
Jinhai Wang, School of Foreign Languages, Zhengzhou University of Aeronautics, China
Yi Tie, School of International Studies, Zhengzhou University, China*
Xia Jiang, School of Foreign Languages, Zhengzhou University of Aeronautics, China
Yilin Xu, School of Business, Zhengzhou University of Aeronautics, China

ABSTRACT

There is a wide connection between linguistics and artificial intelligence (AI), including the multimodal language matching. Multi-modal robots possess the capability to process various sensory modalities, including vision, auditory, language, and touch, offering extensive prospects for applications across various domains. Despite significant advancements in perception and interaction, the task of visual-language matching remains a challenging one for multi-modal robots. Existing methods often struggle to achieve accurate matching when dealing with complex multi-modal data, leading to potential misinterpretation or incomplete understanding of information. Additionally, the heterogeneity among different sensory modalities adds complexity to the matching process. To address these challenges, we propose an approach called vision-language matching with semantically aligned embeddings (VLMS), aimed at improving the visual-language matching performance of multi-modal robots.

KEYWORDS

Computer Vision Advancements, Improved Interaction Capabilities, Multi-Level Matching, Multi-Modal Robots, Text-Image Matching, Visual-Language Matching

INTRODUCTION

Multi-modal robots are intelligent robot systems capable of processing and understanding multiple sensory modalities, typically including vision, hearing, language, and touch (Duduta et al., 2020). These robots integrate different sensory data to interact with their environment and users in a more comprehensive and sophisticated manner. The multi-sensory capabilities allow them to gather information from different perspectives and gain a more complete understanding of their surroundings (Wellhausen et al., 2020; Chen & Du, 2022). Multi-modal robots typically understand and generate natural language, enabling them to engage in language-based interactions with human users. This allows them to answer questions, perform tasks, or provide information (Luo & Qin, 2012). These
robots can perceive and comprehend their surrounding environment, including obstacles, human behavior, and other robots. This enables them to autonomously navigate and make decisions. Multi-modal robots can interact with human users, through emotion recognition and expression, facial expression analysis, and speech synthesis and recognition. This enhances their ability to interact with people (Tang et al., 2023; Du & Chen, 2023).

It is important to note that the development of multi-modal robots involves complex hardware and software engineering to achieve the integration and collaboration of various sensory modalities. Additionally, with technological advancements, multi-modal robots are expected to play a more significant role in various fields in the future, providing innovative solutions. In automation and manufacturing, multi-modal robots play a crucial role in industries and manufacturing by automating tasks on production lines, such as assembly, packaging, welding, and quality control (Anzai & Takahashi, 2020). They enhance production efficiency and consistency in quality. In the healthcare realm, multi-modal robots are used for tasks including surgical assistance, rehabilitation therapy, patient monitoring, and medication dispensing. They improve surgical precision, aid in patient recovery, and reduce the workload of healthcare professionals (Wellhausen et al., 2020). Autonomous vehicles, which are a form of multi-modal robots, can sense their surroundings and drive autonomously. They have the potential to reduce traffic accidents, enhance road safety, and provide smoother traffic flow. With the function of customer service, multi-modal robots can provide assistance, answer questions, and handle customer inquiries, both online and in physical stores (Mohd et al., 2022). They can also provide services in education, such as teaching assistance, support for personalized learning, and assistance for students with special needs (Taunyazov et al., 2009). In summary, multi-modal robots, through the integration of multiple sensory modalities, can perform a wide range of tasks, improving efficiency, safety, and user experiences. As technology continues to advance, their applications will continue to expand.

The role of cross-scene image-text matching algorithms in multi-modal robots is to effectively associate and match image and text information, enabling more intelligent and diverse robot interactions and decision-making. There are several roles of this algorithm in multi-modal robots (Song et al., 2022). First, multi-modal robots can receive information from different sensors, such as images captured by cameras and text descriptions provided by users. Cross-scene image-text matching algorithms help integrate these different modalities of information to better understand user needs and the environment (Kapelyukh et al., 2023). Secondly, the algorithm helps to understand the semantic relationships between images and text. For example, when a user provides a text description, the algorithm can validate or complement these descriptions with images, improving the robot’s understanding of user intent. Thirdly, cross-scene image-text matching algorithms can be used for information retrieval, helping the robot search for relevant information in different scenarios. It can also be used for personalized recommendations, suggesting relevant content or services based on user-provided information and preferences. The next robots in multi-modal environments may need to make complex decisions, such as navigation, object recognition, and sentiment analysis. Cross-scene image-text matching algorithms can provide key information to help robots make more accurate decisions (Khachatryan et al., 2023). Finally, the algorithm can also enhance natural language interaction. When users ask questions or make requests, robots can use cross-scene image-text matching algorithms to analyze the questions and provide answers that are more informative and contextually relevant. In summary, the role of cross-scene image-text matching algorithms in multi-modal robots is to facilitate the integration and interaction of multi-modal information, enhance the robot’s intelligence, and enable it to better adapt to different scenarios and user needs. This helps improve the practicality and user experience of robots, enabling them to better handle a variety of tasks and application scenarios (Al Faraby et al., 2020).

