Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/45/20241179

Principle analysis of five prominent large image classification models

Maier Shi

School of Software and Internet of Things Engineering, Jiangxi University of Finance and Economics, Nanchang, Jiangxi, 330000, China

2202004476@stu.jxufe.edu.cn

Abstract. Deep learning models, based on transformer architectures, have opened up new possibilities for computer vision missions such as object detection, image generation, and image classification. By encoding global context information from an image and capturing long-range dependencies these models achieve state-of-the-art performance in many benchmark datasets. This article focuses on five prominent image classification methods based on deep learning: Contrastive Captioner (Coca), Pathways Language and Image (PaLI) model, EvoLved Sign Momentum (Lion), A Large-scale Image and Noisy-text embedding (ALIGN), and ONE-PEACE. Each of these methods is briefly introduced and the general principles behind them are explained. Additionally, their practical performance in various applications is also reviewed. Most of these methods involve integrating existing datasets to train the model and focusing on the characteristics of the dataset to improve classification accuracy. This work also analyses the advantages and disadvantages of each method, as well as the functions implemented by their individual designs.

Keywords: Deep Learning, Model Comparison, Large Language Model.

1. Introduction

Image classification aims to holistically understand an image and classify it into a specific category. Unlike object detection, which involves classifying and locating multiple objects within an image, image classification typically applies to images of a single object [1,2]. When classification becomes very detailed or reaches the instance level, it is often referred to as image retrieval, which also involves finding similar images in large databases.

Deep learning models perform well in image classification tasks and have broad application prospects, helping to solve various visual recognition and analysis problems. Therefore, it has very important research value. Its advantages include high accuracy, adaptability to complex data, and multi-modal applications [3]. Deep learning models achieve excellent accuracy on image classification tasks because they learn to extract information about objects, features, and patterns from images, allowing for accurate classification.

Deep learning models can handle large-scale and complex image data, including high-resolution images and large-scale datasets. They are able to learn from this data and capture complex visual information thanks to their effective feature extraction and classification capabilities [4]. For multi-modal applications, deep learning models can be used for tasks that associate images with text, audio,

DOI: 10.54254/2755-2721/45/20241179

and other modalities. This multi-modal capability helps solve more complex problems related to image classification by integrating information from multiple sources.

2. Contrastive Captioner (Coca) Model

CoCa pre-trains an encoder-decoder model leveraging text and image pairs, by optimizing contrastive and subtitle losses, integrating contrastive methods [6]. The diagram in Figure 1 provides CoCa's pretraining approach as an image-text foundation model.

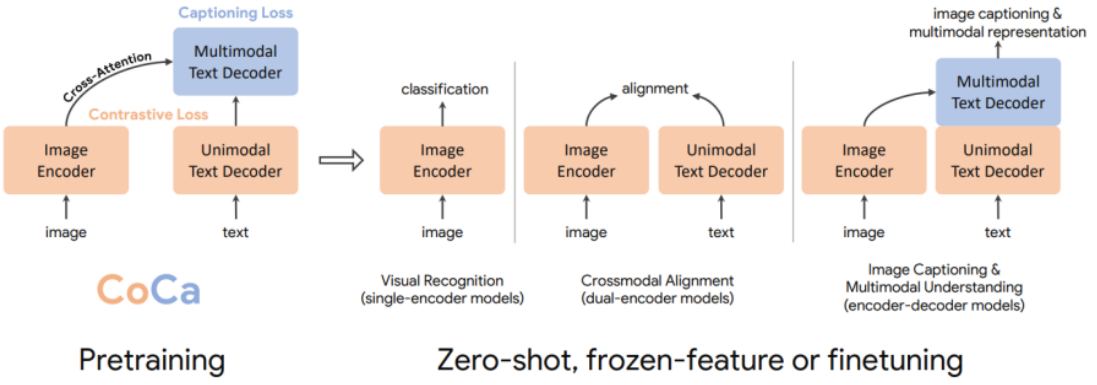


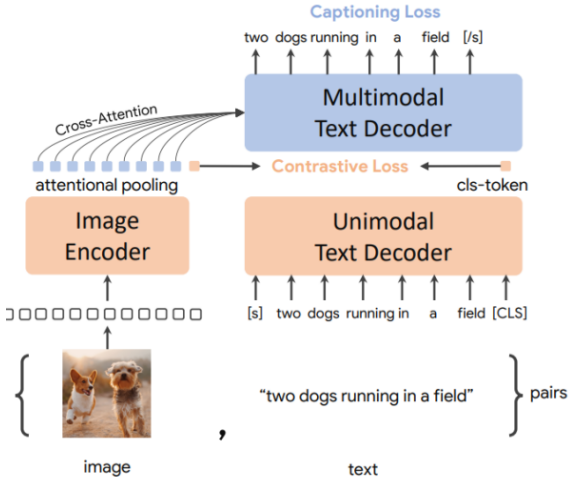
Figure 1. Coca pre-training overview [6].

Previously proposed vision pretraining models have aimed to address visual recognition issues. A common approach is to pretrain ConvNets or Transformers on large-scale annotated data. However, in order to achieve further performance enhancements, vision-language pretraining techniques have been introduced with the objective of jointly encoding vision and language. Early work in this direction extracted visual representations using pre-trained object detection modules. Subsequently, efforts have involved training a multimodal transformer from scratch to unify vision and language transformers.

The CoCa approach is both simple and efficient for training, while also offering additional model capabilities. Firstly, CoCa only requires one forward and backward propagation for a batch of data, while other approaches require two. Secondly, CoCa is trained to optimize two objectives, whereas other approaches are initialized from pre-trained encoders leveraging extra training signals. Thirdly, the decoder enables zero-shot learning and image captioning directly. The CoCa approach has been shown to achieve outstanding performances. Its effectiveness can be attributed to its simplicity, efficiency, and additional model capabilities compared to other vision pretraining techniques.

Figure 2 depicts the operation of the CoCa model, which utilises a simple encoder-decoder framework integrating three distinct training strategies. The model is similar to other image-text encoder-decoder architectures, where CoCa utilises a neural network image encoder, such as a vision transformer (ViT), to encode images into latent representations. However, CoCa can also utilise other image encoders such as ConvNets. The CoCa model then decodes textual data using a causal masking transformer decoder. However, different from conventional decoder transformers, it does not have cross-attention in the first layer of the decoder. The remaining modules use cross-attention with the image encoder to generate multimodal image-text representations. The objective function for CoCa can be expressed as $L_{CoCa} = \lambda_{Con} \cdot L_{Con} + \lambda_{Cap} \cdot L_{Cap}$, where λ_{Con} and λ_{Cap} are loss weighting hyper-parameters. It can be observed that the single-encoder cross-entropy can be seen as a specific instance applied to image annotation data.

DOI: 10.54254/2755-2721/45/20241179



Algorithm 1 Pseudocode of Contrastive Captioners architecture.

```
# image, text.ids, text.labels, text.mask: paired {image, text} data
# con_query: 1 query token for contrastive embedding
# cap_query: N query tokens for captioning embedding
# cls_token_id: a special cls_token_id in vocabulary

def attentional_pooling(features, query):
  out = multihead_attention(features, query)
  return layer_norm(out)

img_feature = vit_encoder(image) # [batch, seq_len, dim]
con_feature = attentional_pooling(img_feature, con_query) # [batch, 1, dim]
cap_feature = attentional_pooling(img_feature, cap_query) # [batch, N, dim]

ids = concat(text.ids, cls_token_id)
mask = concat(text.mask, zeros_like(cls_token_id)) # unpad cls_token_id
txt_embs = embedding_lookup(ids)
unimodal_out = lm_transformers(txt_embs, mask, cross_attn=None)
multimodal_out = lm_transformers(
    unimodal_out[:, :-1, :], mask, cross_attn=cap_feature)
cls_token_feature = layer_norm(unimodal_out)[:, -1:, :] # [batch, 1, dim]

con_loss = contrastive_loss(con_feature, cls_token_feature)
cap_loss = softmax_cross_entropy_loss(
    multimodal_out, labels=text.labels, mask=text.mask)
```

`vit_encoder`: vision transformer based encoder; `lm_transformer`: language-model transformers.

Figure 2. CoCa algorithm and training objectives [6].

This paper proposes a method for pretraining image-text models on large datasets that can transfer to various downstream tasks with little fine-tuning. The resulting CoCa models are more robust to corrupted images. Although current evaluation sets can detect some image corruptions, there are still other vulnerabilities that could be present in both data and model.

CoCa is a new family of image-text models which utilizes existing vision pretraining methods along with natural language supervision. This model combines the contrastive and captioning objectives within one encoder-decoder model, making it efficient for training on image-text pairs from different sources. CoCa has achieved several state-of-the-art performances, and it has proven to be effective.

3. Pathways Language and Image Model (PaLI)

Large language models can excel at many tasks due to their effective scaling and flexible interface. PaLI is a model that takes this approach further by allowing for the joint leraning from language and vision. It allows both image and text input for content generation. With this interface, it can perform various tasks relating to vision, speech, and multimodality in multiple languages [7].

PaLI uses a single "image-and-text-to-text" interface to perform image-only, language-only, and image+language tasks. PaLI is known for its balanced parameter sharing between the language and vision components, with a greater emphasis on the vision backbone. This approach leads to significant improvements in performance. Another crucial aspect of PaLI is its use of extensive unimodal backbones for language and vision modeling, which allows for the transfer of existing capabilities and reduces training costs.

Figure 3 provides an overview of the PaLI model architecture. The core of the model is a text encoder-decoder Transformer, which is capable of encoding and decoding textual data. To incorporate vision as an input, a sequence of visual "tokens" is fed into the text encoder. These tokens are the output of a Vision Transformer. The visual tokens are passed to the model without any pooling being applied to them first.

ViT-e is the largest common ViT architecture in visual components to date. The model with 4B parameters follows the same architecture and training method as the ViT-G model with 1.8B parameters. The only difference between the two models is the application of learning rate cool-down twice. Additionally, the weights are averaged to produce the final output.

DOI: 10.54254/2755-2721/45/20241179

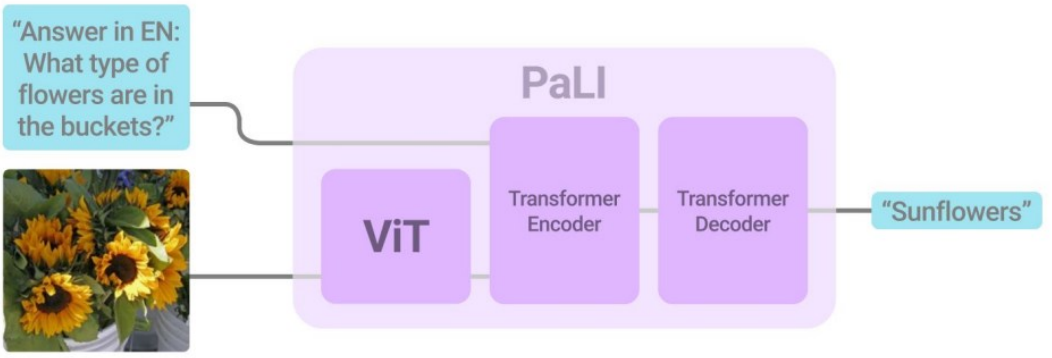


Figure 3. The PaLI main architecture [7].

In image captioning performance, model fine-tuning is first performed using a dataset called COCO Captions, which is widely used for image description tasks. Their model, called PaLI, outperformed state-of-the-art models trained with cross-entropy loss, reaching a new high score of 149.1 as measured by the CIDEr score, a metric that evaluates the quality of image descriptions. Furthermore, there exists an evaluation metric named NoCaps for image description assignments, which is similar to COCO but encompasses a broader range of visual concepts. To evaluate the NoCaps dataset, a model fine-tuned on COCO was employed. PaLI-17B, when tested, garnered a commendable CIDEr score of 124.4 and outperformed other models.

In addition, there are two other related evaluation benchmarks, TextCaps and VizWiz-Cap, which are used to process images containing text and images taken by blind people, respectively.

Regarding some limitations and issues regarding model performance and evaluation, there are limitations in the following areas:

Model performance limitations: Although good model performs well, it has some limitations. For example, when dealing with complex scenes containing many objects, the model may not be able to provide a very detailed description because most of the source data lacks complex annotation information. To alleviate this problem, the model is improved by adding object-aware and location-aware queries.

Multilingual Capability Limitations: It's important to note that fine-tuning language models on only English data may lead to the loss of some multilingual capabilities.

Evaluation Limitations: There are certain limitations concerning how the benchmark is evaluated. For instance, the model might produce the correct answer, but the wording could be a synonym or paraphrase of the target solution rather than an exact match. In such cases, this answer will be considered incorrect. Approaches that use predefined vocabularies sidestep these issues but have limitations when generalizing beyond the answers to a specific dataset.

Bias in evaluation: Comprehensive strategies may be required for unbiased evaluations in benchmarks to avoid bias. Multilingual benchmarks are the first steps towards solving this issue.

4. EvoLved Sign Momentum (Lion)

Lion is a simple and effective optimization algorithm. Unlike adaptive optimizers such as Adam, Lion is more memory efficient because it only tracks momentum. Lion stands out from other adaptive algorithms as it only keeps track of momentum and uses symbolic operations to calculate updates. This approach helps reduce memory usage and ensures consistent update magnitudes across all dimensions. Its parameter updates are computed via sign operations, with the same magnitude for each parameter [8]. When compared to popular optimizers like Adam and Adafactor in image classification tasks, Lion has been shown to improve the accuracy of ViT on ImageNet by up to 2%. Additionally, it can save up to 5 times the amount of pre-training computation on JFT resources. In visual-linguistic contrastive

DOI: 10.54254/2755-2721/45/20241179

learning, The model achieved an accuracy of 88.3% in zero-shot learning and 91.1% in fine-tuning on ImageNet, which is 2% and 0.1% better than the previous best results, respectively. In terms of diffusion models, Lion performs better than Adam, having achieved better FID scores and reducing training computational resources by up to 2.3 times. Lion's performance is similar to or better than Adam's for autoregressive, masked language modeling, and fine-tuning.

Regarding the performance and limitations of Lion, when we calculate the unified update using the symbolic function, it usually results in a relatively large norm. As a result, Lion needs to adopt a lower learning rate and apply more decoupled weight decay to maintain the desired weight decay strength. However, there is no significant difference between Lion and AdamW when it comes to large-scale language and image-text datasets, as shown in Figure 4. Additionally, Lion has less advantage when using strong data augmentation or small batch sizes (<64) in training.

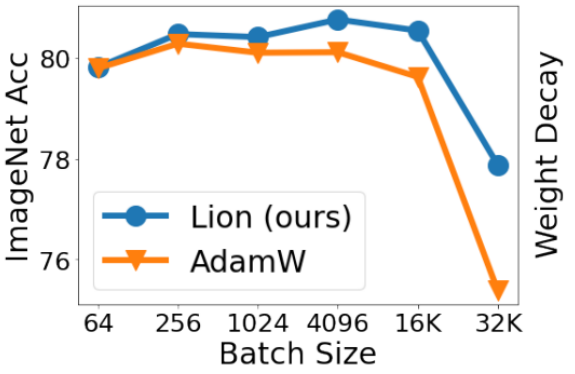


Figure 4. Ablation for the effect of batch size [8].

In the evaluation of the Lion optimization algorithm in different benchmarks, it is mainly compared with AdamW (or Adafactor, when memory is limited), because AdamW is very popular in many learning tasks and is the de facto standard optimization algorithm. This involves fine-tuning each optimizer, including determining the optimal peak learning rate and decoupled weight decay, while leaving the other hyperparameters at Optax's default values. Figure 5 demonstrates that Lion is still the best performing optimizer.

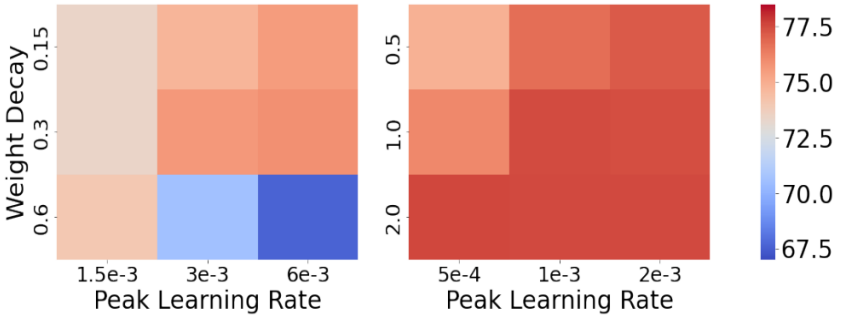


Figure 5. ImageNet accuracy of ViT-B/16 trained from scratch [8].

Here some of the limitations of Lion are listed as follows.

Task selection limitations: While the authors have attempted to assess Lion's performance across a range of tasks, the evaluation is still constrained by a limited set of readings. In the context of vision tasks, the disparity in performance between Lion, AdamW, and Momentum SGD on ResNet is negligible. This could be attributed to the fact that ConvNet is comparatively easier to optimize than Transformer. When using strong data augmentation, the performance improvement brought by Lion is reduced.

DOI: 10.54254/2755-2721/45/20241179

Dataset quality limitations: Some tasks show similar performance between Lion and AdamW, including text-to-image base models, perplexity of autoregressive language models on large-scale in-house datasets, and Masked language modeling on C4. A common feature of these tasks is the large size and high quality of the data sets, which results in reduced differences between different optimizers.

Batch size limit: If the batch size is small, Lion's performance may not be as good as AdamW's (<64). Typically, people increase the batch size to increase parallelism.

Memory overhead limitations: Lion needs to track momentum in bfloat16. It increases the training burden. Factorizing the rate could be a potential solution to save memory.

In summary, Lion may be limited by factors such as data set, data augmentation, batch size, and memory overhead in some cases, resulting in poorer performance than other optimization algorithms. For Lion, it demonstrates superior memory efficiency and generalization performance across diverse domains.

5. A Large-scale Image and Noisy-Text Embedding (ALIGN)

ALIGN is a model that aligns visual and linguistic representations in a shared space to perform well on various tasks [9]. By aligning the representation of images and text, it can achieve strong performance in a variety of tasks. Features of this model include:

Data scale: ALIGN trains on a massive dataset, containing more than a billion samples, eliminating the need for expensive curation and post-processing.

Alignment goal: The main objective of ALIGN is to ensure that the representations of images and text are aligned closely with each other in a shared latent embedding space. This is accomplished via a simple two-encoder architecture and a contrastive loss. The function of the contrastive loss is to bring the embeddings of matching image-text pairs closer and to push the embeddings of mismatched image-text pairs away.

Cross-modal matching: The ALIGN model's alignment representation is ideal for image-text correlation tasks and cross-modal matching and retrieval.

Image representation performance: In addition to its excellent performance in visual-linguistic tasks, the ALIGN model's image representation also excels in downstream visual tasks.

This model enables the joint learning of visual and language, which may contain noise. These representations can be applied to both vision-based and language-based tasks. ALIGN achieves zero-shot visual classification and cross-modal search capabilities without any fine-tuning, enabling several tasks.

The ALIGN model's optimal architecture and hyperparameter configuration can be determined by evaluating its performance in various settings.

Model architecture study: The first step was to evaluate the ALIGN model’s performance leveraging various combinations of image and text backbone networks. During the experiment, different scales of the EfficientNet image encoder and BERT text encoder were tested. Additional fully connected layers were added to match the feature dimensions. Results displayed that as the scale of the backbone network increased, the model's performance gradually improved. It was observed that the image encoder capacity plays a more critical role in vision tasks, whereas both image and text encoder capacity are equally crucial in image-text retrieval tasks. Based on the experimental results, the combination of EfficientNet-L2 and BERT-Large was chosen as the most suitable one.

Research on key architectural hyperparameters: It is essential to study vital architectural hyperparameters of the model, such as the embedding dimensions, the number of negative samples in each batch, and the temperature of the softmax function. Experimental results show that higher embedding dimensions help improve model performance. Reducing the number of negative samples within a batch decreases performance. Different softmax temperature parameter settings were studied and the learnable temperature parameters were found to be competitive in performance and make the learning process easier.

DOI: 10.54254/2755-2721/45/20241179

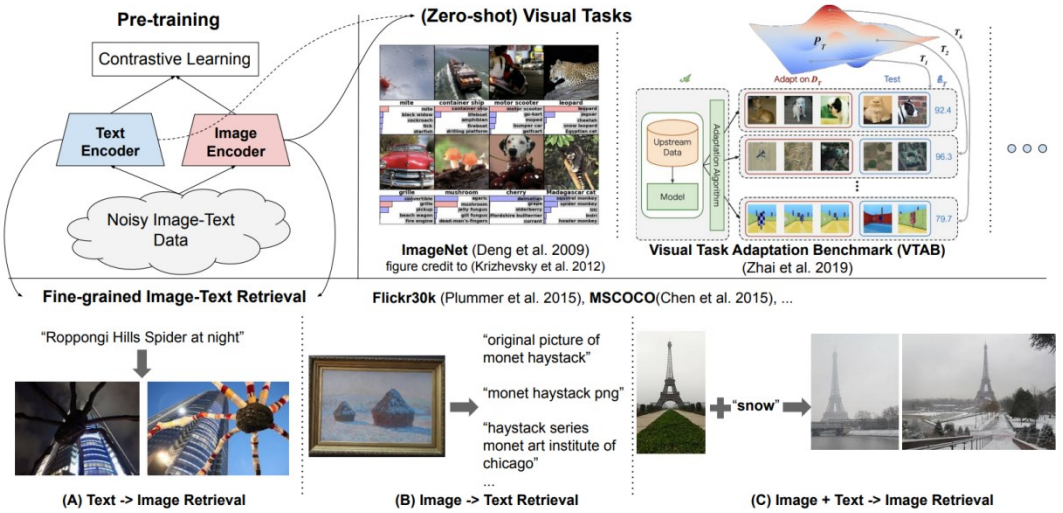


Figure 6. A summary of ALIGN method [9].

The ALIGN model is an exceptional cross-modal retrieval tool that outperforms current state-of-the-art visual semantic embedding (VSE) and cross-attention visual-linguistic models. Moreover, in downstream tasks that involve only vision, the ALIGN model is on par with, or even surpasses, the performance of state-of-the-art models trained using extensive labeled data. This shows that the ALIGN model has strong performance on multiple tasks.

Although this modeling study achieved methodologically promising results, it may still have some implications. For example, consideration needs to be given to the issue that potentially harmful textual data embedded in image descriptions may reinforce undesirable effects. To ensure fairness, it may be necessary to balance the data to avoid reinforcing stereotypes that may be present in network data. In order to prevent the negative impact of mislabeled data, especially when it comes to sensitive religious or cultural items, it is important to perform additional testing and training. The most crucial aspect is to make sure that the model's performance is not influenced by demographic distribution and cultural elements such as clothing, food, and art. If the model is to be used in real applications, it may be necessary to conduct an analysis and balance the data accordingly. Finally, accidental misuse of such models should be prohibited to prevent surveillance or other undesirable purposes.

6. ONE-PEACE Model

The ONE-PEACE model is a general representation model that can handle a variety of different data types, including visual, audio, language and other modalities [10]. With 4 billion parameters, it can seamlessly integrate representations from different modalities. The architecture of ONE-PEACE seems impressive as it includes modal adapters, shared self-attention layers, and modal feedforward neural networks. This design makes it relatively easy to expand new modalities by adding adapters and feedforward neural networks, and the self-attention layers allow multi-modal fusion. Moreover, during the pre-training of ONE-PEACE, the research team developed two modality-agnostic pre-training tasks called cross-modal alignment comparison and intra-modal denoising comparison. Seems like an interesting approach. These tasks can combine the semantics of different modalities. Spatially aligned and capturing detailed information within each modality simultaneously. One-Peace has an easily scalable architecture and pre-training tasks that make it capable of handling various types of data modalities. It has achieved leading results in a several visual tasks. The best part is that it can achieve these results even without using any graphical or language-pre-trained models for initialization.

The ONE-PEACE model architecture is depicted in Figure 7. This model can be decomposed into distinct branches to accommodate a range of different tasks.

DOI: 10.54254/2755-2721/45/20241179

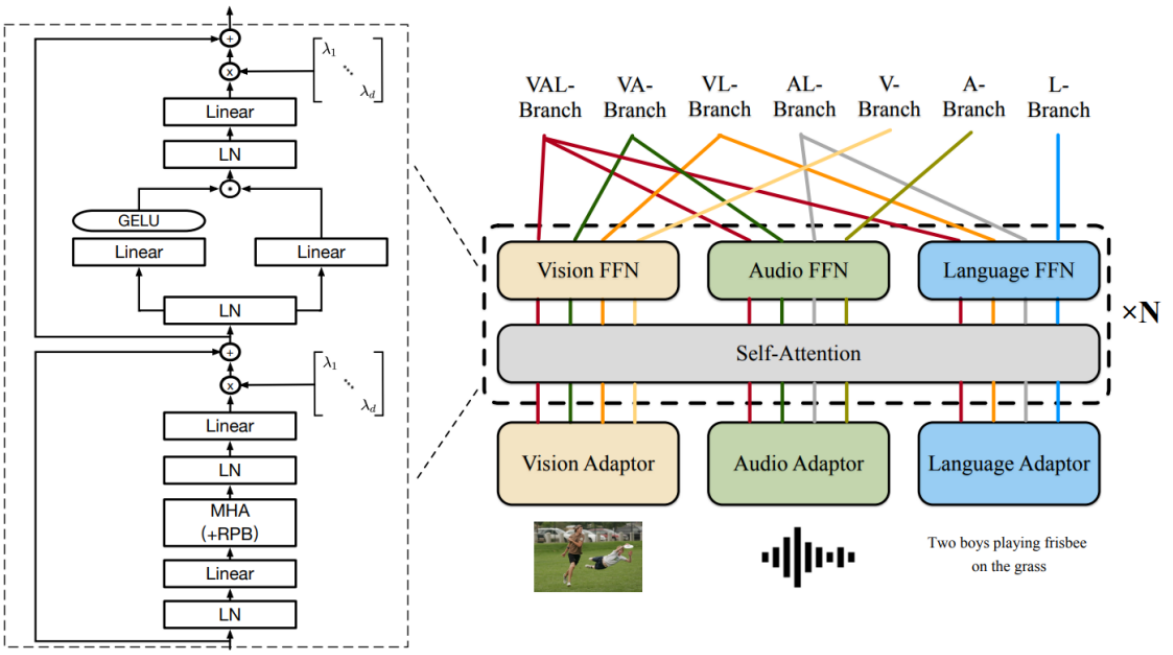


Figure 7. The architecture of ONE-PEACE [10].

Modal adapter: The function of the modal adapter is to convert different original signals into a unified feature representation. There are three different modal adapters, namely visual adapter (V-Adapter), audio adapter (A-Adapter) and language adapter (L-Adapter). Each adapter uses different network structures, such as Transformers, CNNs, RNNs, etc., to process the corresponding modal data.

Modality fusion encoder: The Transformer-based modality fusion encoder consists of a shared self-attention layer and three modality-specific feedforward neural networks (FFNs) that enable modalities to interact through attention mechanisms. The modal feedforward neural network can further extract information within the respective modalities.

This model architecture allows ONE-PEACE to decompose different modalities into different branches that handle various tasks, such as visual branch, audio branch, language branch, etc., to handle various tasks.

ONE-PEACE is a model that can seamlessly integrate representations from different modalities, such as visual, audio, and language, making it a versatile representation model. Experimental results have shown that ONE-PEACE has achieved state-of-the-art performance in various tasks. Additionally, the research has demonstrated that ONE-PEACE has powerful zero-shot retrieval capabilities that enable it to align unpaired modalities.

7. Conclusion

This article is concerned with five leading image classification methods based on deep learning: Contrastive Captioner (Coca), Pathways Language and Image (PaLI) model, EvoLved Sign Momentum (Lion), A Large-scale Image and Noisy-text embedding (ALIGN), and ONE-PEACE. Each of these methods is briefly described, and the main principles behind it are explained. This work also reviews their practical performance in different applications. The majority of these methods involve integrating existing datasets to train the model and focusing on the characteristics of the dataset to improve classification accuracy. This work provides an analysis of the advantages and disadvantages of each method, as well as the specific functions they perform. These models have completely transformed computer vision tasks. They have broken new ground in image classification, object detection, and image generation by encoding information about the overall context of an image and capturing important relationships between elements.

DOI: 10.54254/2755-2721/45/20241179

References

[1] Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. International journal of Remote sensing, 28(5), 823-870.

[2] Wang, W., Yang, Y., Wang, X., Wang, W., & Li, J. (2019). Development of convolutional neural network and its application in image classification: a survey. Optical Engineering, 58(4), 040901-040901.

[3] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[4] Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. IEEE access, 7, 53040-53065.

[5] Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., & Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. Journal of Medical Imaging and Radiation Oncology, 65(5), 545-563.

[6] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., & Wu, Y. (2022). Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917.

[7] Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., et al (2022). Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794.

[8] Jiang, Y., Chan, C., Chen, M., & Wang, W. (2023). Lion: Adversarial Distillation of Closed-Source Large Language Model. arXiv preprint arXiv:2305.12870.

[9] Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., et al. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In International conference on machine learning, 4904-4916.

[10] Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., et al. (2023). ONE-PEACE: Exploring One General Representation Model Toward Unlimited Modalities. arXiv preprint arXiv:2305.11172.