Proceedings of ICBioMed 2024 Workshop: Computational Proteomics in Drug Discovery and Development from Medicinal Plants
DOI: 10.54254/2753-8818/74/2024.LA18779

Analysis and Experimental Study of Zero-Shot Capabilities in Vision-Language Models: An In-Depth Exploration of Contrastive and Masked Methods

Zhi Xu

Boston College, Chestnut Hill, United States

xuds@bc.edu

Abstract. This paper conducts a thorough analysis and experimental evaluation of zero-shot capabilities in vision-language models (VLMs), concentrating on three distinct approaches: contrastive learning, masked learning, and generative modeling, exemplified by CLIP, FLAVA, and CoCa, respectively. CLIP uses contrastive learning to align images and text robustly, FLAVA employs masked learning to improve multimodal reasoning, and CoCa combines generative captioning with contrastive learning for fine-grained multimodal comprehension. Zero-shot learning, a pivotal AI capability, allows models to apply knowledge to new tasks without further training specific to those tasks. The performance of these models is tested through experiments in zero-shot settings, including image classification on datasets like CIFAR-100, Flowers102, and Food101, to evaluate generalization to new image categories. Furthermore, zero-shot image and text retrieval tasks are performed using Flickr30k and MSCOCO benchmarks to measure the models' ability to align and retrieve across modalities without direct supervision. Results from these tests provide a comprehensive look at the VLMs' zero-shot performance, highlighting their potential and limitations in real-world applications on unseen data.

Keywords: Zero-shot capability, image classification, contrastive learning, generative learning.

1. Introduction

Recent advancements in the field of language modeling have led to significant achievements, particularly with the development of Large Language Models (LLMs) such as Llama and ChatGPT. Historically focused on processing and generating text, these models are now evolving due to efforts to expand their capabilities to include visual inputs, thus enabling the integration of textual and visual data. Vision-Language Models (VLMs) represent a powerful advancement in this domain, utilizing large-scale datasets and diverse methodologies to learn representations that effectively bridge the gap between images and text [1]. These models are adept at performing various downstream tasks like image captioning, image-text retrieval, and visual question answering with notable accuracy, underscoring their utility in multimodal learning.

Despite the successes, a significant challenge in VLM development is their ability to generalize to new tasks or data—specifically through zero-shot learning [2]. Zero-shot learning capabilities enable VLMs to perform tasks or make predictions about data or classes not previously encountered during their training. This capability is paramount for creating versatile and robust AI systems, especially in

DOI: 10.54254/2753-8818/74/2024.LA18779

real-world applications where models must adapt to a wide range of scenarios without needing specific fine-tuning. Exploring and enhancing the zero-shot learning abilities of VLMs remain a critical focus for researchers aiming to extend the models' applicability and functionality.

Research Content of This Paper: This paper aims to scrutinize the zero-shot capabilities of three distinct VLMs: CLIP, FLAVA, and CoCa [3, 4, 5]. Each model embodies a unique approach to learning multimodal representations—CLIP leverages contrastive learning, FLAVA utilizes masked learning techniques, and CoCa combines generative with contrastive methods. By conducting experiments focused on zero-shot image classification and zero-shot image and text retrieval, this study will provide a comprehensive analysis of these models' performance in zero-shot scenarios. The findings from these experiments will illuminate the strengths and limitations of current VLMs in handling zero-shot tasks, pointing to potential avenues for future research and enhancements [6]. This comparative analysis intends to contribute significantly to the development of more generalizable and adaptable AI systems, suitable for complex real-world applications.

2. Relevant Theories

2.1. Contrastive learning-based vision-language models

One of the first explored initiatives for VLMs is Contrastive learning, and the core idea behind it is, as the name suggests, to train models to produce similar representations for matching (positive) pairs and different representations for mismatching (negative) pairs. This is done by maximizing the similarity between paired examples, which in this cases would be an image-caption pair, and minimizing the similarity between mismatched pairs, which is implemented using infoNCE contrastive loss introduced by Oord in 2018 [7] such that:

$$L_{\text{InfoNCE}} = -\sum_{(i,j)\in P} log\left(\frac{exp(\text{Sim}(z_i,z_j)/\tau)}{\sum_{k=1}^{N} exp(\text{Sim}(z_i,z_k)/\tau)}\right) \quad (1)$$

The InfoNCE loss equation utilizes a softmax and a temperature parameter to optimize the similarity between matching pairs while reducing the similarity of all other unmatching pairs in the batch.

Contrastive learning is essential to zero-shot capability because it enables models to learn generalized and robust representations by aligning semantically similar pairs while distinguishing them from dissimilar pairs in a shared embedding space. This process allows the model to understand the underlying connections between different modalities, which is crucial for effectively transferring knowledge to unseen tasks or classes without requiring additional task or dataset specific training.

One of the most famous implementations of contrastive learning in VLMs is the CLIP (Contrastive Language-Image Pre-training) model developed by OpenAI [8]. CLIP leverages contrastive learning through Dual-Encoder pertaining by concurrently optimizing an image encoder and a text encoder to bring matching pairs closer together in the embedding space while distancing the embedding vectors of unmatching pairs. For a batch of N image-text pairs, CLIP maximize the similarity or distance of N positive pairs, at the same time minimizing the similarity of the N^2 - N negative pairs. This approach not only allows the model to learn rich, generalizable representations of both visual and textual data but also facilitates zero-shot transfer.

The CLIP model optimizes a cross-entropy loss over the cosine similarities of the image and text embeddings, using a technique based on the InfoNCE loss. This approach is similar to the methods used in deep metric learning, which aims to effectively acquire a distance metric that accurately reflects the similarity between paired examples [9].

A notable achievement of CLIP model is that it is one of the first VLMs to exhibits exceptional performance on zero-shot transfer. By training two aligned encoders on a diverse dataset, CLIP can execute zero-shot classification and retrieval tasks across various domains with relatively good accuracy in the absence of additional fine-tuning on specific datasets. For instance, CLIP raised the zero-shot transfer image classification result on ImageNet to 76.2% in comparison to prior result of 11.5% yield by Visual N-Grams.

DOI: 10.54254/2753-8818/74/2024.LA18779

However, a challenge of CLIP model is its reliance on large and diverse dataset to ensure a diverse set of negative pairs, which is essential to the model's zero-shot capability. CLIP was trained on a dataset of 400 million image-text pairs collected from the internet. This enormous scale allowed CLIP to learn more robust and generalized representations, which contributed to its strong performance across many zero-shot tasks.

2.2. Masked learning-based vision-language models

Masked learning has become increasingly prevalent in vision-language models (VLMs) for its effectiveness in improving model robustness and learning strong visual representations. The central concept behind this approach is to mask certain portions of the input data and train the model to predict the patched information. On the textual side, Masked learning strategies have been popularized by BERT model due to its accomplishment in natural language processing tasks, where Masked Language Modeling (MLM) is used to predict masked word tokens in a sentence leveraging the surrounding context [10]. The approach has been extended to the visual domain through Masked Image Modeling (MIM), which involve masking areas of an image and optimizing the model to reconstruct missing parts through the unmasked patches.

An examples of a VLM with masking objective is FLAVA (Foundational Language and Vision Alignment), which employs masked learning strategies to reach state-of-the-art performance across a broad spectrum of tasks. FLAVA integrates multiple masking objectives, including Masked Multimodal Modeling (MMM), Masked Image Modeling (MIM), and Masked Language Modeling (MLM), into a unified framework. This approach allows FLAVA to learn strong, shared representations of both images and text, making it highly effective for tasks across different modalities.

The architecture of FLAVA is the combination of dual-encoder approach and fusion encoder approach, using the Vision Transformer (ViT) architecture as template. The image encoder processes input images by dividing them into patches, and feed them into a transformer model for feature extraction. Similarly, the text encoder tokenizes and processes textual inputs to output hidden state vectors. The outputs of these encoders are then combined in a multimodal encoder that fuses the visual and textual representations through cross-attention mechanism [11]. The masking objective along with other pertaining objectives, including Global contrastive loss, allows FLAVA to perform well not only on multimodal tasks but also on unimodal tasks, making it a versatile model in the realm of vision-language processing.

The FLAVA model is pretrained on 70 millions publicly available text and image pairs, a much smaller dataset in comparison to the CLIP model mentioned above [12]. On top of that, the model is also designed to be able to learn strong representation from also unpaired unimodal data, therefore further reduces the data requirement of VLMs. Through the application of masked learning objectives, FLAVA showcases exceptional versatility and effectiveness, reaching state-of-the-art results across 35 varied tasks that encompass vision, language, and multimodal benchmarks, thereby highlighting the model's capability to comprehend and synthesize information across multiple domains.

2.3. Generative-based VLMs

Generative-based VLMs offer a distinct paradigm compared to contrastive or masked learning approaches, focusing on the generation of new content, in the form of text or images, rather than aligning existing data. These models focus on generating complete text or image outputs based on learned representations, enabling advanced tasks including image captioning, text to image synthesis, and more complex vision-language understanding.

One prominent example of generative-based VLMs is the Contrastive Captioner (CoCa) model. CoCa is designed to integrate contrastive learning with generative modeling in a single architecture, combining the strengths of both approaches. CoCa employs a dual-objective training method, where it learns to align image and text embeddings through contrastive loss while simultaneously generating contextually appropriate textual descriptions through a generative captioning loss. This dual objectives

DOI: 10.54254/2753-8818/74/2024.LA18779

allows CoCa to excel in both alignment tasks, like image-text retrieval, and generation tasks, such as image captioning.

The CoCa model features an encoder-decoder architecture very standard to generative models. The Image Encoder is a standard convolutional neural network (CNN) or a vision transformer (ViT) that extract the feature vectors of the input image. However, the text decoder is decoupled into two parts: a unimodal decoder that processes text alone and a multimodal decoder that integrates visual information from the image encoder. This decoupled design enables CoCa to seamlessly perform contrastive learning objective and generative learning objective, where the unimodal decoder focuses on aligning image and text representations, while the multimodal decoder focuses on generating coherent and contextually accurate text based on image input. During pretraining, the encoder-decoder is trained with teacher-forcing technique, feeding the model with ground truth text tokens at each step, to minimize the Captioning Loss (L_{Cap}) [13]:

$$L_{\text{Cap}} = -\sum_{t=1}^{T} logP_{\theta}(y_t \mid y_{<t}, x) \tag{2}$$

CoCa's ability to generate and align multimodal data makes it a versatile model equipped to handle a wide variety of vision-language tasks with minimal adaptation.

CoCa is pretrained on two large-scale datasets, the ALIGN dataset and JFT-3B, that include both annotated images with noisy labels and images with alt text. The model is trained by treating all labels as text, which enables the model to learn from a diverse and noisy dataset. By leveraging large-scale datasets and combining different training objectives, the model exhibits state-of-art performance across various vision-language tasks, namely zero-shot image classification, image-text retrieval, and visual question answering (VQA).

3. Experiments and Results

This section presents the setup and results obtained from evaluating three Vision Language Models (VLMs): CLIP, FLAVA, and CoCa. The evaluation consisted of three categories of downstream tasks: image classification, image to text retrieval, and text to image retrieval. Also, all experiment are conducted under zero-shot scenarios, meaning that no further fine-tuning or specific training is done to enhance the models' performance in these tasks.

3.1. Datasets and model setup

To evaluate the zero-shot capabilities of the VLMs, a diverse set of datasets was selected, covering various domains and categories:

Image Classification: Zero-Shot image classification were conducted on CIFAR-100, CIFAR-10, MNIST, Fashion-MNIST, Flowers102, and Food101. These datasets were selected to cover a broad range of image classification challenges, from simple digit recognition (MNIST) to complex and diverse food and flower categories (Food101, Flowers102).

Image-Text Retrieval: The Flickr30K and MSCOCO datasets were used for text and image retrieval tasks, as they are very common and effective benchmarks for retrieval tasks.

Model Setup: All three VLMs (CLIP, FLAVA and CoCa) evaluated utilize the ViT-B/32 Transformer architecture as their visual backbone. This means each model employs the base configuration Vision Transformer with 32x32 image patch size as the image encoder. While this configuration is not the most advanced version available for these VLMs, for example ViT-B/16 with smaller patches and ViT-L/14 using large ViT models offer better performance across various tasks, the ViT-B/32 configuration is more computational efficient and versatile enough to handle basic vision-language tasks from image classification to image-text retrieval. Additionally, it is worth to stress again that these models were evaluated without any task-specific fine-tuning to assess their generalization capabilities in a zero-shot setting. For image classification, the models were tasked with assigning each image to one of the class labels in the dataset. For image and text retrieval tasks, the models were

DOI: 10.54254/2753-8818/74/2024.LA18779

evaluated on their ability to retrieve the correct text given an image and vice versa. As shown in Table 1.

Table 1. Zero-shot performance on Flickr30K and MSCOCO datasets (1K test set).

Model	Dataset	Image → Text (R@1)	Image → Text (R@5)	Image → Text (R@10)	Text → Image (R@1)	Text → Image (R@5)	Text → Image (R@10)
CLIP	Flickr30K	0.7160	0.9050	0.9420	0.6610	0.8860	0.9310
FLAVA	Flickr30K	0.7300	0.9500	0.9740	0.7510	0.9430	0.9740
CoCa	Flickr30K	0.7420	0.9180	0.9480	0.7250	0.9110	0.9520
CLIP	MSCOCO	0.5230	0.8180	0.9120	0.4850	0.7880	0.8720
FLAVA	MSCOCO	0.6200	0.8970	0.9660	0.5830	0.8780	0.9560
CoCa	MSCOCO	0.5810	0.8560	0.9200	0.5510	0.8110	0.9090

3.2. Zero-shot image-text retrieval

The image-text retrieval task under zero-shot scenario was conducted following the setup described in the CLIP paper. First, the images and captions are preprocessed and passed through the model's encoders to extract their respective features, image or text. These features are then normalized, and went through a dot product operation to obtain in cosine similarity scores. Finally, the caption (or image) with the highest similarity score is retrieved. Table 1 illustrates all the experiment results for this task.

CLIP: CLIP demonstrated a relatively moderate success on both datasets. For the Flickr30K dataset, it achieved a top-1 recall (R@1) of 0.7160 and 0.6610 for image to text and text to image retrieval, respectively. On the more challenging MSCOCO dataset, CLIP achieved an R@1 of 0.5230 for the former and 0.4850 for the latter.

FLAVA: FLAVA outperformed CLIP at retrieval on both dataset. On Flickr30K, FLAVA achieved an R@1 of 0.7300 for image to text retrieval and 0.7510 for text to image retrieval. FLAVA's performance on MSCOCO was also strong, with an R@1 of 0.6200 and 0.5830 for the two tasks, repectively.

CoCa: CoCa performs good in retrieval tasks, particularly on the Flickr30K dataset, where it achieved the highest R@1 of 0.7420 for image to text retrieval. On MSCOCO, CoCa maintained strong performance with an R@1 of 0.5810 for image to text retrieval and 0.5510 for text to image retrieval.

Table 2. Zero-shot classification performance on Cifar-100, Cifar-10, MNIST, Fashion-MNIST, Flowers102, and Food101 (1K test set).

Model	CIFAR-100	CIFAR-10	MNIST	Fashion-MNIST	Flowers102	Food-101
CLIP	0.5570	0.8760	0.3160	0.6330	0.6060	0.6560
FLAVA	0.0130	0.1320	0.1160	0.0620	0.0020	0.0010
CoCa	0.7040	0.9310	0.3850	0.7740	0.5980	0.7130

3.3. Zero-shot image classification

The image classification task is also done following the CLIP paper in a similar fashion: the class labels of the corresponding dataset are tokenized and passed through the encoder to extract text feature for each class. These text features are then used to calculate cosine similarity in conjunction with the image embedding, and the class label that has the highest similarity score is predicted. The classification results are presented in Table 2.

CLIP: CLIP demonstrated strong performance across most datasets with with a accuracy of 0.8760 on CIFAR-10, 0.6560 accuracy on Food101, and achieved the highest accuracy on Flowers102 among the three models, 0.6060. However, CLIP's performance dropped on the simpler MNIST (0.3160).

DOI: 10.54254/2753-8818/74/2024.LA18779

FLAVA: FLAVA struggled with zero-shot classification task, particularly on the Flowers102 and Foood101 datasets, where it achieved accuracies of 0.0020 and 0.0010, respectively. FLAVA's performance was similarly low on the other datasets, with the highest accuracy being 0.1160 on MNIST.

CoCa: CoCa outperformed the other models in most classification tasks, achieving a accuracy of 0.9310 on CIFAR-10 and 0.7040 on CIFAR-100. CoCa also performed well on Fashion MNIST (0.7740) and Food101 (0.7130)

3.4. Discussion

Across the two categories of tasks, CoCa consistently outperforms the other two models. Its architecture, which integrates both contrastive and generative learning objectives, appears to provide a solid foundation for handling a diverse range of multimodal objectives, indicating it capability to learn effective and versatile representations of the dataset useful for both distinguishing of classes and aligning images with text.

FLAVA, while excelling in image-text retrieval tasks, benefiting from its masked learning strategy, struggles with image classification tasks. This may be the result of the model's training focus on multimodal alignment that favors tasks involving cross-modal understanding at the expense of its performance in fundamentally unimodal classification task.

CLIP, with its contrastive learning-based approach, remains balanced and competitive in both tasks, especially in classification scenarios. However, its performances are generally overshadowed by the CoCa model across basically all experiments. This could be attributed to the fact that CoCa is one of the earliest VLM and later models, including FLAVA and CoCa, have borrowed from and incorporated CLIP's innovative use of a contrastive learning objective that aligns visual and textual representations in a mutual embedding space into their own design. This further underscores the significance contrastive learning for zero-shot capabilities.

In summary, while each model has its strengths, CoCa' s versatility across different tasks makes it the most well-rounded model in zero-shot scenarios. FLAVA shows great promise in retrieval tasks but struggles with classification, while CLIP remains a solid option for both retrieval and classification. On a side note, the poor performance of all three models on the MNIST dataset—consisting of low-resolution (28x28 pixels) grayscale images of handwritten digits—despite their strong results on more complex datasets, is intriguing. This suggests that these VLMs rely heavily on rich textures, colors, and details to learn strong representation for different classes.

4. Limitation and Bias

Although the experiments carried out in this paper provide valuable insights into the zero-shot capabilities of the three VLMs, several limitations and potential biases must be acknowledged.

Model Configuration: One significant limitation is the use of the ViT-B/32 model configuration instead of the more advanced configuration like ViT-L/14. While computationally efficient, the ViT-B/32 model has a much smaller number of parameters and a lower capacity compared to ViT-L/14. This reduced capacity constraints the models' ability to capture complex patterns and representations, especially in tasks requiring fine-grained visual understanding. Therefore, the performance results observed in this study might underestimate the full potential of these models if more advanced configurations were used and the results should be interpreted with this limitation in mind.

Image and Text Retrieval Setup: Both the Flickr30K and MSCOCO datasets provide multiple captions per image, providing a richer and more comprehensive textual context that could enhance retrieval performance. However, only one caption per image was utilized for the image and text retrieval for the sake of computational efficiency. By limiting the evaluation to one caption per image, the experiment may not fully capture the models' capabilities in understanding and aligning with diverse textual descriptions. This simplification may be particularly limiting in scenarios where different captions highlight different aspects of an image.

The above decisions were made to optimize computational resources to allow for easier replication of the experiments. However, these computational efficient choices may introduce limitations to the

DOI: 10.54254/2753-8818/74/2024.LA18779

findings in this paper. Future studies could address these limitations by exploring more advanced model configurations and more comprehensive data for retrieval task.

5. Conclusion

This paper has conducted a comparative evaluation of three prominent Vision-Language Models—CoCa, FLAVA, and CLIP—highlighting their performance across various zero-shot retrieval and classification tasks. The analysis revealed that CoCa offers remarkable versatility, showing robust performance in both retrieval and classification tasks. In contrast, FLAVA excels specifically in retrieval tasks but shows limitations in classification scenarios. CLIP, utilizing a solely contrastive learning objective, provides balanced and competitive results across both domains. These outcomes underscore the potential benefits of integrating contrastive learning with other training methodologies to boost a model's zero-shot capabilities. This insight is crucial for enhancing the effectiveness of VLMs in handling diverse and complex tasks without additional task-specific training.

There is substantial scope for advancing the research on Vision-Language Models by exploring hybrid training techniques that combine the strengths of contrastive, generative, and other learning strategies. Future studies could focus on developing new models that incorporate these integrated approaches to further improve zero-shot learning capabilities. Additionally, extending the evaluation framework to include a broader range of tasks and datasets could provide deeper insights into the models' versatility and real-world applicability. Investigating the impact of different training data scales and modalities on the performance of VLMs will also be critical. Ultimately, these efforts will contribute to the ongoing refinement of VLM technologies, making them more adaptable and efficient for practical applications in diverse fields such as autonomous navigation, interactive robotics, and digital content management.

References

[1] Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., ... & Chandra, V. (2024). An introduction to vision-language modeling. In arXiv preprint arXiv:2405.17247.

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & others. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning (pp. 8748–8763). PMLR.

[3] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., & Kiela, D. (2022). FLAVA: A foundational language and vision alignment model. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15638–15650).

[4] Wang R., Zhu J., Wang S., Wang T., Huang J., Zhu X. Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking. International Journal of Multimedia Information Retrieval, 2024, 13(4): 39.

[5] Krizhevsky, A., & Hinton, G., et al. (2009). Learning multiple layers of features from tiny images. In Citeseer.

[6] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In Proceedings of the IEEE (pp. 2278-2324).

[7] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms.

[8] Zhu, X., Huang, Y., Wang, X., & Wang, R. (2023). Emotion recognition based on brain-like multimodal hierarchical perception.Multimedia Tools and Applications, 1-19.

[9] Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101 – Mining discriminative components with random forests. In Proceedings of the European Conference on Computer Vision (pp. 557-570).

[10] Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (2024). A Review of Key Technologies for Emotion Analysis Using Multimodal Information. Cognitive Computation, 1-27.

DOI: 10.54254/2753-8818/74/2024.LA18779

[11] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In Proceedings of the European Conference on Computer Vision (pp. 740–755).

[12] Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. In arXiv preprint arXiv:1807.03748.

[13] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the North American Chapter of the Association for Computational Linguistics (pp. 4171–4186).