In the realm of computer vision and natural language processing, a fascinating research area is the intersection of image and text matching. This interdisciplinary study aims to develop algorithms and models that can effectively recognize and understand the relationship between visual and linguistic
data (Wu et al., 2023). One popular approach is to use deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to extract features from images and text, respectively (Shu & Xu, 2019). These feature vectors are then combined and fed into a matching model, which aims to find the most relevant associations between them. Another line of research explores the use of attention mechanisms, which allow the model to selectively focus on specific parts of the input data when making predictions (Liu et al., 2019). By doing so, the model can improve its accuracy and efficiency in handling complex and ambiguous relationships between images and text. Despite these advancements, there are still many challenges to be overcome in this field, such as dealing with imbalanced datasets, improving the interpretability of the models, and developing more robust and efficient matching algorithms. Nevertheless, the potential applications of this technology are vast, ranging from content-based image retrieval and question answering systems to automated image captioning and visual questioning tasks (Wang et al., 2019; Ye & Zhao, 2023).

Our research proposes three key contributions in the context of text-to-image synthesis:

- The VLMS model possesses multi-level matching capabilities, including local matching, global matching, and general matching. This means that it can understand the connections between text and images in different contexts, thereby achieving more accurate matching of multimodal data.
- The VLMS model embeds image and text information into semantically aligned vectors using advanced models such as ResNet-152 and BiGRU. It effectively fuses information from the visual and language domains, enabling multimodal robots to better understand and interact.
- Compared to existing methods, the VLMS model has made significant improvements in visual and language matching. This validation provides strong support for the application of the VLMS model in the field of multimodal robotics, with the potential to enhance efficiency, accuracy, and user experience.

We propose a method for VLMS aimed at capturing multi-level similarities between text and image patterns. It comprises local matching, global matching, and general matching. Experimental results demonstrate a significant and noteworthy improvement in performance compared to previous methods. The subsequent sections of this paper are organized as follows: section two briefly reviews relevant work in the field of text-to-image synthesis. Section three introduces the methodology behind VLMS that we propose. In section four, we present and analyze the experimental results. Finally, in section five, we provide the research conclusions and outline directions for future research.

**RELATED WORK**

Today, there is a rapidly growing field of research in multi-modal algorithms, which has become one of the forefront areas in computer vision and natural language processing (Yao et al., 2021). This field focuses on combining images and text to create more powerful and intelligent computer systems. These multi-modal algorithms not only enable computers to better understand and process image and text data, but also allow for mutual enhancement between these two modalities, driving innovation across various domains. The applications of these multi-modal algorithms are diverse, spanning image captioning, text generation, visual reasoning, image-text retrieval, sentiment analysis, autonomous driving, healthcare, and more (First, 2023). They not only provide computers with deeper perceptual capabilities, but they also offer richer and more natural ways for people to interact with computer systems.

At the forefront of multi-modal research are advancements in deep learning techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), as well as pre-trained models like BERT and GPT, which provide higher performance and flexibility in the fusion of image and text. Additionally, the creation of large-scale multi-modal datasets has provided strong support for research in this field. Five algorithms are noted below (Li et al., 2020).
Language-Image Pre-Training Based on Contrastive Learning

CLIP is one representation of these kinds of method (Radford & Kim, 2021). CLIP’s primary advantage lies in its ability to create a joint image-text representation space, allowing it to effectively bridge the gap between vision and language. This versatility enables zero-shot learning and supports a wide range of vision-language tasks. CLIP learns from a large-scale dataset, making it robust and capable of generalizing across diverse domains. However, one of its disadvantages is its computational complexity, which can be challenging for deployment in resource-constrained environments (Gu et al, 2021). Additionally, fine-tuning CLIP for specific tasks might require a substantial amount of task-specific data, which can be a limitation in scenarios with limited annotated data.

Multi-Scale Matching

DAMSM is a noted algorithm in multi-scale matching. It excels at fine-grained image-text matching, offering precise alignment between image regions and words in text (Lu et al., 2023). This level of granularity makes it suitable for tasks where detailed correspondences are crucial. It leverages dual-attention mechanisms, enhancing its ability to capture compositional and semantic relationships. However, DAMSM’s focus on fine-grained matching may not be necessary or ideal for all applications, and it may be computationally more demanding compared to models with simpler architectures. Its performance may also heavily depend on the quality and quantity of available training data (Liu et al., 2021; Ye et al., 2023).

Semantic Composition

Semantic Composition is based on Semantic Compositional Attention Network (SCAN) (Russin et al., 2019). SCAN addresses the challenge of capturing both semantic and compositional relationships between images and text. It does so by incorporating hierarchical structures and attention mechanisms. This design offers a balanced approach that can handle both fine-grained and holistic matching tasks. SCAN’s hierarchical structure allows it to capture varying levels of abstraction in both images and text. Nevertheless, similar to DAMSM, SCAN’s increased complexity may come with higher computational requirements, and its performance can be sensitive to the quality and diversity of training data (Furrer et al., 2020; Liu & Chen, 2023).

Image-Text Representation Learning

The Universal Image-Text Representation Learning (UNITER) algorithm is working on image-text representation learning. UNITER focuses on creating a unified representation space for images and text, making it versatile for various vision-language tasks (Paraschiv et al., 2022; Deng et al., 2021). It achieves state-of-the-art results on multiple benchmarks, showcasing its effectiveness in cross-scene image-text matching. UNITER’s architecture is designed to capture rich contextual information and relationships between visual and textual content. Despite its strengths, training a model like UNITER may require significant computational resources, which could be a limitation in some settings (Chen et al., 2023; Chen & Zhang, 2022; Liu et al., 2021). Additionally, fine-tuning UNITER for specific applications might still require substantial task-specific data.

METHODOLOGY

The VLMS model is designed to capture multi-level similarities between the text and image modalities. It encompasses local-level matching, global-level matching, and general-level matching. Figure 1 illustrates the architecture of the VLMS model, which comprises three sub-models: the vision encoder, textual encoder, and matching scoring block (MSB). The primary objective of the vision encoder and text encoder is to embed images and text into semantically aligned vectors. This embedding process is pivotal for establishing connections between the domains of vision and language. To ensure a fair
comparison, we utilize the same backbone architecture, ResNet-152, for image features extraction and a bidirectional recurrent neural network (BiGRU) for processing textual information. ResNet-152 is a widely-recognized model for extracting visual features, while BiGRU is chosen for its ability to handle long-distance dependencies in natural language processing. The proposed MSB functions as a transformer encoder responsible for generating a matching score that quantifies the alignment between the image and text modalities. The overall structure of the model is shown in Figure 1.

The visual-language matching model contains three sub-models: vision encoder, text encoder and matching scoring block.

**ResNet-152**

The ResNet-152 neural network model is a state-of-the-art deep learning architecture designed for image classification tasks (Malali & Keller, 2021). It was introduced in 2018 and has since become one of the most widely used models in the field of computer vision. It is based on the ResNet architecture, which consists of a series of stacked residual blocks that progressively learn to extract more complex features from the input data. In contrast to previous ResNet models, ResNet-152 introduces a wider range of filter sizes, resulting in more accurate and robust feature extraction. At its core, ResNet-152 uses dilated convolutions, which enable the model to process input images with varying sizes and aspect ratios.

Additionally, it incorporates depth wise separable convolutions, which reduce the number of parameters and computational requirements, making it faster and more efficient than previous models. One of the key advantages of ResNet-152 is its ability to handle large datasets and achieve high accuracy rates. It has been successfully applied to a wide range of image classification tasks, including object detection, semantic segmentation, and face recognition, among others. The network architecture of ResNet-152 is shown in Figure 2.

**BiGRU**

When engaging in language integration models, one of the most straightforward approaches is to employ the bag-of-words (BOW) model, a commonly used technique in various applications like cross-modal retrieval (Yan et al., 2020). However, BOW lacks the ability to capture the contextual
meaning of text descriptions, leading to a shift towards more sophisticated learning-based models, such as LSTM and Skip-gram (Das & Saha, 2021). In these models, each word in a sentence is transformed into a semantic vector through an embedding layer before being input into the LSTM. Specifically, in text analysis, the word characteristics are represented by the hidden states, and the overall sentence meaning is conveyed by the last hidden state.

In this research, we opted for the bidirectional gated recurrent unit (BiGRU) model, which excels in considering contextual information in text (Chen et al., 2022). Compared to traditional unidirectional
LSTMs, BiGRU has the advantage of capturing both forward and backward information in the text, thus gaining a better understanding of semantic relationships within the text. The strength of choosing this model lies in its ability to comprehensively analyze textual data, thereby enhancing performance in tasks related to semantic comprehension and others. Therefore, our study employs the BiGRU model to further improve the quality and effectiveness of text language embeddings, better supporting the tasks under investigation (Zhang et al., 2022; Liu & Chen, 2021).

The bidirectional and iterative gated recurrent unit (BIGRU) neural network model is a type of recurrent neural network (RNN) that has gained significant attention in recent years due to its ability to capture long-range dependencies in sequential data. Unlike traditional RNNs, which process data in a unidirectional manner, BIGRU models enable information to flow in both forward and backward directions, allowing them to capture longer-term dependencies and improve the accuracy of predictions. The architecture of BIGRU consists of two stacked GRU layers, each followed by a gate layer that controls the flow of information in the opposite direction. The gate layer uses a parameter called reset gate and update gate to determine whether new information should be added or discarded from the input sequence. This bidirectional processing allows BIGRU to capture both past and future context, resulting in more accurate and robust predictions. BIGRU has been successfully applied to a wide range of natural language processing tasks, such as machine translation, text classification, and sentiment analysis, among others. Its ability to handle long sequences and capture complex dependencies makes it a powerful tool for dealing with large amounts of textual data. The BIGRU neural network model represents a significant advancement in the field of recurrent neural networks, offering a promising solution for handling sequential data with long-range dependencies. The network architecture of BIGRU is shown in Figure 3.

Matching Scoring Block

The purpose of the vision-language matching scoring block is to generate a comprehensive matching score for assessing how well an image and text correspond to each other. Transformer has demonstrated strong performance in various tasks, particularly in the realm of vision-language comprehension, much like BERT and UNITER. The self-attention mechanism in Transformer is effective in deeply exploring the semantic relationships between visual and textual features. Consequently, we have chosen to employ this mechanism for learning the matching score between the image and text. This general-level matching calculation takes into account both the global and local features of the image, as well as the sentence and word features of the text. Specifically, we define the vision-language united feature as follows:

$$\psi = F_{cat}(\varphi, \varphi, \phi, \phi)$$

In this context, where $F_{cat}$ represents the concatenation operation, the united feature is subsequently inputted into the vision-language encoder based on the Transformer, as outlined below:

$$\hat{\psi} = F_{Transformer}(\psi)$$

Once the latent features combining visual and textual information have been acquired, we select a fully connected layer to map these features into a hidden space:

$$\tilde{\psi} = W_0 \hat{\psi} + b_0$$
Here, $W_0$ and $b_0$ represent the trainable parameters of the fully connected layer. The ultimate feature can be obtained through mean pooling:

$$\bar{\psi} = F_{\text{mean pooling}}(\psi)$$

Finally, the score assessing the match between vision and language is determined by applying a Sigmoid function following a fully connected layer, which transforms the feature into a single-dimensional value:

$$\text{Score} = F_{\text{Sigmoid}}(W_1\bar{\psi} + b_1)$$

Here, $W_1$ and $b_1$ represent the trainable parameters of the second fully connected layer. The entire procedure for acquiring the visual-language matching score can be represented using the following formula:

$$\text{Score} = F_{\text{MSB}}(\phi, \bar{\phi}, \varphi, \bar{\varphi})$$

**Local-Level Matching**

The local-level matching focuses on assessing the semantic alignment between word features and localized image features. For a particular image containing local regional features and a text description comprising $T_0$ word features, we compute the cosine similarity for every conceivable pairing of image regions and words using the following formula:

$$s(i, j) = \frac{\phi_i^T \varphi_j}{\phi_i \varphi_j}, i \in [1, 2, \ldots, 289], j \in [1, 2, \ldots , T_0]$$

In this context, $s(\phi_i, \varphi_j)$ represents the similarity between the i-th image region and the j-th word. We use $S$ to denote the similarity matrix between word features and localized image features. To capture nuanced similarities, we leverage the widely used attention mechanism. The word context relative to each image region is determined by computing a weighted sum of the image’s visual features, as described below:

$$c_i = \sum_{j=0}^{288} \alpha_{ij} \phi_j$$

where:

$$\alpha_{ij} = \frac{\exp(\gamma_i s_{i,j})}{\sum_{k=0}^{288} \exp(\gamma_i s_{i,k})}$$
Following the approach employed in minimum classification error formulation within the field of speech recognition, we compute the local-level matching score between the image and the textual description using the LogSumExp pooling method, as outlined below:

\[
S_{local}(I, T) = \log \left( \sum_{i=1}^{T-1} \exp \left( \gamma S(c_i, \varphi) \right) \right)^{1/2}
\]

In this context, \( S(c_i, \varphi) \) represents the matching score computed using cosine similarity between the \( i \)-th word and the \( i \)-th region-context:

\[
S(c_i, \varphi_i) = \frac{c_i^T \varphi_i}{c_i \varphi_i}
\]

**Global-Level Matching**

The global-level matching takes into account the overall visual feature and the complete textual feature. Likewise, when dealing with the global visual feature \( \varphi \) and the sentence feature \( \phi \), the matching score is directly computed using cosine similarity:

\[
S_{global}(I, T) = \frac{\varphi^T \phi}{\varphi \phi}
\]

**General-Level Matching**

The matching score at the general level is generated using the pre-trained vision-language matching scoring block, as illustrated below:

\[
S_{general}(I, T) = F_{MSB} (\phi, \varphi, c, \bar{c})
\]

**Objective Function**

Triplet loss is a well-known ranking objective commonly applied to matching tasks, and it has found widespread usage in tasks involving image-text matching. After obtaining matching scores at three different levels, we employ a hinge-based triplet ranking loss to fine-tune and optimize the vision-language matching model. The optimization loss for the general-level matching is defined as follows:

\[
L_{general} = \left[ \alpha + S_{general}(I', T) - S_{general}(I, T) \right]_+ + \left[ \alpha + S_{general}(I, T') - S_{general}(I, T) \right]_+
\]

In this context, we have \( S_{general}(I', T) \) and \( S_{general}(I, T') \) representing the general-level matching scores for non-matching image-text instances, and \( S_{general}(I, T) \) indicating the general-level matching scores for matching image-text pairs. Additionally, \( \alpha \) represents the margin, and in our experiments, we’ve set it to 0.2. If, in the joint embedding space, the distance between the image and text is less than \( \alpha \) compared to any negative pairs, the hinge loss becomes zero. By substituting the general-level
matching score with the local-level and global-level matching scores, we can derive the losses $L_{local}$ and $L_{global}$.

In conclusion, we establish the comprehensive loss function for the VLMS model as:

$$L_{VLMS} = L_{local} + L_{global} + L_{general}$$

**Optimization**

The optimization process for the VLMS encompasses two distinct phases. The initial phase involves training to oversee text-to-image synthesis, optimizing the model exclusively on the training dataset. In the subsequent phase, the focus shifts to optimizing the MSB to acquire the VLMS for performance evaluation. During this phase, training occurs on the complete dataset, comprising both the training and testing data. We employ the Adam optimizer with a learning rate of 0.0002, and the training process concludes after 200 epochs.

**EXPERIMENT**

In this section, we conduct a comprehensive series of experiments to validate the efficacy of our proposed approach. We begin by introducing the experimental configurations, followed by the presentation of both quantitative and qualitative evaluation findings. Finally, we provide ablation studies and engage in further discussions.

**Experiment Settings**

**Datasets**

The proposed method’s capability is showcased using four extensively utilized benchmarks, Microsoft Common Objects in Context (MSCOCO) (Chun et al., 2022), LAION-400M (Ma et al., 2023), Conceptual Captions (Kim et al., 2022), and Visual Genome (Abdulmumin, et al., 2022). MSCOCO is a widely recognized multimodal dataset consisting of a large collection of images along with natural language descriptions. It serves as a benchmark for various computer vision and natural language processing tasks, including image captioning, object detection, and text-image matching. MSCOCO’s images are diverse and rich in context, making it a valuable resource for research in the field of multimodal AI. Language, images, and objects at large scale (LAION-400M) is a relatively new multimodal dataset. It is designed to facilitate research in multimodal understanding and reasoning. The dataset includes a vast amount of textual data, images, and object-centric information, making it suitable for tasks such as text-image alignment, scene understanding, and knowledge extraction. Conceptual Captions is a dataset introduced by Google, containing a large number of images paired with textual descriptions. It serves as a valuable resource for research in image captioning, text generation, and cross-modal understanding. The dataset’s captions are generated by users, providing a diverse and creative set of descriptions for the images. Visual Genome is a comprehensive multimodal dataset that combines images, text, and knowledge graph information. It is designed for tasks such as visual reasoning, text generation, and image understanding. Visual Genome’s unique feature is the inclusion of detailed scene graphs that describe the relationships between objects in images, enhancing its suitability for complex multimodal research.

**Evaluation Metrics**

We quantify the effectiveness of the proposed method in terms of area under the ROC curve (AUC) (Christensen et al, 2023), accuracy RATE, F1-score (Miura et al., 2020), recall rate, and specificity (SPE).
AUC is a widely used metric in deep learning and machine learning to assess the performance of binary classification models. It is a measure of a model’s ability to distinguish between positive and negative classes across different thresholds. Computing AUC needs a ROC curve. The curve is a graphical representation of a binary classifier’s performance at various discrimination thresholds. It plots the true positive rate against the false positive rate at different threshold values. Once the ROC curve is obtained, the AUC is calculated by computing the area under the curve. This area ranges from 0 to 1. The AUC value can be interpreted as follows: 1) AUC = 0.5: random classifier. 2) 0.5 < AUC < 0.7: poor classifier. 3) 0.7 ≤ AUC < 0.8: fair classifier. 4) 0.8 ≤ AUC < 0.9: good classifier. 5) AUC ≥ 0.9: excellent classifier. A high AUC score implies that the model is doing a good job of separating the positive and negative classes, regardless of the threshold chosen. It’s a valuable metric when the goal is to evaluate the overall discriminatory power of a binary classifier without being concerned about a specific threshold:

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]

Accuracy is one of the most straightforward and commonly used metrics in deep learning and machine learning for evaluating the performance of classification models. It measures the ratio of correctly predicted instances to the total number of instances in the dataset. To calculate accuracy, compare the model’s predictions to the actual labels for a given dataset. Every correctly classified instance contributes to the numerator, while every instance, whether classified correctly or incorrectly, contributes to the denominator. Accuracy is usually expressed as a percentage, ranging from 0% to 100%. A higher accuracy value indicates that the model is making more correct predictions, while a lower accuracy value suggests that the model is making more errors. However, accuracy may not always provide a complete picture of a model’s performance, especially in scenarios with imbalanced datasets. For example, in a dataset where one class is much more prevalent than the other, a classifier that predicts the majority class for every instance can still achieve a high accuracy, but it may be practically useless.

The F1-score is a commonly used metric in deep learning and machine learning, particularly for binary and multi-class classification tasks. It combines precision and recall into a single score to provide a balanced measure of a model’s performance. The F1-score is especially useful when dealing with imbalanced datasets or when trying to strike a balance between precision and recall. The F1-score gives equal weight to precision and recall. This makes it useful for balancing the trade-off between minimizing false positives and capturing as many true positives as possible. The F1-score ranges between 0 and 1, with a higher value indicating better model performance. An F1-score of 1 indicates perfect precision and recall, while an F1-score of 0 means that either precision or recall is zero.

Recall, also known as sensitivity or true positive rate, is an important metric in deep learning and machine learning, particularly in binary and multi-class classification tasks. It measures the ability of a model to correctly identify all relevant instances of a particular class within a dataset. Recall quantifies the model’s capacity to minimize false negatives, which are instances of the positive class that are incorrectly classified as negative. Recall is usually expressed as a value between 0 and 1. A value of 1 indicates that the model correctly identifies all positive instances in the dataset, while a value of 0 suggests that it misses all positive instances. High recall is desirable when the cost of false negatives is high, and it is crucial to capture as many positive instances as possible. Recall and precision are often in tension with each other. Increasing recall tends to decrease precision and vice versa. Precision focuses on minimizing false positives, whereas recall concentrates on minimizing false negatives. The choice between the two metrics depends on the specific problem and its requirements.

Specificity, also known as the true negative rate, is a metric used in deep learning and machine learning, particularly in binary classification tasks, to evaluate a model’s ability to correctly identify
negative instances within a dataset. Specificity complements sensitivity and helps in assessing a model’s performance, especially when the cost of false positives is a concern. Specificity typically ranges between 0 and 1, with a higher value indicating better performance in correctly identifying negative instances. High specificity is desirable when the cost of false positives is high. In applications such as medical testing, a high specificity ensures that healthy individuals are not incorrectly identified as having a disease. There is often a trade-off between specificity and sensitivity (recall). Increasing specificity may decrease sensitivity and vice versa. The choice between the two metrics depends on the specific problem and its requirements (Feng & Chen, 2022).

Implementation

This experiment utilized a system comprising a 12900k CPU and 8 NVIDIA GeForce GTX3080 GPUs. The batch size was set to 64 for ResNet-152, BiGRU, and the visual-language matching scoring model. However, for the VLMS model, a batch size of 24 was used, with a learning rate of 0.0001 for the generators and 0.0004 for the discriminators. Notably, training for the MSCOCO dataset was halted after 120 epochs, for the LAION-400M dataset after 100 epochs, for the Conceptual Captions dataset after 160 epochs, and for the Visual Genome dataset after 180 epochs.

Performance Comparison

In Table 1, we provide a comprehensive comparison of our proposed method with the latest methods on several benchmark datasets across multiple visual-language matching tasks, including MSCOCO, LAION-400M, Conceptual Captions, and Visual Genome. Evaluation metrics include AUC and accuracy, which offer insights into the model’s performance in ranking and classification. Our method (“Ours”) achieved significant scores across all datasets. On the MSCOCO dataset, we achieved an AUC of 0.95 and an accuracy of 0.90, demonstrating the effectiveness of our method in capturing complex visual-textual relationships. Similarly, on the LAION-400M dataset, our method obtained an AUC of 0.90 and an accuracy of 0.89, showcasing its robustness in large-scale multimodal understanding. For Conceptual Captions, our method achieved an AUC of 0.88 and an accuracy of 0.89, indicating its capability to handle diverse textual descriptions related to images. On the Visual Genome dataset, we surpassed existing methods with an AUC of 0.93 and an accuracy of 0.91, highlighting our model’s ability to capture detailed relationships within images. It’s worth noting that our method consistently outperformed or matched the performance of state-of-the-art methods such as CLIP, SCAN, DAMSM, and UNITER on all datasets and evaluation metrics. These results demonstrate the effectiveness and versatility of our proposed method in various visual-language matching scenarios. Figure 3 presents the contents of the table, and from the graph, the outstanding performance of our method is evident.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSCOCO AUC</th>
<th>MSCOCO Accuracy</th>
<th>LAION-400M AUC</th>
<th>LAION-400M Accuracy</th>
<th>Conceptual Captions AUC</th>
<th>Conceptual Captions Accuracy</th>
<th>Visual Genome AUC</th>
<th>Visual Genome Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLIP (Radford et al., 2021)</td>
<td>0.95</td>
<td>0.89</td>
<td>0.89</td>
<td>0.87</td>
<td>0.86</td>
<td>0.88</td>
<td>0.91</td>
<td>0.90</td>
</tr>
<tr>
<td>SCAN (Russin, et al., 2019)</td>
<td>0.92</td>
<td>0.88</td>
<td>0.88</td>
<td>0.89</td>
<td>0.88</td>
<td>0.89</td>
<td>0.87</td>
<td>0.88</td>
</tr>
<tr>
<td>DAMSM (Alshehri et al., 2022)</td>
<td>0.91</td>
<td>0.87</td>
<td>0.87</td>
<td>0.91</td>
<td>0.86</td>
<td>0.87</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>UNITER (Zhang &amp; Wang, 2022)</td>
<td>0.93</td>
<td>0.89</td>
<td>0.82</td>
<td>0.86</td>
<td>0.81</td>
<td>0.89</td>
<td>0.87</td>
<td>0.85</td>
</tr>
<tr>
<td>Ours</td>
<td>0.95</td>
<td>0.90</td>
<td>0.90</td>
<td>0.89</td>
<td>0.88</td>
<td>0.89</td>
<td>0.93</td>
<td>0.91</td>
</tr>
</tbody>
</table>
Table 2 and Figure 4 present the comparative results of our proposed method with state-of-the-art approaches in downstream vision-language tasks. These tasks encompass visual question answering (VQA) (Gao et al., 2022), natural language visual reasoning (NLVR) (Huang et al., 2021), and text-image entailment (SNLI-VE) (You et al., 2023). Different rows in the table represent various methods, while different columns denote different datasets and test sets. In the table, we can observe the performance of various methods across different tasks and test sets. Methods like CLIP, SCAN, DAMSM, and UNITER exhibit varying degrees of performance across these tasks. However, it is worth noting that our proposed method (“Ours”) consistently demonstrates outstanding performance on most tasks and test sets. Specifically, our method achieves high scores in various tests for VQA, NLVR, and SNLI-VE. Notably, in the SNLI-VE tests, our method excels with an accuracy of 78.91, surpassing other methods.

The results in this table underscore the exceptional performance of our proposed method in vision-language fusion tasks, highlighting its wide applicability across multiple domains. This holds significant implications for advancing research and applications in the field of vision and language integration.

As shown in Table 3, we report the results of zero-shot image-text retrieval. It is worth noting that our model achieved state-of-the-art performance, significantly surpassing other pretrained models. In this task, our approach demonstrates exceptional zero-shot retrieval capabilities, indicating that our model can effectively retrieve relevant images or text without specific training data. Specifically, in the text retrieval task, our model stands out with an impressive @1 score of 88.7, far exceeding...
other methods, including DAMSM, UNITER, and SCAN. In the image retrieval task, our model also performs remarkably well, achieving @1, @5, and @10 scores of 66.5, 88.1, and 88.3, respectively. These results not only highlight the outstanding performance of our proposed method in zero-shot retrieval tasks but also showcase its tremendous potential in efficiently retrieving relevant information. Our model can play a pivotal role in various cross-modal information retrieval applications, providing a powerful tool for addressing real-world challenges.

### Ablation Study

As shown in Table 4 and Table 5, we conducted extensive ablation experiments to assess their impact on the studied task. The objective of these experiments was to dissect the performance characteristics...
of different combinations. The combination of VGG16+LSTM achieved an accuracy, F1-score, and recall of 0.85, indicating a fairly balanced performance, while the AUC of 0.89 suggests good discriminative capability. Combining VGG16 with BiGRU improved accuracy to 0.87, although the F1-score and recall were slightly lower at 0.82 and 0.84, respectively, while maintaining an AUC of 0.90. Combining ResNet152 with LSTM increased accuracy to 0.88, with an F1-score slightly below 0.81 and a recall of 0.83. This combination achieved an AUC of 0.91. The configuration of GoogLeNet+BiGRU performed exceptionally well across all metrics, with an accuracy of 0.90, an F1-score of 0.87, a recall of 0.88, and an AUC of 0.94. Our model combining ResNet152+BiGRU also demonstrated robust performance, with an accuracy of 0.90, an F1-score of 0.88, a recall of 0.88, and the highest AUC of 0.95.

As shown in Table 5, in this ablation experiment, we investigated the impact of different combinations of loss functions on model performance, emphasizing the rationality and excellence of our triple loss function. In the next stage of this ablation experiment, we delved further into the superiority of the triple loss function and compared it with differences among other common combinations of loss functions. To comprehensively understand the performance of different loss functions, we designed a series of experiments using both single loss functions and combinations of dual loss functions. The goal of these experiments was to evaluate the impact of each loss function combination on model performance and determine whether they could achieve optimal performance in their respective tasks.

**Case Study**

Figure 5 demonstrates the exceptional performance of our model in comprehending complex scenes and contextual inquiries. It also showcases a keen understanding of information across multiple levels and dimensions. Samples 1-4 represent pre-training examples with scene text descriptions, samples
5-8 feature long pieces of scene text, and samples 9-12 highlight the diverse knowledge present in the pre-training data, including logos, celebrities, landmarks, products, and more. This series of qualitative examples not only highlights the immense potential of VLMS but also further substantiates its distinctive advantage in addressing complex real-world challenges.

CONCLUSION

In this study, we delved into the field of multi-modal machine learning, focusing on a challenging problem: how to effectively capture and leverage the information matching relationship between vision and language. To address this, we introduced the VLMS model, a model that not only possesses forward-looking theoretical foundations but also demonstrates outstanding performance across various real-world tasks and datasets. The design of the VLMS model encompasses several key aspects, one of which is its multi-level matching capability. The model can perform local-level matching, capturing subtle associations within the data, and also perform global-level matching, understanding the overall semantic structure of the data. Additionally, it can engage in more general matching, enabling the
model to generalize to different domains and tasks. This multi-level matching capability empowers the VLMS model to excel in the realm of multi-modal information matching.

Nevertheless, despite the remarkable performance demonstrated by the VLMS model in various aspects, it is essential to acknowledge its limitations. First, the model’s performance may be constrained when dealing with complex multi-modal data, such as data containing noise or ambiguity (Yuan et al., 2022). In such cases, the model’s performance may not meet expectations, highlighting the need for further research to enhance its robustness in adapting to more complex real-world scenarios. Additionally, the training and inference processes of the VLMS model require substantial computational resources and data, which may not be practical in resource-constrained application scenarios (Feng et al., 2021). Therefore, one of our future research directions is to explore more efficient model training and deployment strategies to enhance the model’s usability and scalability.

Looking ahead, the field of multi-modal machine learning continues to offer vast research opportunities. We plan to further refine and extend the VLMS model to address emerging challenges in the analysis of multi-modal data. Additionally, we will focus on the practical applications of the model, including fields such as automation and manufacturing, healthcare, and autonomous driving, with the aim of translating the achievements of multi-modal machine learning into innovative solutions. The significance of this study lies in its contribution to the field of multi-modal machine learning, and we anticipate that future work will continue to drive progress in this field, providing more innovations and possibilities for addressing complex real-world problems. The development of multi-modal machine learning not only enriches our understanding of data but also fosters breakthroughs and advancements in the field of artificial intelligence.

**AUTHOR NOTE**

This project is supported by Social Science Foundation of Ministry of Education of China (Grant No.22YJC740075) and by Culture Rejuvenation Project in Henan Province (Grant No.2023XWH037).
REFERENCES


Gao, F., Ping, Q., Thattai, G. (2022). A thousand words are worth more than a picture: Natural language-centric outside-knowledge visual question answering. https://doi.org/arXiv.2201.05299.10.4855


