scientific reports

Check for updates

OPEN

Clustering swap prediction for image-text pre-training

Sun Fayou[1,2,3]✉, Hea Choon Ngo[4], Yong Wee Sek[4] & Zuqiang Meng[1]✉

It is essential to delve into the strategy of multimodal model pre-training, which is an obvious impact on downstream tasks. Currently, clustering learning has achieved noteworthy benefits in multiple methods. However, due to the availability of open image-text pairs, it is challenging for multimodal with clustering learning. In this paper, we propose an approach that utilizes clustering swap prediction strategy to learn image-text clustering embedding space by interaction prediction between image and text features. Unlike existing models with clustering learning, our method (Clus) allows for an open number of clusters for web-scale alt-text data. Furthermore, in order to train the image and text encoders efficiently, we introduce distillation learning approach and evaluate the performance of the image-encoder in downstream visual tasks. In addition, Clus is pre-trained end-to-end by using large-scale image-text pairs. Specifically, both text and image serve as ground truth for swap prediction, enabling effective representation learning. Concurrently, extensive experiments demonstrate that Clus achieves state-of-the-art performance on multiple downstream fine-tuning and zero-shot tasks (i.e., Image-Text Retrieval, VQA, NLVR^2, Image Captioning, Object Detection, and Semantic Segmentation).

Keywords Model pre-training, Clustering learning, Swap prediction, Cluster number

The zero-shot approaches are reasonable in practice use. Meanwhile, a lot of methods (e.g., VLMo[1], Coca[2], BEiT-3[3], etc.) proved that image-text fusion can obviously improve the performance of feature representation. In view of this, this paper deeply explores a novel image-text cluster fusion method to achieve progress on a broad range of downstream tasks. As is known to all, the supervised learning methods can only be applied to a range of categories, which are difficult to achieve zero-shot inference[4]. In recent years, multimodal methods can learn image feature representations from image-text pairs, which achieved great success (e.g., image classification[5], object detection[6], semantic segmentation[7], image captioning[60–62], VQA[63], cross-modal retrieval[64], etc.). Currently, due to the diversity of downstream tasks, researchers use different parts of a pre-trained model to adapt to various tasks by building blocks[3,8]. Furthermore, community solutions focus on encoder-only or encoder-decoder manner[9].

CLIP[4] used open-vocabulary as labels with encoder-only manner to unseal an era of zero-shot computer vision tasks, which has inspired many influential researches in the recent past. Currently, CLIP is effective during performing some tasks (e.g., image classification, image-text retrieval, etc.). Meanwhile, the inferiority of CLIP is that modal intersection only uses a simple cosine similarity, which performs poorly in tasks with complex modal interactions (e.g., visual reasoning, etc.). However, cross- modal feature deep fusion is profitable[10]. To solve above problem, researchers proposed lots of methods (e.g., VLMO[1], BEiT-3[3], etc.) that only use a transform encoder for multimodal fusion. Specially, these methods support different experts to handle different types of input (i.e., text, image), known as MoME (mixture of multi-expert). In addition, lots of multimodal models adopt cross-attention method for modal fusion and an encoder produces the final outputs[13,14]. Although these methods are very powerful and achieve promised performance, there are unable to efficiently perform generative tasks (e.g., image captioning).

On the other hand, researchers design encoder-decoder architecture for generative tasks. Notably, the decoder fuses feature, which are from image and text encoders, and the decoder auto-regressively generates feature representation[8,15]. However, some methods (e.g., SimVlm[12], OFA[16], etc.) are less efficient due to the lack of text-only representation of image embedding alignment. To address this issue, CoCa[2] adopts image-text comparative (ITC) loss for cross-modal alignment before multimodal fusion, which obviously improves inference performance. Currently, image-text alignment is popular in most multimodal models. Another typical method is BLIP[11], which utilizes 2 text encoders, 1 text decoder and an image feature encoder to calculate 3 loss function

[1]Guangxi University, Nanning 530004, Guangxi, China. [2]Huzhou College, Huzhou 313000, Zhejiang, China. [3]Linyi University, Linyi 276000, Shandong, China. [4]Universiti Teknikal Malaysia Melaka, 76100 Durian Tunggal, Melaka, Malaysia. ✉email: 314565679@qq.com; 171313540@qq.com

Content courtesy of Springer Nature, terms of use apply. Rights reserved

for different downstream tasks. However, although encoder-decoder manner performs well in generative tasks, it performs inconspicuously in generic tasks[10] (e.g., image classification, object detection, etc.).

Inspired by above approaches, we discuss a question: is it possible to deep mine image-text fusion to learn a high-performance image-text model for most downstream tasks? In order to achieve this objective, this paper utilizes a lot of tricks. In fact, our idea draws on previous research: (1) distillation learning can enhance generalization, (2) image-text alignment can capture the correlation between features, (3) cluster center with explicit semantics, (4) swap prediction is beneficial for the consistency of image and text features. Specifically, we use V-FFN and L-FFN[3] as teachers to train image and text encoders and then carry out image-text alignment by comparative learning manner. Furthermore, this paper utilizes swap prediction method to produce image-text feature cluster prototype, which can evidently improve robustness and performance. Finally, we employ LongNET[20] as a cross-attention module to fuse long-sequence input tokens and generates features for downstream tasks.

All in all, this paper adopts encoder-only architecture and multiple tricks (e.g., distillation learning, cluster learning, swap prediction, etc.) to achieve benefits on lots of downstream tasks. The performance of our method (Clus) is illustrated in Fig. 1. Our contributions are summarized as follows:

1. This study fills the gap of cluster learning for large-scale multimodal model pre-training.
2. Swap prediction is beneficial to improve the explicit semantics of each clustering center.
3. Our method (Clus) achieves the SOTA performance on downstream tasks and proves that the generic large-scale image-text model is useful for practical tasks.

Related work

In this section, we review the existing visual-language approaches, which are related to this research. To overcome the challenging for our objective, this section discusses image and text encoders, image-text fusion and cluster learning.

Image-encoder and text-encoder

Specifically, image and text encoders are pre-trained separately that is a widely used method. As is known to all, image-encoder plays an important role in vision-language (VL) tasks. Currently, the prevalent methods for image-encoder are ViT-based or distillation learning manner. CLIP[4], LSeg[17], CLIPasso[18], ActionCLIP[19], et al. used ViT-based manner to train image-encoder on different datasets. The ablation studies proved that the

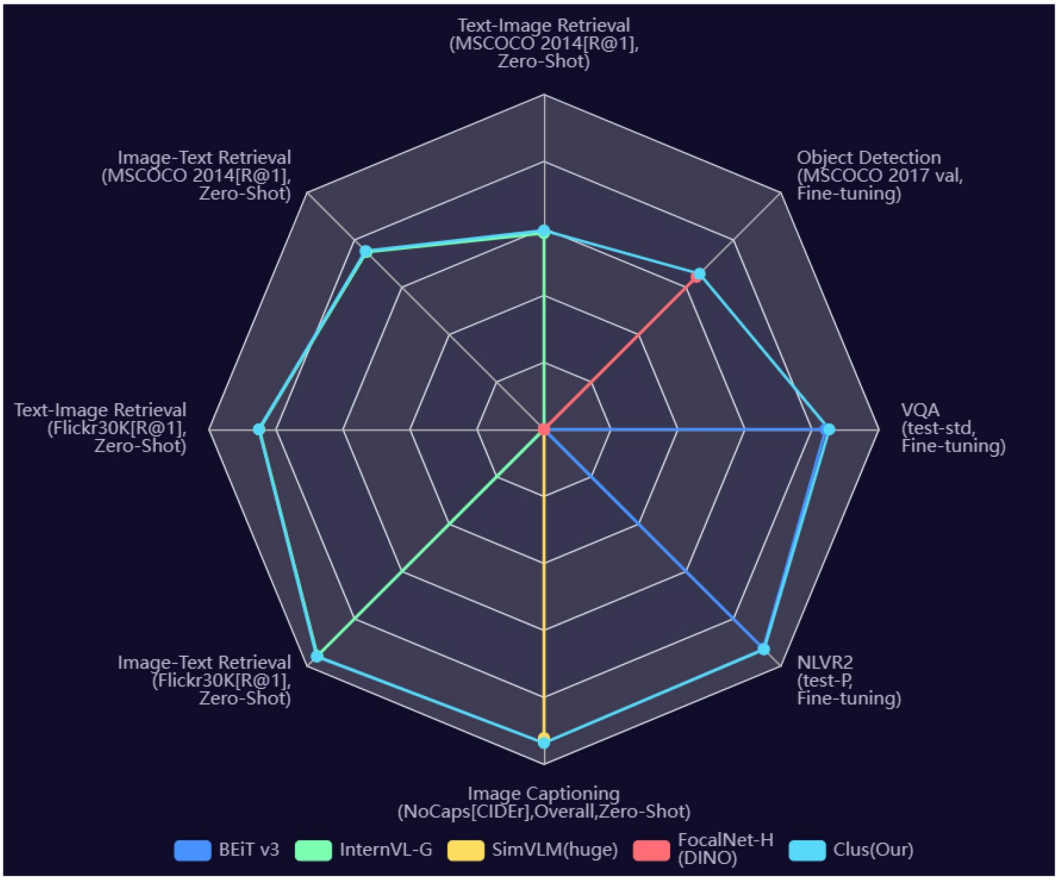

Figure 1. Performance illustration of Clus. Our method achieves the promised results.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

performance of the network has obviously improved when image-encoder employed some tricks. Similarly, ViLD[21] trains image-encoder by distillation learning, which sped up the training and generalization ability of the network. Thus, visual-embed module needs to be a complex network architecture. However, these two manners are independent at present.

Relatively speaking, text-encoder network has a simple architecture and high maturity. Specifically, BERT[22] is a typical method for language model (LM) pre-training and fine-tuning on downstream tasks. At the same time, ALBEF[14] et al. adopt ViT as the text-encoder, which uses 50% of transformer layers as the text-encoder and other layers as the image-text fusion module. This approach improves performance by image-text alignment and multimodal fusion. However, this method requires a lot of computing cost. Recently, VLMo[1] used well pre-trained vision expert (V-FFN) and multi-head self-attention module to train language expert (L-FFN) to achieve SOTA performance.

In view of this, we adopt distillation learning and the parameters of ALBEF[14] for initialization image and text encoders to improve the training efficiency.

Image-text fusion

Intuitively, single modal cannot achieve clear and accurate feature representation. Currently, multimodal feature fusion methods can obtain a more comprehensive and reasonable representation. In order to better utilize the features of various modalities, researchers need to consider the correlation and weight between different modalities. Nowadays, there are four manners for image-text fusion, as shown in Fig. 2.

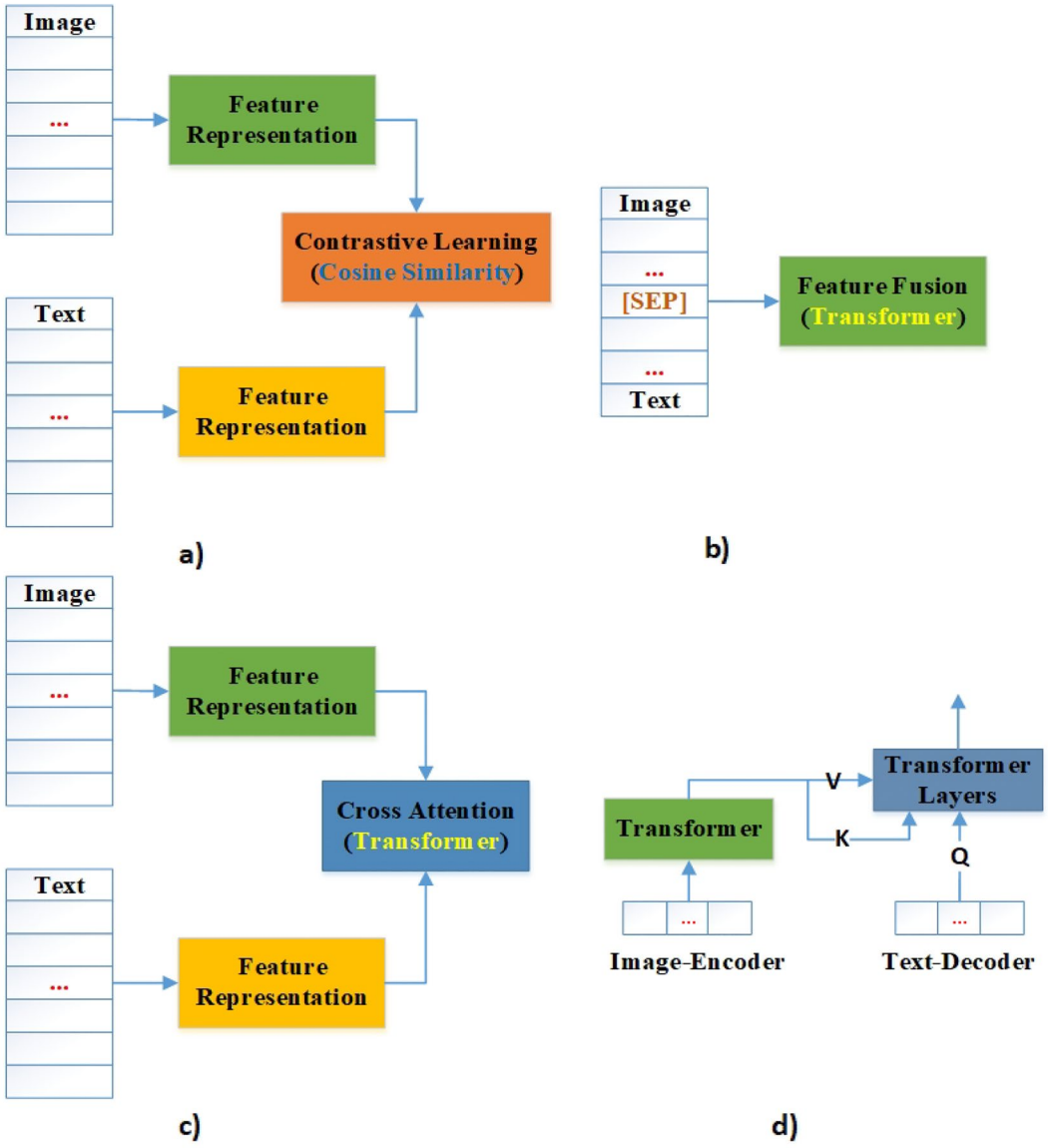

Figure 2. Illustration of the four architecture types.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

The 1st method is similarity-based (Fig. 2a), which is contrastive learning manner. However, due to the simplistic process of modal interaction, these methods are ordinary in performance for difficult tasks (e.g., VQA, Visual Reasoning, etc.). Later, researchers were aware of the important role of modal interaction for multimodal learning. Therefore, the community adopted networks with efficient representation ability to replace cosine similarity for modal interaction. The 2nd method (e.g., VisualBERT[13], UNITER[15], ViLT[23], etc.) is transformer-based multimodal fusion (Fig. 2b), which is single-stream method where the two modalities are concatenated and separated by a special token (e.g., [SEP]). This approach achieves unconstrained multi-modal fusion. The 3rd method (e.g., ALBEF[14], CoCa[2], etc.) is dual-stream (Fig. 2c), where the image and text features are first processed by two independent transformer layers, and then all features are fed into multimodal fusion module (e.g., transformer layers, etc.). Dual-stream method explicitly constrains interactions between modalities and effectively introduces an inductive bias in each model, but it also introduces additional parameters. The 4th method (e.g., BLIP[11]) is based on the encoder-decoder architecture (Fig. 2d), which is a good generative model (e.g., image captioning, etc.). Typically, encoder utilizes multi-head self-attention to fuse inputs from the encoder and decoder. Currently, single-stream methods and dual-stream methods have their own superiority in different downstream tasks. Thus, it is hard to draw a conclusion. In view of this, the dual-stream approach is adopted according to our model architecture in this study.

Cluster learning

Cluster learning is used to construct the meaningful cluster center from unlabeled datasets. Currently, researchers use it to produce better discriminative feature representations.

Currently, cluster Learning methods are widely used in computer vision tasks. ClusDet[24] unifies object cluster and detection. ORE[25] adopts contrastive clustering and unknown-aware proposal network for Object Detection. Specially, Contrastive Learning methods also used cluster Learning. GroupViT[8] integrates grouping blocks in transformer layers as image-encoder for semantic segmentation. SwAV[26] uses "swapped" prediction manner to compare image features. PiCIE[27] uses invariance and equivariance in Clustering for unsupervised semantic segmentation. HAIS[28] introduces the hierarchical aggregation to makes full use of spatial relation of points and point sets for 3D Instance Segmentation. In view of this, this study adopts cluster learning to improve the consistency between modalities.

Method

This section introduces the proposed Clus which contains four modules, i.e., image and text encoders, multi-modal fusion block, image-text clustering, and reasoning.

The architecture of Clus is shown in Fig. 3. From Fig. 3, it can be observed that Clus consist of distillation learning block, co-attention block, clustering swap prediction block, and LongNET block. Specially, ViT is the

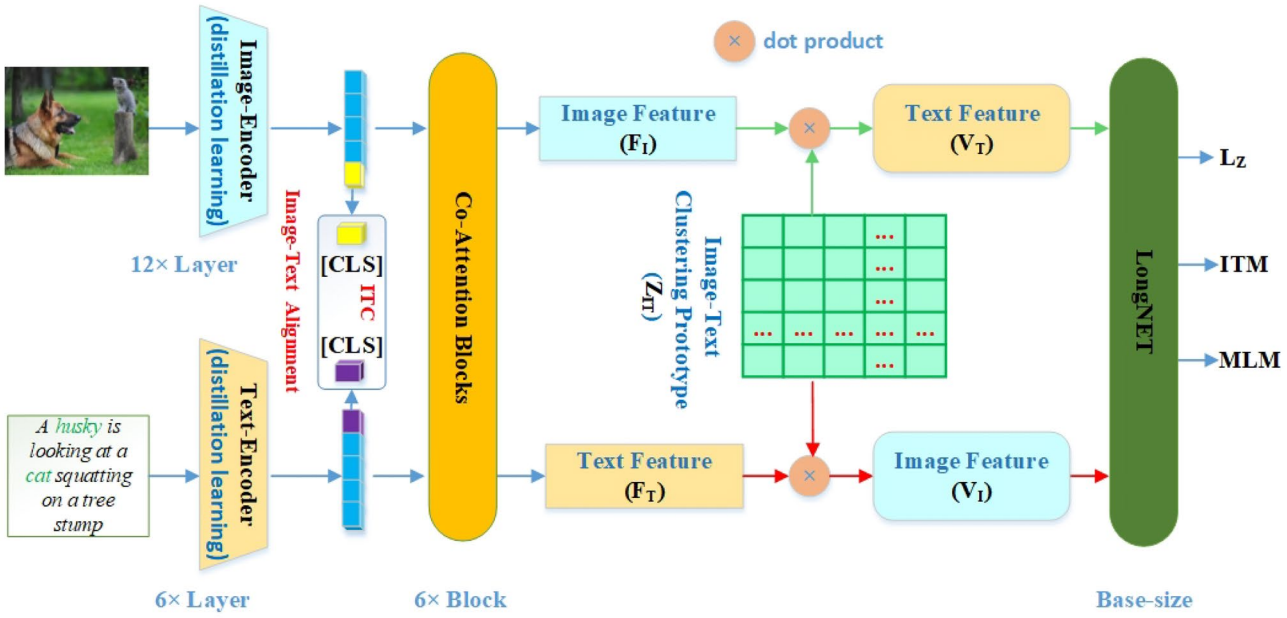

Figure 3. An overview of the Clus. Image and text encoders are distilled sub-networks. Image-text alignment adopts lower-dimensional [CLS] tokens to solve the unimodal representations. There is deep cross-modality middle-fusion with 6 co-attention blocks. The clustering prototype vector (Z_{IT}) is a trainable image-text pair feature vector, which is trained with cross-prediction method. In addition, L_Z is the clustering swap prediction loss. Especially, Clus is trained in an end-to-end way. Finally, the loss function is ITC (image-text contrastive learning loss) + ITM (image-text matching loss) + MLM (masked language modeling loss) + L_Z. In addition, ITM adopts global hard negative mining[3] method.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

backbone of encoders that are trained by distillation learning manner. Then, co-attention blocks are used for image-text feature fusion. Subsequently, we adopt clustering swap prediction method to achieve cross-prediction of image feature (V_I) and text features (V_T). Finally, the large number of feature vectors from the cluster are brought into LongNET[20]. Moreover, ITM loss and LM loss are used to achieve optimal matching and prediction. Note that the image-text alignment is the same as ALBEF[14], where the dimension of [CLS] tokens is $R^{256\times1}$.

Note that BEiT-3[3], VLMo[1], etc., treat images as a foreign language and adopts mask strategy for pre-training. Specially, an image is split into non-overlapping patches and then some patches are randomly masked. Although these methods achieved the SOTA performance at the time, this strategy harms the local neighboring structures especially when discriminative regions are split. To alleviate this issue, this paper employs unmasked images and achieves SOTA performance by strategy of clustering swap prediction. Specially, due to the huge cluster number, we adopt LongNET[20] that scales token length to 1 billion with linear computational cost.

Image and text encoders

As is known to all, knowledge distillation enables the network with novel semantic representations for downstream tasks. Nowadays, V-FFN and L-FFN[3] have good performance for vision and text feature representation respectively in open-vocabulary tasks. Thus, this study uses them as teachers, as shown in Fig. 4. At the same time, in order to improve the training efficiency and performance, the encoders are initialized with the parameters of ALBEF[14]. In a word, distillation learning enables our model (Clus) to be general and energy-efficient. Specifically, this study adopts soft distillation loss. The loss function consists of Kullback–Leibler divergence loss (soft loss) and cross entropy loss (hard loss) are as follows:

$$\textbf{Loss} = \alpha L_{\text{soft}} + (1-\alpha)L_{\text{hard}} \tag{1}$$

where $\alpha = 0.45$.

As shown in Fig. 4b, this study adopts the "Prompt Template" method to produce extra text labels for each image in addition to the original sentence label. In other words, k nouns are randomly selected from a sentence, and every noun word is prompt with a set of handcrafted prompt templates. Specifically, there are many templates for downstream tasks (e.g., object detection, VQA, etc.). The motivation is that objects in images are more likely to be described by nouns, and it is beneficial for supervised labels.

Multimodal fusion

Clus studies two multimodal fusion methods and investigates their performance, as shown in Fig. 5. In the co-attention method, image and text features are fed into different encoders respectively, where each encoder consists of self-attention module, cross-attention module, and one feed-forward module. However, the self-attention method only adopts transformer encoder layer. Concurrently, compared to self-attention method, co-attention method utilizes cross-attention to achieve multimodal interaction, and image and text modalities can be transformation independently. Specially, VOLTA[29] demonstrated that these two methods can achieve comparable performance. Thus, this study uses co-attention method to match dual-stream architecture. Furthermore, our experiments prove that co-attention performs better. In addition, this study designs 6 co-attention blocks so that the number of parameters of these two models are roughly close to each other.

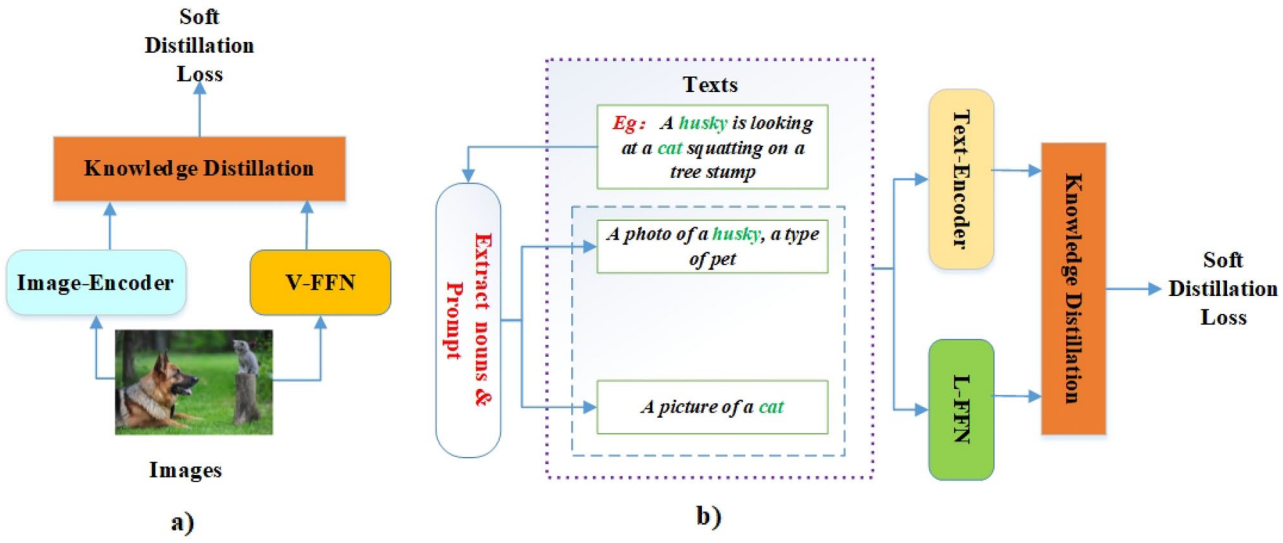

Figure 4. The distillation learning illustration of image-encoder (**a**)) and text-encoder (**b**)). The natural language toolkit (NLTK) is adopted for extracting nouns. Specifically, the encoders are trained together with the network, and the trained encoders are used for reasoning.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

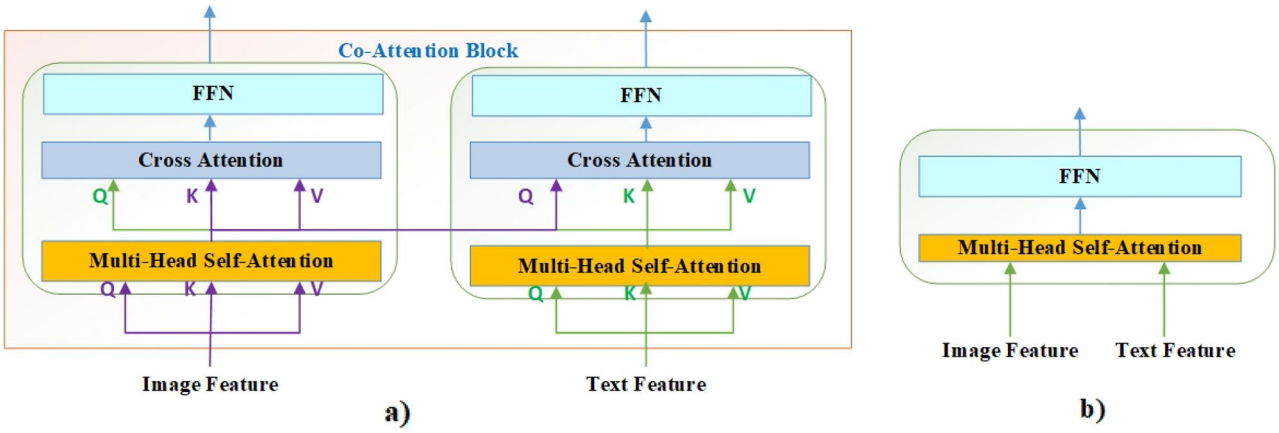

Figure 5. Illustration of two types of multimodal fusion modules: (**a**) co-attention, and (**b**) self-attention.

$$\mathbf{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathrm{softmax}(\frac{\mathrm{QK^T}}{\sqrt{\mathrm{d_k}}}) \tag{2}$$

$$\mathbf{Att_I} = (\mathrm{Q_T}, \mathrm{K_I}, \mathrm{V_I}) \tag{3}$$

$$\mathbf{Att_T} = (\mathrm{Q_I}, \mathrm{K_T}, \mathrm{V_T}) \tag{4}$$

where $\mathrm{Att_I}$ and $\mathrm{Att_T}$ denote image and text cross-attention respectively, as shown in Fig. 5a. From Eqs. (3)–(4), it can be observed that cross modal feature fusion is achieved by Q-vector from different modalities.

Image-text clustering

Currently, clustering method is effective in multimodal pre-training. SOHO[55] and SwAV[26] only performed an online clustering on visual feature maps, which are lack of the representation of image-text feature consistency. TL;DR[56] used K learnable image-text embedding vectors to achieve a small, high-quality set for vision-language pre-training. However, TL; DR[56] is not enough for zero-shot tasks with only K embedding vectors. Meanwhile, SwALIP[57] employed swapped prediction method, yet it used representations from the other modality as prototypes. One objective of this study is to use images and text as supervised labels respectively for feature learning. Thus, we design a multimodal jointly embedding space method by online clustering manner. Simply, this study proposes an online swap prediction method that utilizes the advantages of contrastive learning without comparing the feature of image-text pair. Specifically, Clus boosts the consistency of positive samples for clustering while learning features of the image-text pairs rather than directly comparing features as the contrastive learning. Furthermore, Clus adopts "swap" manner to predict another modality representation from one modality feature and the clustering prototype vector. Specially, we use the DBSCAN[30] method to achieve automatic clustering. Furthermore, the loss function for clustering swap prediction is as follows:

$$\mathbf{L_Z} = \mathrm{L}(\mathrm{P_T}, \mathrm{G_T}) + L(P_\mathrm{I}, \mathrm{G_I}) \tag{5}$$

$$\mathbf{V_T} = dot(F_\mathrm{I}, \mathrm{Z_{IT}}), \mathrm{F_I} \in \mathrm{R}^{i\times \mathrm{d}}, \mathrm{Z_{IT}} \in \mathrm{R}^{\mathrm{d}\times \mathrm{k}} \tag{6}$$

$$\mathbf{V_I} = dot(F_\mathrm{T}, \mathrm{Z_{IT}}), F_\mathrm{T} \in \mathrm{R}^{\mathrm{t}\times \mathrm{d}} \tag{7}$$

$$\mathbf{LongNET_{Input}} = concat(V_\mathrm{T}, \mathrm{V_I})^\mathrm{T} \tag{8}$$

where $L(V_T, G_T)$ and $L(V_I, G_I)$ are cross entropy loss. k is the number of cluster and $\mathrm{Z_{IT}}$ is clustering prototype vector in Eq. (6). The G_T and G_I are the ground truth of text and image respectively. Thus, this study uses $L(V_T, G_T)$ as a case.

$$\mathbf{L}(\mathbf{P_T}, \mathbf{G_T}) = -\sum_\mathrm{k} \mathrm{G_T^{(K)}} \log^{\mathrm{P_T^{(k)}}}, \mathrm{P_T^k} \in \mathrm{softmax}(\mathbf{V_T}) \tag{9}$$

Specifically, Eqs. (6)–(7) represent cross-prediction of different modalities respectively. At the same time, we use the loss function (i.e., Eq. (5)) to achieve the consistency of positive samples.

We argue that the advantages of cluster swap prediction are, (1) improve the performance of the image and text encoders, (2) each clustering center has a clear semantic, which helps to efficiently determine the positive image-text pairs, (3) the size of clustering prototype vector ($\mathrm{Z_{IT}}$) can be dynamically adjusted with downstream tasks, and memory cost is acceptable and not explosive due to the dimension = 2.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Reasoning

Our another objective is to train a model with flexible architecture by end-to-end manner for downstream tasks. The advantage of flexibility can be shown in reasoning. Clus performs monomodal and multimodal downstream tasks, as shown in Fig. 6. For example, Fig. 6c is suitable for carrying out retrieval multimodal tasks. Figure 6d is masked image-text multimodal model for VQA, NLVR^2, etc. Figure 6e masks some words during training and only images are input for reasoning, which is used for generation multimodal tasks (e.g., image captioning, etc.). Specially, our method like stacking blocks to solve architecture inconsistency issue for achieving a unified model.

Computational complexity

The computational complexity of Clus mainly involves three parts, i.e., co-attention, clustering prototype vector, and LongNET. Specifically, Swin Transformer[59] proposed that the computational complexity of vision transformer can be denoted as:

$$\Omega(\mathbf{ViT}) = o(4Ld^2 + 2L^2d) \tag{10}$$

where L is the sequence length and d is the hidden dimension.

Due to the architecture of co-attention is ViT, its computational complexity is consistent with Swing Transformer.

$$\Omega(\mathbf{co-attention}) = \Omega(\mathbf{ViT}) \tag{11}$$

Meanwhile, the computational complexity of the clustering prototype vector is as follow:

$$\Omega(\mathbf{clustering}) = o((i+t) \times d \times k) = o(Ldk) \tag{12}$$

where k is the number of cluster.

In addition, it has been proved in the paper[20] that LongNET can successfully scale up the sequence length with almost constant runtime and save memory. The computational complexity is as follow:

$$\Omega(\mathbf{LongNET}) = o(Lk) \tag{13}$$

Thus, the computational complexity of Clus is as follow:

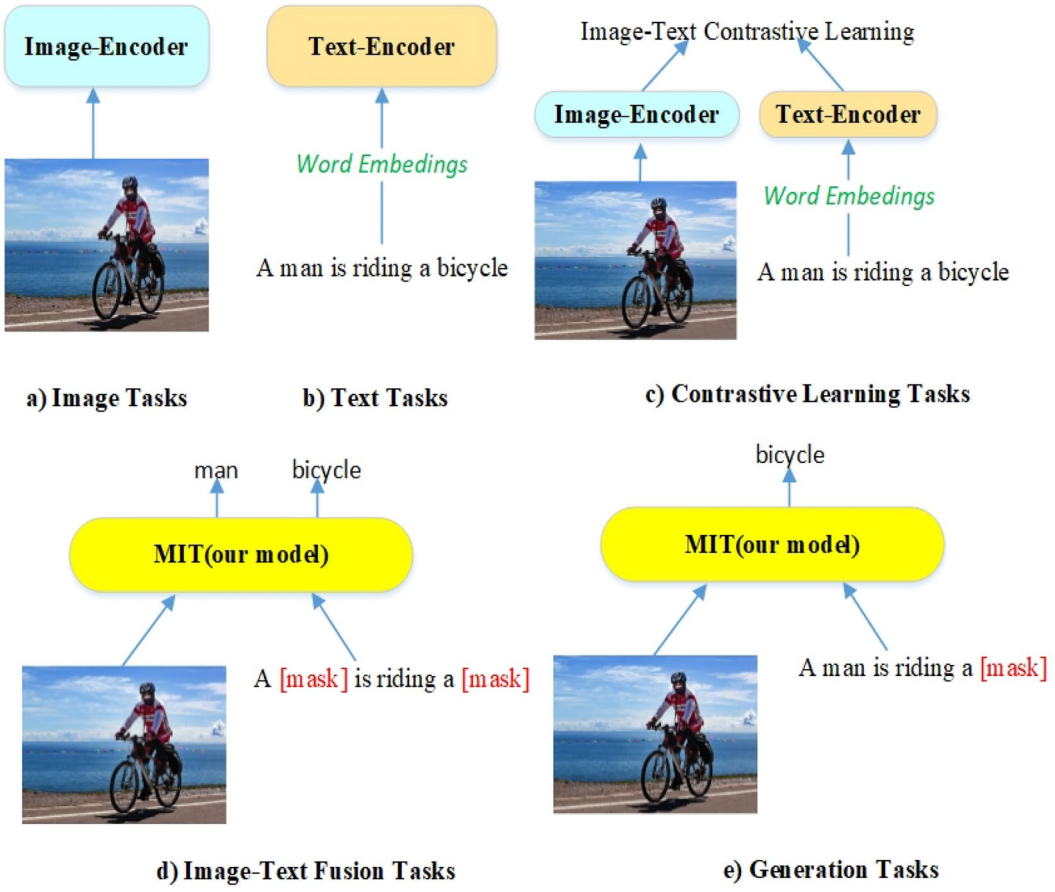

Figure 6. An illustration of Clus (our model) for downstream tasks.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

$$\begin{aligned}\Omega(\textbf{Clus}) &= \Omega(co-attention)+\Omega(clustering)+\Omega(LongNET)\\ &= o(4Ld^2+2L^2d)+o(Ldk)+o(Lk)=o(Lk)\end{aligned} \tag{14}$$

where d is constant.

In view of this, the complexity of Clus is near linear manner and reasonable.

Experiments

Nowadays, BEiT-3[3] has achieved the promised performance, but its image-encoder, text-encoder and visual-language encoder are trained independently. Therefore, inspired by ALBEF[14], CoCa[2], etc., this study utilizes large-scale image-text pairs to pre-train our models with end-to-end manner. Specially, our model will be evaluated on both monomodal and multimodal downstream tasks by fine-tuning or zero-shot manner. In addition, the results of the comparison methods are from the paperswithcode.com (deadline: 8/1/2024). The url of our codes is https://github.com/dlearing/Clus.git.

Pre-training data

In order to achieve end-to-end pre-training, we directly use large-scale image-text pairs to pre-train image-text encoders and our multimodal model. In addition, we adopt widely used web datasets, i.e., CC12M[31], SBU captions[32], COCO[33], Visual Genome[34], LAION-400M[35] and RedCaps-12M[36].

Pre-training settings

Following ViT[37], the resolution of input image is 224×224 and the dimension of token is 1×768. The batch size is 2048 image-text pairs. Furthermore, we adopt AdamW[38] optimizer with $\beta_1=0.9$, $\beta_2=0.999$ and a weight decay of 0.01. Concurrently, the learning rate is warmed-up to the peak value of 1e−4 in the first 5% of training steps with a cosine schedule. In addition, we configure a learnable temperature parameter with an initial value of 0.06 in ITC loss. Specially, DBSCAN[30] is applied to automatically generate the number of cluster centers. Finally, we employed 8 Nvidia A100 GPU 80 GB cards and spent about 6 days for pre-training.

Downstream visual-text tasks

We demonstrate the performance of our method through 4 downstream experiments. Specially, the hyper-parameters of visual question answering (VQA), NLVR^2, and image captioning tasks are shown in Table 1.

Image-text retrieval

There are two sub-tasks, i.e., image-to-text retrieval, and text-to-image retrieval. Likewise, CoCo and Flickr30K[39] datasets are adopted for this task and the Karpathy split[40] is used for both datasets. We conduct fine-tuning our model for 10 epochs with 1024 batch size and the input image resolution is 384×384. In addition, the learning rate reaches peak value 3.5e−5 within the 1st epoch by a cosine schedule. The weight decay is 0.01. The architecture of the modal is shown in Fig. 6C. The experimental results are as follows.

From Fig. 7, it can be observed that our method gains a leading advantage in both fine-tuning and zero-shot tasks. We believe that the strategy of clustering multimodal embedding space improves the performance of the image and text encoders. Meanwhile, the performance of image-to-text retrieval tasks demonstrates that images have richer discriminative features.

VQA and NLVR^2

Our method is suitable for VAQ and NLVR^2 tasks, and the strategy is shown in Fig. 6d. We think that VQA is a classification issue and fine-tune and evaluate on the VQA 2.0 dataset[41]. For the VQA task, image-question pairs are fed into network and a MLP classifier is appended for prediction. Likewise, For NLVR^2, this study uses each triplet (one caption and two images) to construct two image-text pairs as input. The final outputs of the two pairs are concatenated and then fed into a MLP to predict the label.

Hyper-parameters	VQAv2	NLVR^2	Image captioning
Optimizer	AdamW		
Peak learning rate	2e−5	3.5e−5	5e−5
Fine-tuning epochs	8	10	10
Warmup epochs	1	4	1
decay schedule	Cosine schedule		
Weight decay rate	0.01	0.05	0.01
batch size	256	256	128
AdamW β_1	1e−8		
AdamW β_2	0.9,0.995		
Input resolution	256	384	512

Table 1. Hyper-parameters for VQA, NLVR^2, and image captioning tasks.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Figure 7. Results of image-text retrieval on COCO and Flickr30K.

From Fig. 8, it can be observed that our method outperforms BEiT-3[3] by 1.13 points in test-std for VQA. In addition, our model achieves 0.32% gain compared to BEiT-3 in test-P for NLVR[2].

Image captioning

Generative tasks are challenging that relies on the clues of image-text and text-text. Specially, we follow the previous methods (e.g., VLMO[1], UNILM[42]) by a masked fine-tuning manner to learn clues. The strategy is shown in Fig. 6e. Meanwhile, we input image-text pairs into model and randomly mask some caption words for fine-tuning. During reasoning, we only feed images into model to generate the caption tokens in an autoregressive manner. Furthermore, we adopt language model (LM) loss without CIDEr optimization.

From Table 2, it can be observed that our model gets a further 9% improvement compared to BEiT-3 in CIDEr. Furthermore, our model outperforms previous methods in 4 metrics and achieves SOTA performance.

As shown in Table 3, the zero-shot performance of our model is competitive with fine-tuned model as PaLI-17B. Specially, our model outperforms PaLI-17B by 3.2% in in-domian metric and all prior methods in zero-shot. This is because In-domain contains lots of COCO Captioning data. However, Near-domain and Out-of-domain sets have a lot of strange data. Meanwhile, PaLI adopts an encoder-decoder architecture with innate advantage for image captioning. Specifically, the performance of Clus is very close to PaLI.

In this task, we believe that clustering can help construct image-text clues. Concurrently, language model (LM) loss is beneficial for improving the generalization of Generative tasks.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

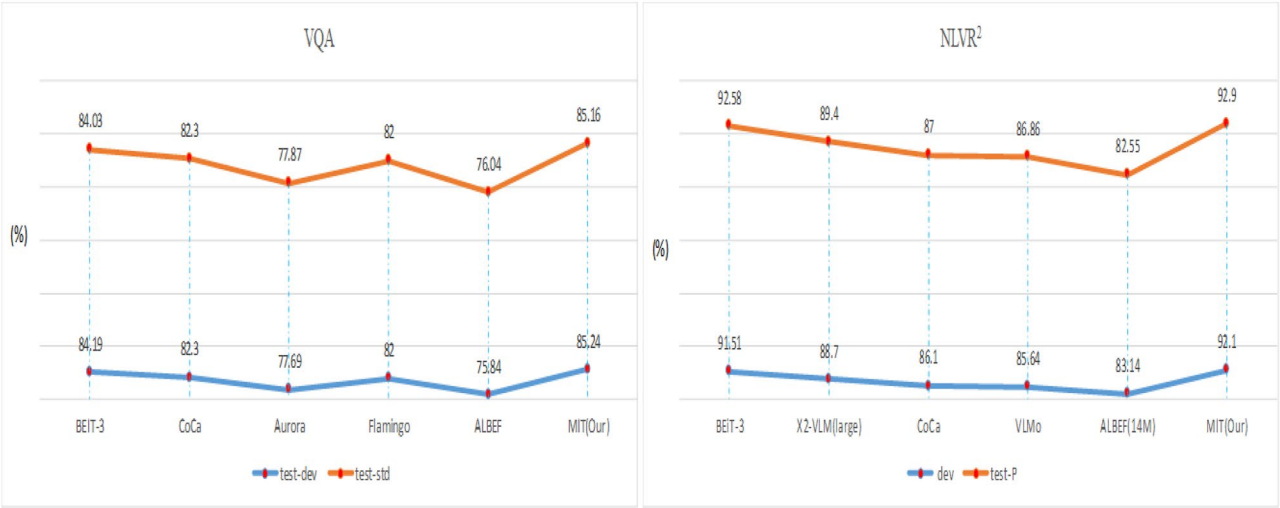

Figure 8. Results of VQA and NLVR^2 tasks.

Model	Spice	Meteor	BLEU-4	CIDEr
mPLUG[43]	26	32	**46.5**	155.1
OFA[16]	**26.6**	32.5	44.9	**154.9**
BEiT-3[3]	25.6	32.4	44.1	147.6
CoCa[2]	24.7	**33.9**	40.9	143.6
Clus(Our)	27.8	34.2	47.2	156.6

Table 2. Results of image captioning on fine-tuned COCO captioning. Significant values are in bold.

Model	In-domain	Near-domain	Out-of-domain	Overall
SimVLM(base)[12]	83.2	84.1	82.5	83.5
SimVLM(huge)[12]	101.2	100.4	102.3	101.4
mPLUG[43]	86.34	81.5	90.49	84.02
BLIP-2[11]	**123.7**	120.2	124.8	121
PaLI-17B(Fine-tuning)[8]	121.1	**124.4**	**126.7**	**124.4**
Clus(Our)	124.3	122.8	125.3	123.6

Table 3. CIDEr results of image captioning on zero-shot NoCaps caption. Significant values are in bold.

Vision downstream tasks

In order to demonstrate the impact of clustering-swap strategy on the performance of image-encoder, we carry on object detection and semantic segmentation experiments.

Object detection

For a fair comparison, we pre-train image-encoder on the Object365[44] and conduct experiments on the COCO2017[33] benchmark. we adopt our image-encoder as backbone and use the strategy of ViTDet[45] for object detection. Likewise, Soft-NMS[46] is employed for reasoning.

From Fig. 9, it can be observed that the AP of image-encoder is improved by 1.3% compared with FocalNet-H (DINO)[47] and is only 0.3% lower than the SOTA supervised model Co-DETR[48]. Ultimately, it proves that clustering swap is beneficial for encoder training.

Semantic segmentation

Generally, natural image pre-trained models are hard to achieve amazing results in medical image segmentation. Notably, the scarcity of public available medical imaging data affects the progress of medical models. In view of this, we transfer the parameters of our image-encoding to SAM-Md3D[49] model and validate the performance on the BraTS2021 dataset. The input resolution is $240 \times 240 \times 155$. Specially, we use AdamW[38] optimizer with

Content courtesy of Springer Nature, terms of use apply. Rights reserved

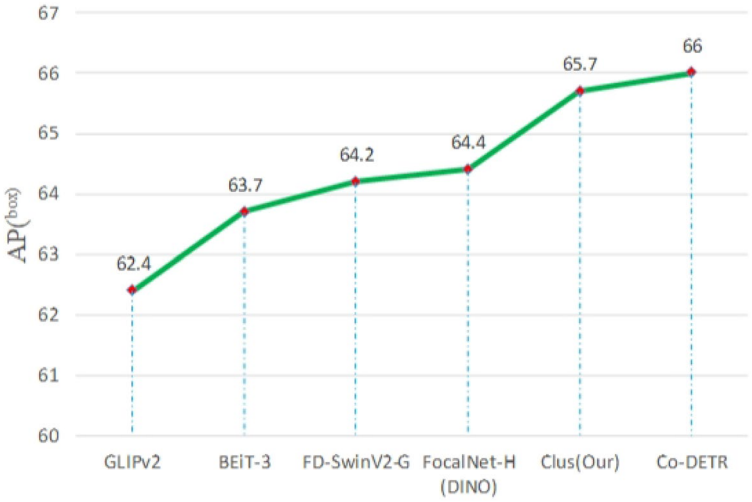

Figure 9. Results of object detection on COCO2017.

$\beta_1 = 0.9$, $\beta_2 = 0.98$ and weight decay is 0.01. The learning rate is warmed-up to 4e−3 within 10 epochs and decayed to 2e−3 following a cosine schedule. In addition, the loss is the combination of dice loss and cross-entropy loss.

From Table 4, it can be observed that the model using the parameters of our image-encoder is significant advantages in 3 tests (i.e., WT, ET, TC). Specially, it brings 3% AVG gains compared to Swin UNETR[53]. Therefore, we believe that natural images pre-trained models are beneficial for improving the performance of other tasks. Notably, Fig. 10 shows the segmentation results of four sequences.

Image-text clusters

Currently, DBSCAN[30] can find any shape clusters based on density. It has two key parameters eps and minPts. Moreover, Erich et al.[58] proposed a method to find minPts based on the 2×dimensionality, and the appropriate value for eps based on the elbow in the k-distance. Since the dimension of clustering prototype vector is 2, the value of minPts is 4. Meanwhile, we use this method to get eps from Fig. 11.

From Fig. 11, it can be observed that the value of eps is 108. Note that Fig. 11 is an enlarged view of the eps part. Therefore, we can obtain the number of clusters.

Notable, the clustering prototype vector (Z_{IT}) embedding space fuses the features of image- text. Our pre-train model Clus with 1.2million cluster number, which is much greater than the cluster number on ImageNet[54]. Due to the large quantity, we are unable to draw the distribution of all clusters by colors. Therefore, we select partial clusters to represent the distribution density by a heat map.

From Fig. 12, it can be observed that the image-text can be effectively fused and the clustering density matches the common sense of daily life. For example, we often see cat-dog, but we rarely see bamboo-frisbee. Therefore, we believe that this clustering is reasonable and valuable.

Memory requirement

Given a lot of clusters and the use of LongNET for scaling token length, it is essential to discuss memory constraints associated with the model. The experimental results are as follow:

From Fig. 13, it can be observed that it is near a linear relationship between memory requirement and scaling token length. Specifically, memory requirement is acceptable. We analyze that this is because the dimension of the clustering prototype vector is 2 and LongNET saves memory[20].

Model	Whole tumor	Enhancing tumor	Tumor core	Avg
nnU-Net[50]	0.929	0.88	0.917	0.909
SegResNet[51]	0.93	0.878	0.912	0.906
TransBTS[52]	0.915	0.867	0.893	0.892
Swin UNETR[53]	0.933	0.891	0.917	0.914
Image-encoder (our)	0.951	0.934	0.946	0.944

Table 4. fivefold cross validation for dice metrics.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

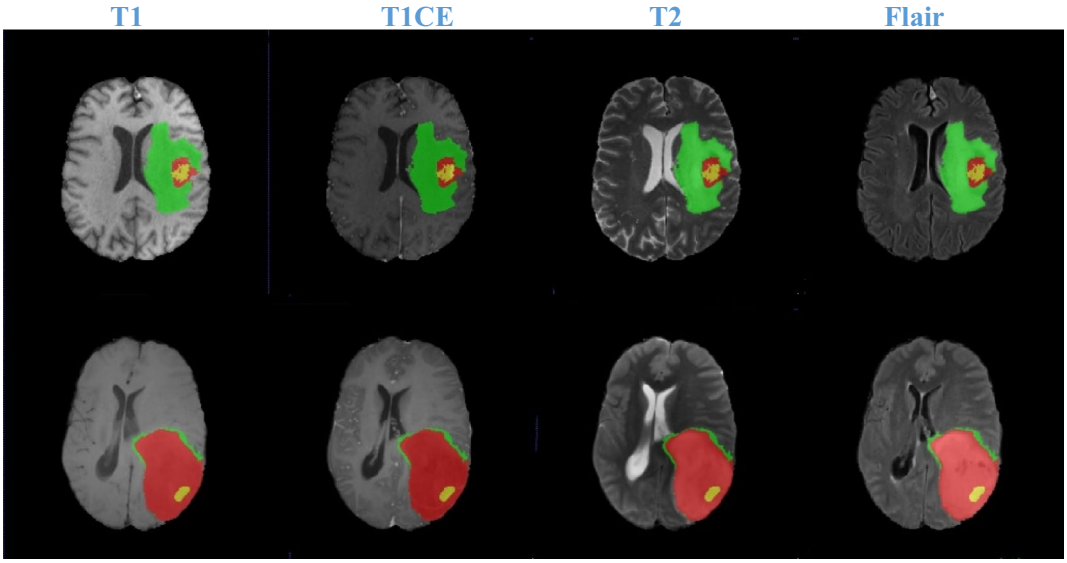

Figure 10. The semantic segmentation results of four sequences for Brain Tumors. The yellow is enhancing tumor regions. Tumor core is composed of yellow and red regions. The green, yellow and red consist of whole tumor.

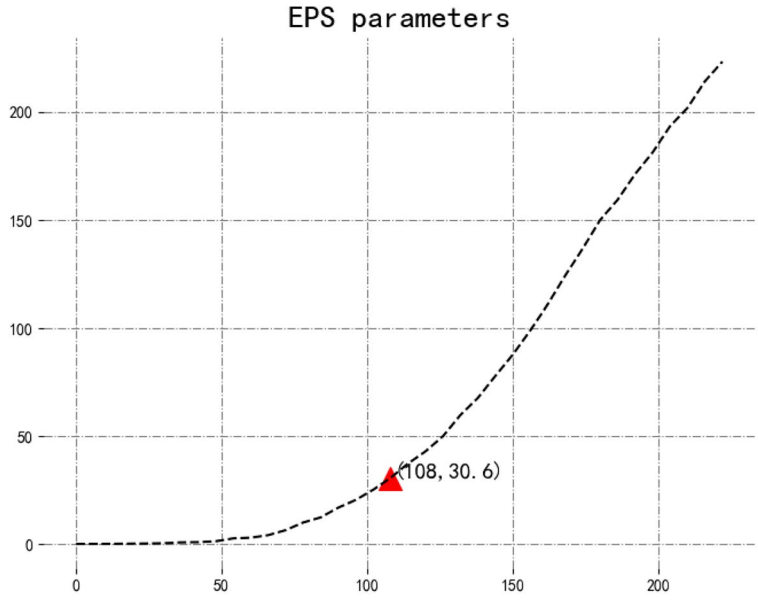

Figure 11. Sorted k-distance plots on pre-training data.

Ablation studies

In order to verify the contribution of different components to the overall performance, we conduct detailed ablation studies. Due to the similar results on any dataset, our method is evaluated on object detection (i.e., COCO2017).

From Table 5, it can be observed that each component is beneficial for performance. Firstly, if we directly adopt image and text encoders from ALBEF, performance is reduced by 0.6. Thus, it is essential to select a good teacher network. Secondly, the co-attention mechanism is helpful, which is consistent with other methods (e.g., Coca, BEiT, etc.). Thirdly, when this study uses generic transformer layers (i.e., dim = $R^{197\times768}$), the result decreases by 1.4. This demonstrates that an increase in the number of active features can lead to a remarkable improvement in revenue. Similarly, if we give up clustering swap method, the ablation result is worse than replacing LongNET. We believe that the clustering center of image-text can better achieve feature semantic representation. Finally, the performance is unsatisfactory when we drop clustering swap prediction and LongNET, which further prove that the key ingredient of Clus is these two components.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

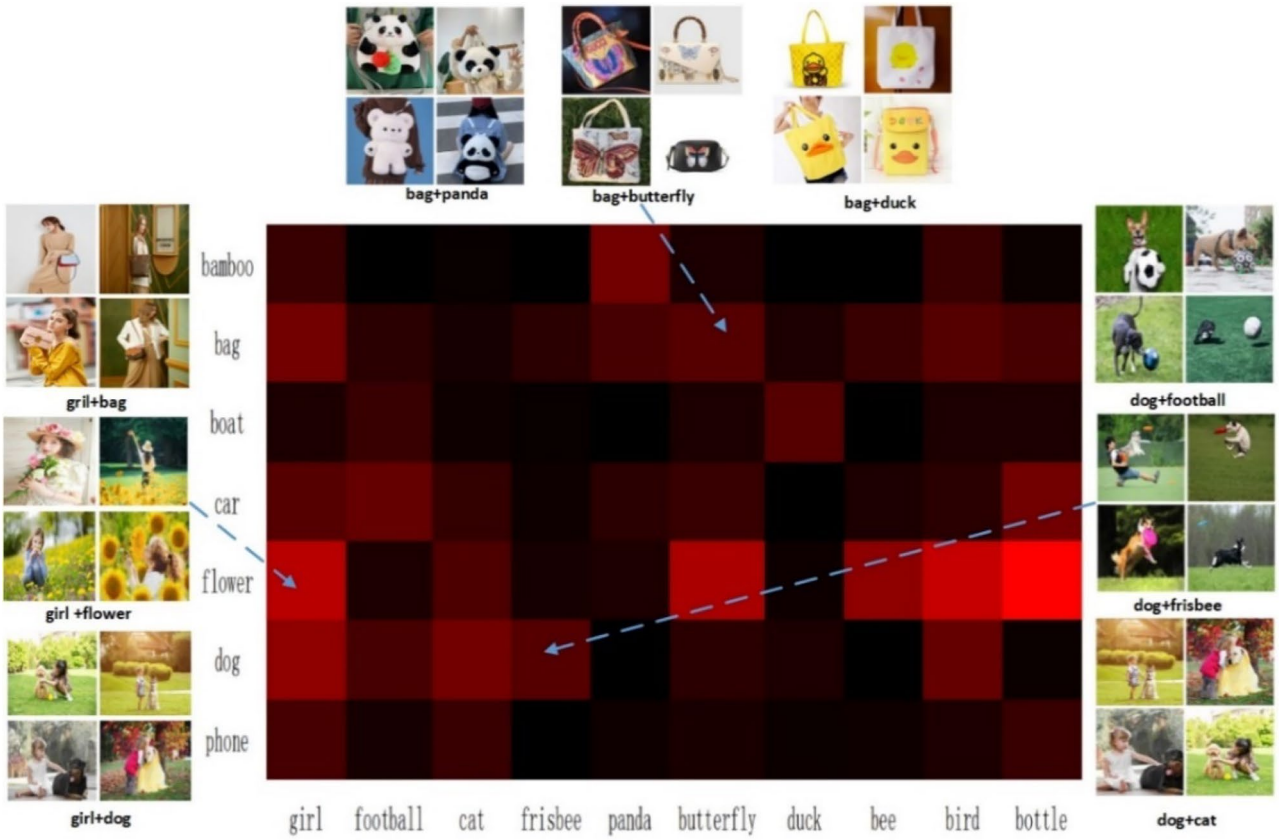

Figure 12. Partial clustering heat map. Red color is strong clustering density.

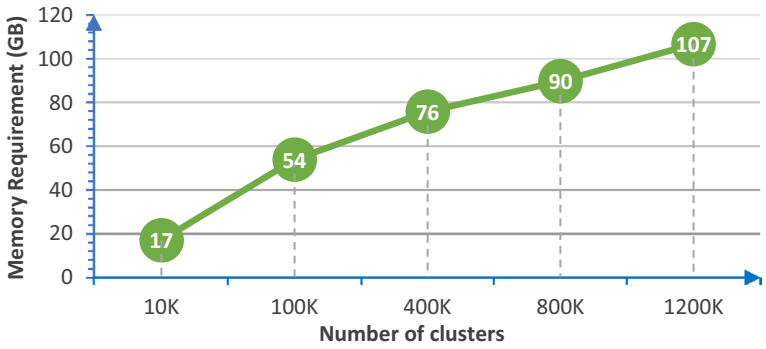

Figure 13. Memory requirement for scaling token length.

Models	COCO2017
Clus (our)	65.7
w/o—Distillation learning	65.1
w/o—Co-attention blocks	65.4
w/o—LongNET(Replace by normal transformer layer)	64.3
w/o—Clustering swap prediction	63.4
w/o—Clustering swap prediction and LongNET	62.2

Table 5. Ablation studies for Clus pre-training on object detection.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Discussion of experimental results

We conduct extensive experiments (i.e., Retrieval, VQA, NLVR2, Image Captioning, Object Detection, and Semantic Segmentation tasks) to verify the Clus. The experimental results demonstrate that the swap prediction method can help cluster centers learn appropriate image-text fusion features. In other words, a bit of improvement is great progress in performance. In addition, due to the difference among the three datasets in Nocaps, Clus only took the lead on the image captioning task in the in-domain dataset. Specifically, due to PaLI-17B[8] with encoder–decoder architecture, it is an innate advantage in image captioning. However, Clus is already close to PaLI-17B[8] in terms of metrics. In the future, we will increase the number of pre-training datasets and the number of clusters to improve the ability of image captioning generation.

Conclusion

In this paper, we propose an image-text clustering swap prediction method to conduct multimodal fusion. At the same time, our model achieves SOTA performance on both downstream visual-text and vision tasks, which demonstrates that the increase of cluster number is beneficial. In particular, our method fills the gap of lack of clustering in multimodal methods and attempts to transfer generic models to practical tasks. Concurrently, Clus is with the same limitations as other methods, such as stopping knowledge updates after pre-training, and generating unfit advice or content, etc. Furthermore, Clus was pre-trained on partial open-source datasets with image-text, but we believe that other modality sets (e.g., audio, video, etc.) are very helpful for improving performance. Currently, due to extensive pre-trained data, Clus may generate offensive, biased content or be used maliciously. In order to mitigate negative impacts, we are attempting to address these issues by fine-tuning. Eventually, we attempt to find a balance between security and practicality.

In the following work, we will actively explore the application of generic models in industrial, electric power, medical, and other fields.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper. In addition, the data can be provided by corresponding authors.

Received: 25 January 2024; Accepted: 28 April 2024
Published online: 24 May 2024

References

1. Wenhui, W., Hangbo, B., Li, D., & Furu, W. VLMo: Unified vision-language pre-training with mixture-of-modality-experts, computing research repository. abs:2111.02358 (2021).
2. Jiahui, Y., Zirui, W., Vijay, V., Legg, Y., Mojtaba, S., & Yonghui, W. CoCa: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.* (2022).
3. Wenhui, W., Hangbo, B., Li, D., Johan, B., Zhiliang, P., Qiang, L., Kriti, A., Owais Khan, M., Saksham, S., Subhojit, S., & Furu, W. Image as a foreign language: BEIT pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19175–19186 (2023).
4. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 139: 8748–8763(2021).
5. Hangbo, B., Li, D., Songhao, P., & Furu, W. BEiT: BERT pre-training of image transformers. In *International Conference on Learning Representations*, abs/2106.08254 (2022).
6. Liunian Harold, L. *et al.* Grounded language-image pre-training. *Proceedings—IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1**, 10955–10965 (2022).
7. Jiarui, X. *et al.* GroupViT: Semantic segmentation emerges from text supervision. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1**, 18113–18123 (2022).
8. Xi, C., Xiao, W., Soravit, C., AJ, P., Piotr, P., Daniel, S., Sebastian, G., Adam, G., Basil, M., Lucas, B., Alexander, K., Joan, P., Nan, D., Keran, R., Hassan, A., Gaurav, M., Linting, X., Ashish, T., James, B., Weicheng, K., Mojtaba, S., Chao, J., Burcu Karagol, A., Carlos, R., Andreas, S., Anelia, A., Xiaohua, Z., Neil, H., & Radu, S. PaLI: A jointly-scaled multilingual language-image model, ICLR 2023 (2022).
9. Zi-Yi, D. *et al.* An empirical study of training end-to-end vision-and-language transformers. *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **1**, 18145–18155 (2022).
10. Chunyuan, L., Zhe, G., Zhengyuan, Y., Jianwei, Y., Linjie, L., Lijuan, W., & Jianfeng, G. Multimodal foundation models: From specialists to general-purpose assistants. CoRR. abs/2309.10020 (2023).
11. Junnan, L., Dongxu, L., Silvio, S., & Steven, H. (2023) BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742.
12. Zirui, W., Jiahui, Y., Adams Wei, Y., Zihang, D., Yulia, T., & Yuan, C. (2022) SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*. abs/2108.10904.
13. Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. VisualBERT: A simple and performant baseline for vision and language. arXiv preprint (2019).
14. Junnan, L., Ramprasaath R., S., Akhilesh Deepak, G., Shafiq, J., Caiming, X., & Steven, H. Align before fuse: Vision and language representation learning with momentum distillation. In *Conference on Neural Information Processing Systems*. abs/2107.07651, 9694–9705 (2021).
15. Chen, Y., Li, L., Yu, L., Kholy Ahmed, E., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. UNITER: Learning universal image-text representations. In *European Conference on Computer Vision*, 104–120 (2019).
16. Peng, W., An, Y., Rui, M., Junyang, L., Shuai, B., Zhikang, L., Jianxin, M., Chang, Z., Jingren, Z., & Hongxia, Y. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, vol. 162 (2022).
17. Boyi, L., Kilian Q., W., Serge, B., Vladlen, K., & René, R. Language-driven semantic segmentation. In *International Conference on Learning Representations* (2022).
18. Yael, V. *et al.* CLIPasso: Semantically-aware object sketching. *ACM Trans. Graphics* **41**(4), 1–11 (2022).
19. Mengmeng, W., Jiazheng, X., & Yong, L. ActionCLIP: A new paradigm for video action recognition. CoRR. abs:2109.08472 (2021).

Content courtesy of Springer Nature, terms of use apply. Rights reserved

20. Jiayu, D., Shuming, M., Li, D., Xingxing, Z., Shaohan, H., Wenhui, W., & Furu, W. LONGNET: Scaling transformers to 1,000,000,000 tokens. arXiv (Cornell University) (2023).
21. Xiuye, G., