第一页为封面页

参赛学生姓名： 樊宬睿

中学： 上海中学国际部

省份： 上海市

国家/地区： 中国

指导老师姓名： 何天舟

指导老师单位： 上海中学国际部

论文题目： **CelsiaNet: Collaborative Understanding of Images and Text–A Multi-Modal Vision-Language Model Framework**

CelsiaNet: Collaborative Understanding of Images and Text–A Multi-Modal Vision-Language Model Framework

Chengjui Fan
Shanghai High School International Division

Abstract

In computer vision, combining visual and language models has become the future of image-text processing and captioning. Recently, due to the high demand for Artificial Intelligence in modern life and work, accuracy and precision are needed for every model that is brought to work. Furthermore, combining multiple lightweight models and algorithms for partial computing can yield a highly efficient yet accurate Multimodal framework. Our approach combines the strengths of each model to address complex problems that require further knowledge and relational understanding of both Vision and Language.

In this paper, we propose a novel framework according to this description: CelsiaNet, a Multimodal Vision-Language Model Framework that integrates partial techniques from LLaVA, BLIP-2, BERT, and Q-Former and employs a multi-stage training strategy to bolster the model's generalization capabilities. Initially, we utilize a pre-trained vision transformer, convolution network, and a large-scale language model for efficient bootstrapped multimodal representation learning. Subsequently, the entire model undergoes end-to-end fine-tuning on a vast multimodal dataset to achieve a rich integration of visual and linguistic modalities. Moreover, we adopt a Context-Object Split (COS) factorization that partitions the latent space into contextual and object-specific components, diversifying generated descriptions and the model's ability to handle novel objects not encountered during training. Finally, we introduce a context-based pseudo-supervision module that diversifies the contextual descriptions of similar images in the latent space.

This architecture performs well in generating diverse and accurate image captions handling various queries, and our model achieves zero-shot captioning through unsupervised training(with pseudo-supervision) on the Visual Genome (VG), VG-COCO, and RefCOCO datasets. Our experimental results show that our framework significantly outperforms existing models in various vision-language tasks, highlighting Multimodal VLM's potential, which should be explored further.

Key words: Multimodal Learning, Vision-Language Model, Context-Object Split, Bootstrapped Learning, End-to-End Fine-Tuning

Contents

1. Introduction

1.1. Purpose of the Multimodal Vision Language Models

Multi-modal vision language models(VLMs) are developed for the need of an integrated system of programs(neural networks) that could understand and react to the relationship between Visual and Lingual data when state-of-the-art models in both fields could cooperate in synergy. This approach reduces the need for extensive, specially crafted datasets that directly relate text to images and vice versa, instead utilizing existing text-only and image-only datasets along with techniques to align these modalities [1].

One of the primary purposes of VLMs is to accurately generate and translate visual data to language outputs. These systems are needed in fields such as image captioning services, visual question-answering services, automobile autopilot systems, and Artificial General Intelligence(AGI) development. For instance, when these models are applied in image captioning, the system would need to generate a descriptive text that could explain the relationships and actions in the image, like how and what humans can observe [2]. Another purpose for designing these systems is to efficiently leverage the vast amount of training data on the internet into developing a multi-purposed model composed of respective related networks. By training on a large amount of data, VLMs can learn to generalize across different contexts and relationships, which could improve their zero-shot capabilities. This means that the models working in coherence inside the main system could predict the results of unfamiliar situations without needing an excessive amount of situation-specific training [1]. Furthermore, VLMs help develop techniques to integrate multiple models into a single framework. This involves sophisticated mechanisms such as cross-attention layers, contrastive learning objectives, and embedding alignment, ensuring that the model can coherently fuse visual and textual features and understand them. All of the improvements stated above have allowed VLMs to be versatile and accurate in various applications while being (cost and data) efficient [3].

1.2. Previous Weakness in VLMs

Although Vision-Language Models(VLMs) have received many research developments and improvements, they still encounter difficulties understanding detailed regional visual components. These challenges have occurred due to many older technical problems unsolved but were still implemented into the task:

During previous research development on VLMs, VLMs have not been good at capturing fine-grained spatial relationships within the images [4]. For example, when a VLM is called on to analyze the interactions of the objects inside an image, the VLM can only interpret the physical existence of the objects inside the image due to the lack of spatial and action understanding. This is very problematic in tasks that would require the model to clearly identify the layout and situation, such as in the case of indoor navigation for autonomous systems [5], or the interpretation of the arrangements of compositions in an antique artwork.

At the same time, the quality of training data has also often been a constraining factor in VLM development. Since the quality of a dataset depends on the images' content and resolution, the tags' ambiguity(the less, the better), and the diversity in both images and tags, it is hard to create an initial dataset. Many existing datasets provide only general descriptions without keywords that could pinpoint the specific characteristics of the regions described within the images. This results in a model that would generate non-precise answers when given a task [6]. Furthermore, human bias is another crucial factor to be solved as different tag rankings would cause the model's interpretation to skew while also affecting future models that train on an extended dataset crafted by former captioning models [7].

Specific regional analysis has been another struggle point for VLMs, as older VLMs find it hard to link or anchor particular objects and elements in a region of an image to the part of the text that exactly describes them, which results in the disconnection between the visual and textual modalities. This is due to the failure to make connectors that could translate and map features between text and images. More disconnections can lead to incorrect or incomplete interpretations, especially in tasks where the model must explain a verbally given region inside the image or other kinds of analysis regarding good relatability between vision and text [8]. Another significant weakness in previous VLMs is that they cannot understand the holistic meaning of an entire picture after identifying the components in the image. Even if the model was capable of seeing what is visually presented upon the image, the inability to reason what the purpose or the background situation can result in disjointed captions or failure to make a valuable analysis of the function of the visual content [9].

All of the mentioned problems during VLM development have caused VLMs to sometimes tend to generate overgeneralized captions that lack the specificity needed to describe an image fully. These are parts of the difficulties we also deal with to overcome in this paper, where a multi-modal VLM will require higher accuracy, precision, and collaboration between models.

1.3. Proposed Strategy

In this study, we aim to present our Vision Language Model(VLM) architecture framework that we designed for the model's better recognition of the relationships between text and images and, hence, generating fine image captions. Our strategy used a multi-stage method, resulting in strength in efficient pretraining integrating BLIP-2's method and end-to-end fine-tuning inspired by LLaVA's approach [10] [11]. Most importantly, we have added a Context-Object Split (COS) latent space factorization [12], which segments the representation space to boost caption diversity and accurately manage object detection in the image. Furthermore, we also added a context-based pseudo-supervision that ensures the accurate matching of features and data between multiple models and datasets. For the pretraining process, we applied K-means clustering for unsupervised feature grouping. During the alignment phase, we incorporate BERT's ability for textual embeddings [13] and pair it with the excellent feature extraction capabilities of the pre-trained vision transformers and CNNs to assist the alignment.

Architecturally, our model is based on a visual transformer for image encoding and uses a multimodal mixture of encoder-decoder (MED) framework, seamlessly integrating BLIP-2's alignment mechanism through a lightweight Querying Transformer (Q-Former) to connect image encoding with language modeling. The pretraining process applies the three losses: Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), and Language Modeling Loss (LM)—to amplify the model's interpretative and generative ability [14]. We also implement the Captioning and Filtering (CapFilt) method, enhancing the textual dataset's quality through a dual mechanism of web image captioning and noise elimination [14]. In the final phase, the model employs a caption generation mechanism that utilizes the Q-Former as the primary encoder [15]. This is complemented by a refining encoder that extracts and processes complex feature relationships. The system culminates in a decoder that integrates visual and textual information, producing contextually appropriate and visually grounded captions.

