined loss functions, it seemed that only using BCE loss was suitable and other approaches like BCE+IOU, BCE+DiceLoss didn't work well

Our code is available here - **https://github.com/sunkustar/MultiModal-Image-Masking-by-text**

References

1. TokenCompose: Text-to-Image Diffusion with Token-level Supervision (Zirui et. al.)
2. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On.
3. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations (Liu et. al.)
4. CLIP - https://github.com/sunkustar/MultiModal-Image-Masking-by-text
5. Stanford proj - https://web.stanford.edu/class/cs224n/final-reports/256711050.pdf
6. Pytorch Documentation