The contributions of this study are listed as follows:

- **Context-Object Split Architecture:** Enhancing our VLM architecture with Context-Object Split (COS) factorization to adeptly capture diverse contexts and improve performance on novel objects, inspired by the COS-CVAE approach [12].

- **Lightweight Model and High Performance:** By Implementing multiple state-of-the-art modules together and stating our own highly efficient and accurate linking mechanisms, our model demonstrates state-of-the-art performance at 4.2B parameters compared to many previous works, such as Osprey, which operates on 7B parameters [16].

- **Unsupervised Image Training with Pseudo-Supervision Training:** Throughout our model, we separated the images and the text to be trained separately, whereas the images are trained unsupervised with clustering algorithms. Later, the bootstrapped features are trained using a self-supervision mechanism(or pseudo-supervision) that relates captions from a captioned image to the ones that had similar underlying caption relations, inspired by the paper, Unicom [17].

- **Multi-stage Training with Attention Mechanisms:** Implementing a multi-stage training process that incorporates self-attention mechanisms and a projection layer from the frozen image encoder at the end of the processes. This approach allows the model to better relate the final output with the original visual content, improving the coherence between generated text and input images.

- **Better Zero-Shot Capability with Multi-Modal Structure:** Our model is trained under an unsupervised manner with a pseudo-supervision module, which allows the captions of different content to be related or be used to learn the underlying relationships in each image or content. The basis of our framework is on strong performing models and methods, which has enabled the model to perform exceptionally well when our introduced modules can connect them with perfect consistency and attention mechanisms.

2. Related Work

2.1. Vision Language Models

2.1.1 Development of Current State Visual Language Models (VLMs)

Vision Language Models(VLMs) have become a significant focus of the development of Artificial intelligence; these models integrate visual and lingual processing capabilities to process complex understanding tasks. These tasks include generating captions for images and responding to visual and lingual questions [9]. This kind of model could be created by integrating vision models alongside Large Language Models(LLM) as a multi-modal or all-in-one approach by making a model relatable between vision and text contents.

2.1.2 Classification of VLMs

If we classify based on their functionalities, there are three main types of VLMs: models dedicated to vision-language relational understanding, models that process multiple types of inputs (images and texts) to generate unimodal outputs and models that accept and produce multimodal inputs and outputs of vision-language data. VLMs are mainly designed in these categories, and there is a wide range of subtasks that each distinctive VLM model chooses one or many to achieve. Their common aim is to create a model that evolves from only being able to deal with specialized understanding tasks to be more versatile in multi-input-output tasks. [2].

2.1.3 Architectural Choices and Training Techniques

Recent developments in visual-based Multimodal Large Language Models (MLLMs) have taken multiple architectural, alignment, and training approaches [18]. The goal is to fit the language model towards the language processing section of the task while aligning the vision model to the corresponding visual task, using a newly trained feature mapping mechanism that can relate the image and text features of the two or multiple models.

The newest architectures for vision-based MLLMs have been developed to make a more efficient cross-modal attention mechanism for better fusing different models. For example, the LAVIS framework [19], which we will use for training our network, is a unified structure that can be applied to many different vision-language tasks. The architecture can integrate many different advantages of the vision model. For example, it can apply vision networks like BLIP [14]that use novel pretraining strategies to improve image-text alignment; the Flamingo model [20] introduces an interleaved architecture for more efficient multimodal reasoning; and CLIP [21] which employs contrastive learning on vast amounts of image-text pairs from the internet, enabling zero-shot classification capabilities(which BLIP refines the network by using captioning and filtering to create high-quality training data).

The developments in architectures and training methods of vision-based MLLMs have improved multiple capabilities and the model's performance on various tasks. However, cross-modal attention and understanding abilities remain major points that need to be solved, as understanding linked models would cause accuracy deficits or potential biases in large-scale pretraining datasets.

2.1.4 Performance Analysis

Knowing that Vision-Language Models (VLMs) have the capabilities to be used in a diverse set of tasks, standardized performance analysis tests should be made in order to reduce the incomplete reports and problems on lack of transparency in order to demonstrate the effectiveness of new VLMs that are being developed [22].

To address these limitations, such analysis frameworks, like the Holistic Evaluation of Vision-Language Models (VHELM) v1.0 [22] and scalable solutions that rely on already annotated benchmarks are implemented to test the overall ability of the VLM. Task-specific tests also have evaluation benchmarks, such as the METEOR and CIDEr benchmarks, regarding the captioning ability of VLMs, which also demonstrate the model's ability to handle specific needs and datasets. These frameworks aim to increase transparency and provide a more fair and comprehensive understanding of VLMs [23].

Based on the given frameworks, VLMs have been analyzed to have deficits in their abilities to handle training and validation on imbalanced datasets, where the given dataset is skewed or poor in maintaining a normal distribution, causing minor classes to be ignored due to model bias [24] [25]. Encountering this performance error was likely caused by the dataset preparation, preprocessing, and alignment steps.

As larger models with more parameters and layers have been believed to learn more relations between contents, and the learning abilities drop when too many layers are added [?], larger VLMs are hard to fine-tune and correct minor errors once it was trained. No current method could escape the long fine-tuning process; however, researchers have designed benchmark probing tasks that would classify tasks with specific properties as each evaluation field with annotations provided as ground truth. However, these methods are not scalable and may not cover all limitations of a certain VLM [23].

Regarding VLMs' performance, many tests and benchmarks can evaluate certain aspects of the model. However, analysis and fine-tuning the model are expensive, resulting in improvements in model architecture more often than retraining on a slightly optimized model. As we used in our experiments, we will use the standard benchmarking methods such as BLEU-N, ROUGE-L, CIDEr, METEOR, and SPICE to evaluate our model. with previous works.

2.1.5 Applications

Vision-Language Models (VLMs) can be applied across various visual recognition tasks, such as Visual Grounding.

In this task, models need to understand texts and their regions of interest inside the image, aiding in object detection, scene understanding, and human-computer interaction. This also relates to image captioning, where obtaining a decent answer to a situation requires knowledge of all the relationships of the features inside each region group of a single image.

VLMs have been adapted for specific domains to meet production requirements or daily life usage. In the automotive industry, for example, VLMs assist in autonomous driving by interpreting visual data with textual inputs like traffic signs or navigation commands, and this would be the implementation of multiple models that would be working on image captioning and system task understanding. In security, they help surveillance by analyzing video feeds and generating reports. In the photography industry, they could assist in making suggestions for modifications to the color of an image or to include other generative components.

Another major field that is separated from individual research VLMs is medical applications, particularly in medical report generation and visual question answering, where super-leveled VLMs with perfect accuracy are needed that would require government project investment or company research. These models can interpret medical images, generate detailed diagnostic reports, and answer questions related to patient data and medical imagery, thus aiding healthcare professionals in diagnosis and treatment planning.answering [2].

2.2. Captioning

For captioning images, early methods to caption image-to-text tags were based on retrieval and template methods. They were very limited in their ability to generate diverse and contextual accurate captions [26]. This was primarily due to the incoherence between the image and text models, as each model can finish their task in feature extracting but faces problems when collaborating on a single query.

Later on, the introduction of deep learning has revolutionized the field of image captioning. Deep learning-based techniques typically employ an encoder-decoder framework, where the encoder extracts features from the image, and the decoder generates the corresponding caption. Attention mechanisms have also supported the model by focusing on the region of interest (RoI) on the image; these mechanisms allow the model to generate more contextually accurate captions [27]. Training strategies, including reinforcement learning and adversarial training, have been explored to improve captioning performance.

Large-scale models with more layers, parameters, or training contexts are more effective for the learning ability of visual neural networks. This could be implemented during the pre-training and fine-tuning phases, but they increase the cost of training the models and the challenge to make a perfect dataset or its processing program [28].

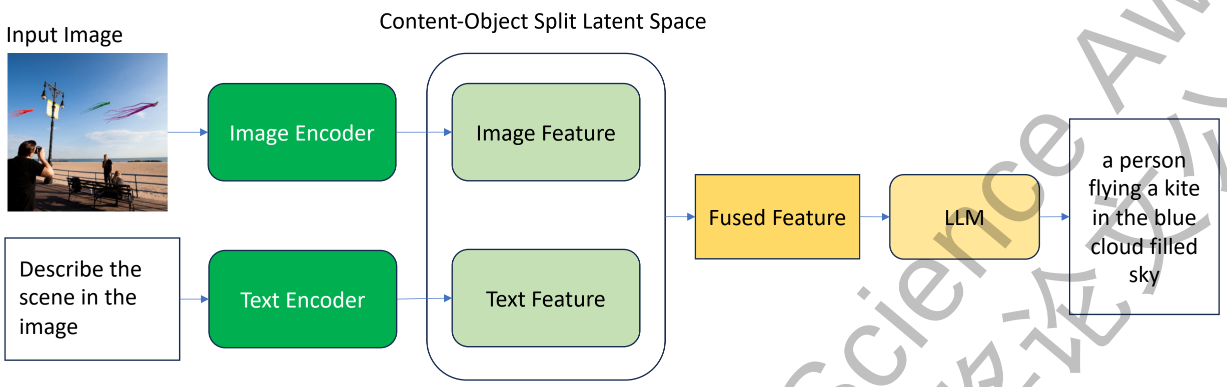
Input Image
Content-Object Split Latent Space
Describe the scene in the image
Image Encoder
Text Encoder
Image Feature
Text Feature
Fused Feature
LLM
a person flying a kite in the blue cloud filled sky

Figure 1. Overview of the proposed method. Encoded Image and Text passes through a content-object split latent space into one feature and LLM into the final context.

3. Method

3.1. Overview

Our proposed model framework has integrated multiple new cutting-edge methodologies presented recently by other researchers while reducing their inherent problems.

In the preprocessing stage, we have standardized all the images into the same dimension while removing all the tags on the images to train the unsupervised model. Next, in the pre-training stage, we encoded the image and the text separately. The images were encoded with a pre-trained ResNet. The encoded feature vectors will pass through a K-Means clustering through the Instance and Cluster Discrimination processes implemented from the paper Unicom [17]. On the other hand, the text will be preprocessed and encoded using the BERT model [13].

We have employed a multi-stage training process inspired by LLaVA and BLIP-2. The model brings in the first step of bootstrapping multimodal representation learning from BLIP-2 [10]. Here, the program projects the outputs of the pre-trained Convolutional Neural Network(ResNet as the image encoder [29]) with the pre-trained language encoder's output. This ensures that the pre-training method is relatively computationally efficient compared to training a multipurpose unimodal model, and it enhances the understanding of the model's image and textual content.

After that, the features are forwarded into the context-object split (COS) factorization that we integrated from the COS-CVAE paper [12]. This module partitions the latent space into the contextual and object-specific components for the model to capture a broad view of the entire picture depicted in the inputted feature and the specific components of all the details in the feature array. This can ensure the accuracy and the relational awareness of the generated captions. This also greatly enables the model to predict and handle objects not encountered during the training procedure. Additionally, the model uses a pseudo-supervision mechanism to retrieve partial captions from other similar images when one image lacks a caption while diversifying the captions that are possible to be accurately generated from each picture.

In the second step, we used the finetuning method introduced from LLaVA [11]. A projection layer is added from the original frozen image encoder(which will not be finetuned). Then, the projection layer and the language model are fused for finetuning. This helps the language model learn the relationships between the features of the projection layer and the text without making a combined network that can interpret text and images at the start.

We are integrating the abovementioned mechanisms in a robust and versatile model. This architecture will have the advantages of efficient pre-training methodologies, end-to-end multimodal integration, multi-image processing capabilities, and context-object split latent space factorization.

3.2. Vision-Language Model Pre-training

To begin with, the pre-training process for vision-language models consists of standardizing all input images to uniform dimensions to ensure consistent input for the model. At the same time, all the labels for the images are removed to enable the model to undergo unsupervised training. After the preprocessing, the images and text are encoded separately. We will first discuss the pre-training method for the image and then for the text.

3.2.1 Vision Model Pre-training

For image encoding, the pre-trained vision convolution networks(ResNet [29]) are dedicated to extracting the high-dimensional feature vectors, capturing visual characteristics and spatial relationships in the whole picture level of the images. Furthermore, this prevents requiring training from scratch(which is very costly and computationally expensive). These feature vectors are combined into a single dataset, which will encounter a clustering process using the K-means algorithm. This unsupervised learning approach enables the model to observe relationships between image features with underlying patterns and associations within the visual information.

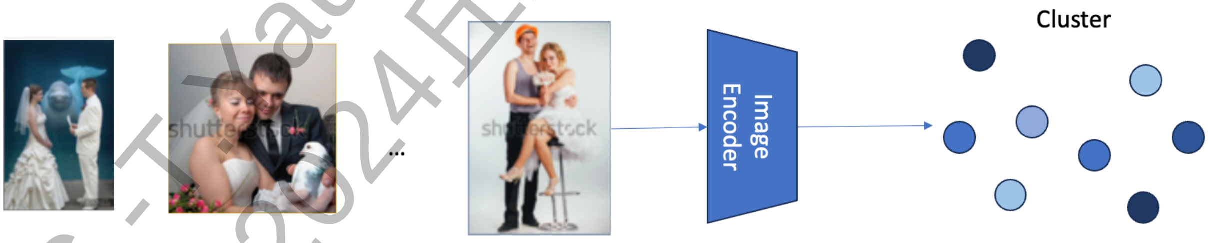

Figure 2. Encoding and clustering process for input images

Here, we have applied and adapted the Instance and Cluster Discrimination in feature representation learning from Unicom's paper [17].

A set of n images used for training is represented as $X = \{x_1, x_2, \ldots, x_n\}$. The primary objective of feature representation learning is to train a mapping function f that can project images X onto a set of embeddings $E = \{e_1, e_2, \ldots, e_n\}$, where each $e_i = f(x_i)$. This mapping is able to relate similar features among the set of images.

Instant Discrimination uses the contrastive loss function in order to complete the mapping process, where the formula is:

$$\mathcal{L}_{\text{instance}} = -\sum_{i=1}^{n} \log \frac{\exp(e_i'^T e_i)}{\sum_{j=0}^{m} \exp(e_j'^T e_i)}, \tag{1}$$

where e_i and e_i' represent the positive embeddings of instance i, and e_j' includes one positive embedding for i and m negative embeddings from different instances [17].

In contrast, Cluster Discrimination involves two key steps: clustering and discrimination. During the clustering stage, pseudo-class labels are assigned to each instance to facilitate a pseudo-supervised training process. K-Means Clustering is performed automatically on features $e_i = f(x_i)$ to define k clusters, with each centroid w_i representing the archetype of the i-th cluster. The training dataset $\{x_i\}_{i=1}^{n}$ is thus divided into k classes, each represented by the prototypes $W = \{w_i\}_{i=1}^{k}$.

The discrimination stage refines a conventional softmax classification loss:

$$\mathcal{L}_{\text{cluster}} = -\sum_{i=1}^{n} \log \frac{\exp(w_i^T e_i)}{\sum_{j=1}^{k} \exp(w_j^T e_i)}, \tag{2}$$

where e_i is the embedding of the image x_i, and x_i is an example of the class represented by w_i.

The k-means clustering algorithm is applied here, which would divide the set of vectors or the features we have obtained into k distinct groups based on proximity. To enhance the accuracy of the representation, we combine image and text features as provided by the pre-trained ResNet model, leveraging their combined attributes. The clustering step simultaneously learns a $d \times k$ centroid matrix W and the cluster assignments y_i for each image x_i by solving the following optimization problem:

$$\min_{W \in \mathbb{R}^{d \times k}} \frac{1}{n} \sum_{i=1}^{n} \min_{y_i \in \{0,1\}^k} \|\Phi(f(x_i), f'(x_i')) - W y_i\|_2^2 \quad \text{s.t.} \quad y_i^\top \mathbf{1}_k = 1, \tag{3}$$

where $f(x_i)$ and $f'(x_i')$ are the image and text feature embeddings, respectively, Φ is a feature ensemble function, $W \in \mathbb{R}^{d \times k}$ is the centroid matrix, $y_i \in \{0,1\}^k$ is a binary label assignment vector, and $\mathbf{1}_k$ is a vector of ones of dimension k. A simple averaging mechanism is used as the ensemble function, utilizing the aligned visual-textual representation provided by the pre-trained vision model [17].

3.2.2 Language Model Pretraining

The pretraining method for the language model is based on the BERT (Bidirectional Encoder Representations from Transformers) methodology, which creates contextual embeddings for the text in the datasets [13]. The text or labels associated with each image are tokenized using BERT's WordPiece tokenizer, forming out of vocabulary words as sub-word units. The tokenized text is then processed through a pre-trained BERT model.

BERT's self-attention mechanism allows it to consider the entire context related to each word and output a high dimensional vector(usually 768 or 1024 dimensions). To align with the image features, the representation of the special `[CLS]` token or the aggregate embedding derived from the mean pooling of all token vectors as the final text representation.

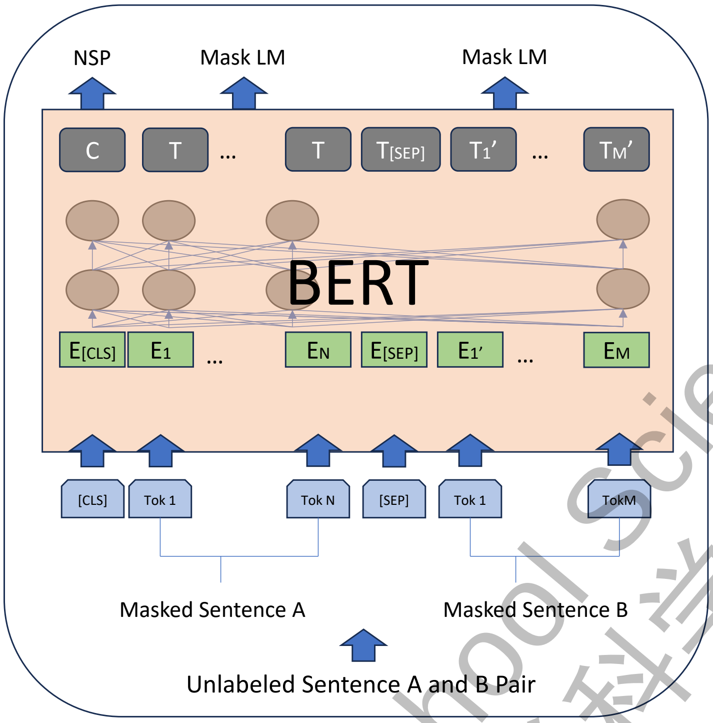

Figure 3. Pre-processing of textual data using the Bidirectional Encoder Representations from Transformers(BERT) method [13]

This encoding procedure allows the textual features to be understood with linguistic constructs and relationships. In contradiction to the image preprocessing phase, no clustering is executed on the linguistic features.

As delineated in Figure 3, we designate the input embedding as E, the terminal hidden vector of the special `[CLS]` token as $C \in \mathbb{R}^H$, and the terminal hidden vector for the i^{th} input token as $T_i \in \mathbb{R}^H$. The input question and passage are encapsulated as a singular packed sequence, with the question harnessing the `A` embedding and the passage leveraging the `B` embedding. Solely during the fine-tuning phase are a start vector $S \in \mathbb{R}^H$ and an end vector $E \in \mathbb{R}^H$ introduced. The likelihood of the word i marking the inception of the answer span is calculated as a scalar product between T_i and S, succeeded by a softmax operation across all the words in the paragraph:

$$P_i = \frac{e^{S \cdot T_i}}{\sum_j e^{S \cdot T_j}} \tag{4}$$

A parallel computational mechanism is implemented for the termination of the answer span. The quantitative assessment of a candidate span extending from position i to position j is formulated as $S \cdot T_i + E \cdot T_j$, where the span exhibiting the maximal score under the constraint $j \geq i$ is employed for prediction. The optimization criterion is defined as the summation of log probabilities corresponding to the true initiation and termination positions. The model undergoes parameter optimization for 3 epochs utilizing a learning rate of 5×10^{-5} and a batch size of 32 [13].

3.3. Vision and Text Feature Alignment

To effectively align visual and textual information, we used both visual transformers and convolution networks as image encoders. Firstly, the static encoder(CNN) will encode the image into features. Then, visual transformers partition the features into patches and encode them into a sequence of embeddings. After that, the Q-Former aligns the image and text features to input them into the LLM. The following section explores the implementation of the alignment.

3.3.1 Visual Transformer Assisting Image Encoder

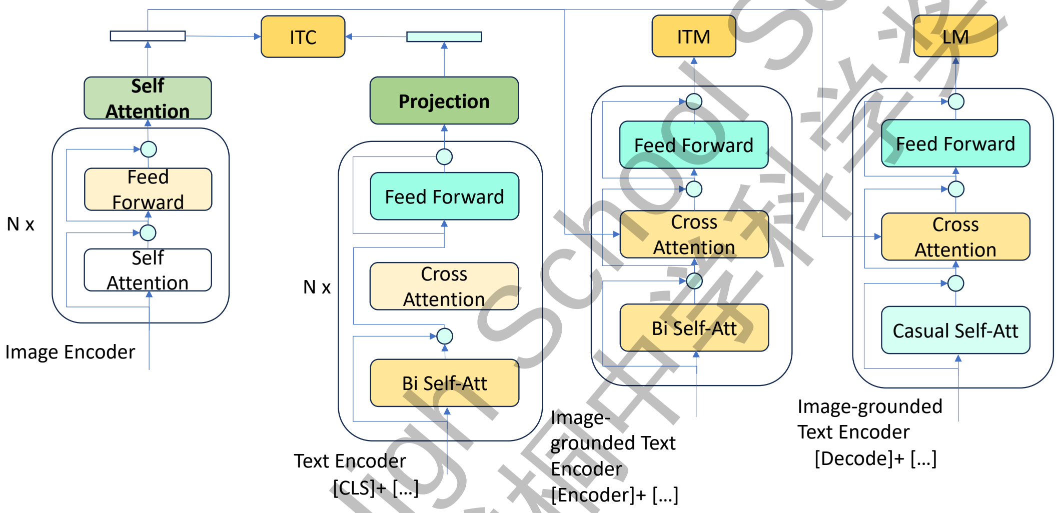

Figure 4. Image and Text encoded passes through a content-object split latent space into one feature and LLM into the final context.

On the bottom level, we use a pre-trained ResNet model as a static image encoder to extract high-dimensional feature vectors from images. We fine-tune and optimize the network during training to integrate multiple computing modules in our VLM model. Extracted high-level features are then passed to the visual transformers.

Visual transformers can efficiently help the encoding process before feeding the encoded features into the LLM. This approach, first introduced by BLIP v1, significantly reduces computational load and memory usage compared to utilizing pre-trained object detectors for visual feature extraction [10] and has been increasingly adopted by recent methodologies in comparison to utilizing pre-trained object detectors for visual feature extraction. A visual transformer will be used to partition an input image(here, it will be the encoded feature list from the static CNN encoder instead) into discrete patches and encode them into a sequence of embeddings. An addition `[CLS]` token will be added [14].

Continuing and adapting the method proposed by BLIP V1 into parts of our framework for our model to understand the image-text pairs and generate captions requires a multimodal mixed encoder-decoder (MED) architecture. The mechanism processes text and images separately in their respective encoder models. Specifically, we use an Image Grounded Text Encoder, which incorporates an additional cross-attention (CA) layer within each transformer block of the text encoder, positioned between the self-attention (SA) layer and the feed-forward network (FFN). A task-specific `[Encode]` token is appended to the text, and the resulting embedding of `[Encode]` serves as the multimodal representation of the image-text pair. For decoding, the bidirectional self-attention layers of the image-grounded text encoder are replaced with causal self-attention layers. A `[Decode]` token marks the beginning of a sequence, and an end-of-sequence token indicates its termination.

3.3.2 Vision-Language Representation Learning

After the visual transformer outputs the embeddings, we utilize the Q-Former, which employs learnable query vectors to identify the most relevant visual features. This aligns visual representations with textual information, reducing the language model's computational load for synchronized feature learning.

We integrate BLIP-2's alignment mechanism into our pretraining mechanism to enhance the alignment between visual and textual modalities. This approach consists of two pretraining strategy phases using a lightweight Querying Transformer (Q-Former). The Q-Former is the bridging mechanism that facilitates the understanding of information between trained image encoders and large-scale language models (LLMs).

The Q-Former is set in this phase with the pre-frozen large language model (LLM). The Q-Former is conditioned to generate text from visual inputs by interpreting the visual representations in a manner perceptible to the LLM. Consequently, in the generative phase, the visual model's outputs are channeled to the Q-Former for transmutation into textual form.

Pre-training Objectives: We jointly optimize three objectives during pretraining, with two understanding-based objectives and one generation-based objective. Each image-text pair only requires one forward pass through the computationally heavier visual transformer, and three forward passes through the text transformer, where different functionalities are activated to compute the three losses as delineated below:

Image-Text Contrastive Loss (ITC): The ITC loss activates the unimodal encoder. It aims to align the feature space of the visual transformer and text transformers' feature space by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs. It is an effective method for improving vision and language understanding. We follow the ITC loss, where a momentum encoder is introduced to produce features. Soft labels are created from the momentum encoder as training targets to account for the potential positives in the negative pairs.

Image-Text Matching Loss (ITM): The ITM loss activates the image-grounded text encoder. Specifically, we have used the Contrastive Loss method for the image model according to ITM. This aims to learn an image-text multimodal representation that captures the fine-grained alignment between vision and language while allowing a wide diversity of image features and language tags to exist in the model. ITM is a binary classification task where the model uses an ITM head (a linear layer) to predict whether an image-text pair is positive (matched) or negative (unmatched) given its multimodal feature.

In order to find more informative negatives, we adopt the hard negative mining strategy, where negative pairs with higher contrastive similarity in a batch are more likely to be selected to compute the loss.

3.4. Language Modeling Loss

To complete the alignment and integration process, we focus on generating textual descriptions from visual inputs using Language Modeling Loss(LM Loss), which activates the image-grounded text decoder to generate textual descriptions from visual inputs. It undergoes a Cross-Entropy Loss (CEL) optimization, which allows the model to maximize the likelihood of the text in an auto-regressive way. A label smoothing technique with a coefficient of 0.1 is integrated during the loss computation.

In pursuit of efficient pre-training with multi-task learning, the text encoder and image decoder share parameters, except the Self-Attention (SA) layer, since the SA layers are pivotal in capturing the nuances between encoding and decoding tasks. Specifically, the encoder utilizes bidirectional self-attention to construct representations of the current input tokens, whereas the decoder employs causal self-attention to predict subsequent tokens. Conversely, the embedding layers, Cross-Attention (CA) layers, and Feed-Forward Network (FFN) perform similar functions in both tasks; thus, sharing these layers can augment training efficacy and capitalize on the synergies of multi-task learning.

4. Caption Decoder

4.1. CapFilt: Enhancement of Text Corpus Quality

Due to the high annotation costs, there is a need for high-quality, human-annotated image-text pairs $\{(I_h, T_h)\}$, such as those found in the COCO dataset. Current researches tries to leverage more extensive datasets of image and alt-text pairs $\{(I_w, T_w)\}$, procured automatically from the web. However, these alt-texts must be more aligned with the visual content, reducing the noise that detracts from learning vision-language relationships.

Here, we will use the method Cationing and Filtering (CapFilt) [14] to enhance the quality of the textual dataset. It encompasses two distinct modules:

- **Captioner**: Generating descriptive captions from web-sourced images.
- **Filter**: Eliminating non-relevant or incorrect image-text associations from the dataset.

Both modules are initialized with a pre-trained MED model and undergo independent fine-tuning using the COCO dataset. The captioner, an image-grounded text decoder, is refined with the LM objective to transcribe texts given images, yielding synthetic captions T_s for web images I_w.

The filter, an image-grounded text encoder, is fine-tuned with objectives ITC and ITM to detect text-image correspondence. This eliminates noisy texts from the original web texts T_w and the synthetic texts T_s, identifying a text as noisy if the ITM head detects it as incongruent with the image.

The filtered image-text pairs are combined with manually annotated pairs to create a new novel dataset that can train complex and advanced models.

Unlike traditional pre-trained Convolutional Neural Networks or the more efficient R-CNN frameworks, we have used the Transformer architecture as our model's primary visual encoder. The encoder extracts a set of grid features $V_G = v_1, v_2, \ldots, v_m$ from input images, where $v_i \in \mathbb{R}^D$, D represents the embedding dimension of each grid feature, and m denotes the total number of grid features.

Based on the transformer architecture, a novel encoder is implemented to refine these initial grid features, capturing how they relate. A mean-pooled global feature v_g is also integrated into the Window Multi-Head Self-Attention (W-MSA) and Shifted Window Multi-Head Self-Attention (SW-MSA) mechanisms. The refining encoder includes N sequentially stacked blocks, each alternating between W-MSA and SW-MSA modules, followed by a feedforward layer. The formulation for the l-th block is as follows:

$$\hat{V}_G^l = \left(V_G^{l-1} + \left(W_Q^l V_G^{l-1}, W_K^l \left[V_G^{l-1}; v_g^{l-1}\right]_s, W_V^l \left[V_G^{l-1}; v_g^{l-1}\right]_s\right)\right), \tag{5}$$

$$\hat{v}_g^l = \left(v_g^{l-1} + \left(W_Q^l v_g^{l-1}, W_K^l \left[V_G^{l-1}, v_g^{l-1}\right]_s, W_V^l \left[V_G^{l-1}, v_g^{l-1}\right]_s\right)\right), \tag{6}$$

$$V_G^l = \left(\hat{V}_G^l + (\hat{V}_G^l)\right), \tag{7}$$

$$v_g^l = \left(\hat{v}_g^l + (\hat{v}_g^l)\right). \tag{8}$$

The output refined grid features V_G^N and global feature v_g^N are subsequently input to the decoder for visual content processing [14].

4.2. Multi-Modal Decoder

The decoder sequentially generates captions based on the encoder's processed global and grid features. This stage is essential for integrating visual and textual modalities. The decoder architecture comprises N sequentially arranged blocks, each containing four primary components that we referenced from the paper on End-to-end transformer based model [30]:

Pre-Fusion Module: This stage starts the inter-modal interaction by combining the previously generated words with the refined global features.

Language Masked MSA Module: This module will help the related interactions between the generated captioning words, which will result in better caption coherence.

Cross MSA Module: Including a Multi-Head Self-Attention mechanism followed by a FeedForward layer, this is the second inter-modal interaction in the process to further relate visual and textual data.

Word Generation Module: Finally, to generate the texts sequentially in a normal probability of the learned content, a linear layer with a softmax function is employed in this module

By leveraging both global and grid features from the encoder and employing multiple stages of inter-modal interaction, our model can produce a more nuanced and comprehensive integration of visual and textual information with better caption coherence by the Language Masked MSA Module. This decoder structure complements the Context-Object Split (COS) architecture we used during the encoding process, providing an effective mechanism to utilize contextual and object-specific information during caption generation.

5. Experiment

5.1. Implementation Details

Our model, named CelsiaNet, is derived from our previous project's name, CELSIA [31]. However, it distinctly differs as it reoriented the goal to focus on multi-modal vision-language models (VLMs) rather than the computer vision automation framework of the earlier CELSIA project. CelsiaNet is developed within the LAVIS framework [19], which uses BLIP-2's visual encoder for image processing, LLaVA's language model for text generation, and a Q-former for multimodal alignment. The Q-former, similar to BERT architecture, utilizes learnable query vectors to refine the extraction of visual features, thereby enhancing multimodal alignment. Our model is trained on eight NVIDIA A100-80GB GPUs, uses the Adam optimizer, and has a batch size of 768. Our models are trained for five epochs with an initial learning rate of 1×10^{-4}, utilizing a cosine decay schedule. The LLM's beam size is set to three during inference, and a solitary caption is produced for each referenced region. This training periodicity is similar to the one stated by ControlCap [32], which we used in order to compare statistics in a controlled manner.

5.2. Datasets

For dense captioning, CelsiaNet is trained to utilize the Visual Genome (VG) and VG-COCO datasets, while the model is trained on the Visual Genome (VG) and RefCOCOg datasets for referring expression generation. The VG dataset consists of annotations of objects, attributes, and relationships, while VG-COCO combines the data from parts of VG V1.2 and MS COCO. RefCOCOg includes detailed descriptions of specific regions from diverse perspectives [32].

5.3. Evaluation

We follow the evaluation methods established in previous research, "ControlCap" and "DynRefer" [32] [33] to measure the performance of CelsiaNet in dense captioning tasks on VG and VG-COCO datasets, as well as in referring expression generation tasks on VG and RefCOCOg datasets. We use Mean Average Precision (mAP) as the primary metric for dense captioning, calculated across various localization and language accuracy thresholds. We use Intersection Over Union (IoU) thresholds of 0.3, 0.4, 0.5, 0.6, and 0.7 for evaluating localization and METEOR score thresholds of 0, 0.05, 0.1, 0.15, 0.2, and 0.25 for evaluating language generation. We use a GRiT model trained on VG to determine object locations, as CelsiaNet does not specialize in object location detection.

To evaluate the region-level captioning performance without the influence of localization, we asses the model using ground-truth bounding boxes during inference. For referring expression generation, we use METEOR and CIDEr scores to measure the quality of the captions produced by CelsiaNet. Unlike previous methods, CelsiaNet can generate customized captions based on interactive controls. The first noun in the ground-truth caption is utilized as the interactive control during inference to test this feature. For example, in the caption, "A tall building stands against the skyline," the word "building" is used as the interactive control [32].

5.4. Results and Analysis

By running the experiment and testing the mAP, evidenced by the results in Table 1, our approach achieves the highest mAP scores of 43.2%, 44.0%, and 44.4% on VG V1.0, VG V1.2, and VG-COCO datasets respectively on the With Ground Truth Test, surpassing the previous best results by a substantial

Table 1. The Results for Different Methodologies on Visual Genome Dataset. (Metrics from other works obtained from: [32])

Methods	GT localization	mAP(%)		
		VG V1.0	VG V1.2	VG-COCO
FCLN	N	5.4	5.2	-
JIVC	N	9.3	10.0	-
ImgG	N	9.3	9.7	-
COCD	N	9.4	9.8	7.9
COCG	N	9.8	10.4	8.9
CAG-Net	N	10.5	-	-
TDC	N	11.5	11.9	11.9
GRiT	N	15.5	16.4	-
CapDet	N	-	15.4	14.0
DCMSTRD	N	13.6	13.4	16.1
ControlCap [32]	N	18.2	18.5	18.4
Ours	**N**	**19.5**	**19.4**	**20.2**
FCLN	Y	27.0	-	-
JIVC	Y	33.6	-	-
CAG-Net	Y	36.3	-	-
GRiT	Y	40.0	40.3	-
BLIP2	Y	37.7	37.9	36.9
ControlCap [32]	Y	42.4	42.8	43.2
Ours	**Y**	**43.2**	**44.0**	**44.4**

margin, with another improvement from the Control Cap's results. The results correlate to higher performance at a better and larger dataset(from VG V1.0 to VG-COCO).

We analyze this change and improvement to be baselined by the end-to-end finetuning and the Content-Object Split(COS) that we have introduced from previous research. This explains the overall performance to increase a long range from BLIP V2(37.7—37.9—36.9 VS 43.2—44.0—44.4). Additionally, our design of the attention mechanisms derived from BLIP V2 at the end of each image encoding process redirects the embeddings to be more concentrated. Consequently, adding preprocessor methods from language processing algorithms also improves the total performance, as we have used the combinations of BERT preprocessor alongside the finetuned Q-Former at the end of the text encoding stage. As demonstrated in the test, older models tend to perform indifferently with a finer and larger dataset; due to the absence of self-attention and projection at the end of their encoding processes, their understanding of the original content is insufficient to differentiate from the experimental results. With the newly developed Control Cap algorithm, which has also integrated the BLIP V2 method in their encoders, we have improved the statistics by a substantial amount. This is due to the projection method, which better relates the textual encodings to the text when compositing in the ITC. We also have demonstrated a momentum on correlation to the given dataset, which symbolizes that the model does take in more understanding each time when better data are given to relate to. This has contributed to the increase in precision that our model has made.

Table 2. Parameters and Results for Different Methodologies on RefCOCOh and VG tests. (Metrics from other works obtained from: [32])

Method	Model size	RefCOCOg		VG	
		METEOR	CIDEr	METEOR	CIDEr
SLR+Rerank	<1B	15.9	66.2	-	-
GRiT	<1B	15.2	71.6	17.1	142.0
Kosmos-2	1.6B	14.1	62.3	-	-
GPT4RoI	7B	-	-	17.4	145.2
RegionGPT	7B	16.9	109.9	17.0	145.6
GLaMM	7B	16.2	105.0	18.6	157.8
Alpha-CLIP+LLaVA	7B	16.7	109.2	18.9	160.3
Osprey	7B	16.6	108.3	-	-
ControlCap	4.2B	17.0	111.4	20.4	181.9
Ours	**4.2B**	**17.3**	**112.6**	**21.7**	**182.6**

As for the No Ground-Truth Localization of our model, our model performs scores around 1 point more than the ControlCap model on VG V1.0 and VG V1.2 datasets while gaining a larger gap of around two more points than ControlCap on the VG-COCO dataset. As our model's images are trained under the unsupervised learning approach (that drops out the annotations at the start) in contradiction to ControlCap's supervised approach, our model has performed exceptionally well on the VG-COCO dataset by the outstanding accuracy on unannotated, no ground truth test. However, the absence of annotations did not cause much deficiency in understanding the model of images that come with annotated results(VG V1.0 and VG V1.2 datasets). From the results, with or without ground truth, our model demonstrates a better understanding ability than previous supervised learning models such as ControlCAP, DCMSTRD, and CapDet, even without the annotations given in the dataset. This proves our model's zero-shot capability in determining the caption for the images and our approach's improvement in the research of VLMs.

We evaluated our model's understanding ability using METEOR and CIDEr on RefCOCOg and VG datasets to test our model's understanding ability. As shown in Table 2, our model significantly outperforms recently developed models (CVPR 24'), especially on the VG test, achieving the highest improvement. With a smaller model size (4.2B vs. 7B), these strong results are attributed to our fine-tuned language methods (Q-Former and BERT preprocessor) and the integration of projected features with image features (post-self-attention), which enhances information exchange in the ITC portion. The new attention and projection methods improve performance compared to previous research.

In our model's offline evaluation results of our proposed model and other existing state-of-the-art models on the MSCOCO "Karpathy" test split(Table 3), for fair comparisons, we evaluated our model using both the single model and ensemble model test by training four models using SCST. Our results in Table 3 show that our model has achieved state-of-the-art performance across all metrics compared to previous works. For the single model experiment, we attained a CIDEr score of 139.2%, overcoming the scores of RSTNet [34] and DLCT [35] by 3.6% and 5.4% respectively; It also shows an improvement of 0.8% in BLEU-1, 1.3% in BLEU-4, 0.9% in METEOR, 1.1% in ROUGE-L, and 1.3% in SPICE, compared to the best statistics of the listed newly developed state-of-the-art models.

Table 3. Results of the offline evaluation for our model alongside other leading models on the MSCOCO “Karpathy” test split. The metrics B-N, M, R, C, and S correspond to BLEU-N, METEOR, ROUGE-L, CIDEr, and SPICE, respectively. (Metrics from other works obtained from: [30])

Models	Single Model						Ensemble Model					
	B-1	B-4	M	R	C	S	B-1	B-4	M	R	C	S
CNN-LSTM based models												
SCST	-	34.2	26.7	55.7	114.0	-	-	35.4	27.1	56.6	117.5	-
RFNet	79.1	36.5	27.7	57.3	121.9	21.2	80.4	37.9	28.3	58.3	125.7	21.7
Up-Down	79.8	36.3	27.7	56.9	120.1	21.4	-	-	-	-	-	-
HAN	80.9	37.6	27.8	58.1	121.7	21.5	-	-	-	-	-	-
GCN-LSTM	80.5	38.2	28.5	58.3	127.6	22.0	80.9	38.3	28.6	58.5	128.7	22.1
SGAE	80.8	38.4	28.4	58.6	127.8	22.1	81.0	39.0	28.4	58.9	129.1	22.2
AoANet	80.2	38.9	29.2	58.8	129.8	22.4	81.6	40.2	29.3	59.4	132.0	22.8
X-LAN	80.8	39.5	29.5	59.2	132.0	23.4	81.6	40.3	29.8	59.6	133.7	23.6
CNN-Transformer based models												
ORT	80.5	38.6	28.7	58.4	128.3	22.6	-	-	-	-	-	-
ETA	81.5	39.9	28.9	59.0	127.6	22.6	81.5	39.9	28.9	59.0	127.6	22.6
X-Transformer	80.9	39.7	29.5	59.1	132.8	23.4	81.7	40.7	29.9	59.7	135.3	23.8
$\mathcal{M}^2$ Transformer	80.8	39.1	29.2	58.6	131.2	22.6	82.0	40.5	29.7	59.5	134.5	23.5
RSTNet	81.1	39.3	29.4	58.5	133.3	23.0	-	-	-	-	-	-
RSTNet	81.8	40.1	29.8	59.5	135.6	23.3	-	-	-	-	-	-
GET	81.5	39.5	29.3	58.9	131.6	22.8	82.1	40.6	29.8	59.6	135.1	23.8
DLCT	81.4	39.8	29.5	59.1	133.8	23.0	82.2	40.8	29.9	59.8	137.5	23.3
Ours	**82.6**	**41.4**	**30.7**	**60.6**	**139.2**	**24.7**	**83.9**	**42.6**	**31.0**	**61.3**	**142.0**	**24.8**

Table 4. Results from the online evaluation of our proposed model compared to other state-of-the-art models on the MSCOCO dataset. (Metrics from other works obtained from: [30])

Models	BLEU-1		BLEU-2		BLEU-3		BLEU-4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
SCST	78.1	93.7	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
GCN-LSTM	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
Up-Down	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
SGAE	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
AoANet	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
X-Transformer	81.9	95.7	66.9	90.5	52.4	82.5	40.3	72.4	29.6	39.2	59.5	75.0	131.1	133.5
$\mathcal{M}^2$ Transformer	81.6	96.0	66.4	90.8	51.8	82.7	39.7	72.8	29.4	39.0	59.2	74.8	129.3	132.1
RSTNet	82.1	96.4	67.0	91.3	52.2	83.0	40.0	73.1	29.6	39.1	59.5	74.6	131.9	134.0
GET	81.6	96.1	66.5	90.9	51.9	82.8	39.7	72.9	29.4	38.8	59.1	74.4	130.3	132.5
DLCT	82.4	96.6	67.4	91.7	52.8	83.8	40.6	74.0	29.8	39.6	59.8	75.3	133.3	135.4
Ours	**83.4**	**97.7**	**68.6**	**92.3**	**54.1**	**84.4**	**41.9**	**74.6**	**30.6**	**40.4**	**61.4**	**76.4**	**137.5**	**139.3**

On the other hand, our ensemble model has demonstrated high consistency across different training runs and outperforms other models by more than 1.0% across all metrics. Notably, the ensemble model achieves a CIDEr score of 142.0%, outperforming DLCT and GET by 4.5% and 6.9%, respectively.

In Table 4, we list the performance of our model in comparison to previous state-of-the-art models on the MSCOCO official online test server. We used five reference captions(c5) and forty reference captions(c40) for each testing method. Our model has achieved the highest scores across all metrics by a decent improvement of around 0.8% to 1.0%, while our CIDEr scores for c5 and c40, which are 137.5% and 139.3%, respectively, representing improvements of 4.2% and 3.9% over the metrics of DLCT. These results underscore the robustness of our model across different evaluation criteria. The consistently excellent performance under the tests for having either 5 or 40 reference captions demonstrates the model's ability to generate captions that align well with human-generated descriptions while outperforming other models by having a significant lead under situations when fewer reference captions were given(c5). Another notable improvement is that the CIDEr scores(sensitive to caption quality and relevance), similar to the last test, perform with excellent accuracies that surpass the results of the other models. Our model has outperformed other models in all the comparison metrics, especially for CIDEr, demonstrating our framework's effectiveness and efficiency. Moreover, in order to take into account this, we analyze this in two aspects: From a module-level analysis, in contrast to models utilizing region-level features or a combination of region and grid-level features, our approach offers a better computationally efficient method using Content Object Split. The end-to-end fine-tuning method allows the model to better relate the possible relationships in image scenarios. On the other hand, the basis of the program has enabled impressive performance improvements and a stable architecture in our model. By integrating a multi-stage training process inspired by LLaVA and BLIP-2 [10] [11], along with the context-object split (COS) factorization [12], the model achieves a more detailed understanding of both image context and specific object details. For the training method, the unsupervised training with the pseudo-supervision mechanism enhances caption diversity while maintaining accuracy, while the fine-tuning method allows for better text-image relationship learning. These synchronous components produce more coherent, accurate, and diverse captions, resulting in gains across all test scores and state-of-the-art results in image captioning tasks.

6. Conclusion

In this research, we have developed an advanced Multimodal Vision Language Model for better zero-shot image captioning. We have integrated advantageous modules from previous researchers' works and adapted them with our mechanisms that assure coherence. These include the methods: Mechanisms derived from BLIP V2, implementing a Content-Object-Split(COS), creating an effective combination of BERT preprocessor with the fine-tuned Q-Former for text encoding, and utilizing the Image-Text Composer(ITC). All of these are carried out using an end-to-end fine-tuning approach.

Our proposed model consistently outperforms existing state-of-the-art methods across various datasets and evaluation metrics, including Visual Genome, RefCOCOg, and MSCOCO. We achieve the highest scores in mAP, METEOR, CIDEr, BLEU, and ROUGE-L, showing the effectiveness of our approach in both region-level captioning and overall image captioning tasks. This superior performance is attributed to its ability to effectively leverage model size and architectural innovations, as evidenced by the significant improvements over solid baseline models like ControlCap and DLCT.

There are several avenues for developing VLMs, as we can use more complex attention techniques, such as multi-headed attention, to improve the model's ability to focus on relevant image features. Unlike basic image regional analysis, we can also leverage a layer-of-depth calculation model to give the model spatial awareness that it can precisely compute. Integrating more recent language models or fine-tuning techniques could improve the text generation component. These are all ways the VLM may be improved, and we will continue exploring them in future research.

References

[1] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024.

[2] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. arXiv preprint arXiv:2404.07214, 2024.

[3] Gaudenz Boesch. Vision language models: Exploring multimodal ai, June 2024. Accessed on July 20, 2024.

[4] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision language model. arXiv preprint arXiv:2406.01584, 2024.

[5] Mingsheng Yin, Tao Li, Haozhe Lei, Yaqi Hu, Sundeep Rangan, and Quanyan Zhu. Zero-shot wireless indoor navigation through physics-informed reinforcement learning, 2023.

[6] Yuhan Zhou, Fengjiao Tu, Kewei Sha, Junhua Ding, and Haihua Chen. A survey on data quality dimensions and tools for machine learning, 2024.

[7] Amit Das, Zheng Zhang, Fatemeh Jamshidi, Vinija Jain, Aman Chadha, Nilanjana Raychawdhary, Mary Sandage, Lauramarie Pope, Gerry Dozier, and Cheryl Seals. Investigating annotator bias in large language models for hate speech detection, 2024.

[8] Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. Regiongpt: Towards region understanding vision language model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13796–13806, 2024.

[9] Hugging Face Team. Vision language models explained. https://huggingface.co/blog/vlms, 2023.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR, 2023.

[11] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. Advances in neural information processing systems, 36, 2024.

[12] Shweta Mahajan and Stefan Roth. Diverse image captioning with context-object split latent spaces. Advances in Neural Information Processing Systems, 33:3613–3624, 2020.

[13] Llinet Benavides Cesar, Miguel-Ángel Manso-Callejo, and Calimanut-Ionut Cira. Bert (bidirectional encoder representations from transformers) for missing data imputation in solar irradiance time series. Engineering Proceedings, 39(1):26, 2023.

[14] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International conference on machine learning, pages 12888–12900. PMLR, 2022.

[15] Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[16] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning, 2024.

[17] Xiang An, Jiankang Deng, Kaicheng Yang, Jaiwei Li, Ziyong Feng, Jia Guo, Jing Yang, and Tongliang Liu. Unicom: Universal and compact representation learning for image retrieval. arXiv preprint arXiv:2304.05884, 2023.

[18] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The (r) evolution of multimodal large language models: A survey. arXiv preprint arXiv:2402.12451, 2024.

[19] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C. H. Hoi. Lavis: A library for language-vision intelligence, 2022.

[20] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[22] Tony Lee, Yifan Mai, Chi Heem Wong, Josselin Somerville Roberts, Michihiro Yasunaga, Faarzan Kaiyom, Rishi Bommasani, and Percy Liang. The first steps to holistic evaluation of vision-language models, May 2024.

[23] Santiago Castro, Oana Ignat, and Rada Mihalcea. Scalable performance analysis for vision-language models. arXiv preprint arXiv:2305.18786, 2023.

[24] Yidong Wang, Zhuohao Yu, Jindong Wang, Qiang Heng, Hao Chen, Wei Ye, Rui Xie, Xing Xie, and Shikun Zhang. Exploring vision-language models for imbalanced learning. International Journal of Computer Vision, 132(1):224–237, 2024.

[25] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models, 2024.

[26] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence, 45(1):539–559, 2022.

[27] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6):1–36, 2019.

[28] Chao Zeng and Sam Kwong. Learning cross-modality features for image caption generation. International Journal of Machine Learning and Cybernetics, 13(7):2059–2070, 2022.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[30] Yiyu Wang, Jungang Xu, and Yingfei Sun. End-to-end transformer based model for image captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 2585–2594, 2022.

[31] Chengjui Fan. Computer vision enhanced lightweight system for interface automation (celsia). Applied and Computational Engineering, 2024.

[32] Yuzhong Zhao, Yue Liu, Zonghao Guo, Weijia Wu, Chen Gong, Fang Wan, and Qixiang Ye. Controlcap: Controllable region-level captioning, 2024.

[33] Yuzhong Zhao, Feng Liu, Yue Liu, Mingxiang Liao, Chen Gong, Qixiang Ye, and Fang Wan. Dynrefer: Delving into region-level multi-modality tasks via dynamic resolution. arXiv preprint arXiv:2405.16071, 2024.

[34] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Rstnet: Captioning with adaptive attention on visual and non-visual words. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 15465–15474, 2021.

[35] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning, 2021.

7. Acknowledgment

In recent years, significant improvements in computer vision and textual analysis networks have been made. As a result of the development, there are more and more needs in multiple fields that would potentially require one robust model that suits both while remaining highly accurate and relatable in the two topics. Previous studies have continued advancing computer vision or language processing networks, while new methods use multiple combined networks to solve the issue. From the given basis of state-of-the-art vision and textual networks, we want to establish an optimal network that will integrate each network's advantages while maintaining high consistency between the networks. Furthermore, we aim to achieve zero-shot capability and higher relation knowledge of the content given in both image and text. By researching the multimodal vision language framework and conducting multiple experiments that will prove the model's superiority under the guidance of our instructor, we have successfully chosen the topic, completed this research paper, and established our state-of-the-art model, CelsiaNet.

I, Chengjui Fan, was responsible for information research, framework interpretation(before and after improvement), programming, experimental analysis, and paper conclusion.

Tianzhou He is the instructor of my paper. He has provided me with possible ideas for network integration, how to narrow the research down to a specific range, and how to conduct our experiments. His guidance is voluntary and without remuneration, and I am grateful for his support during the research.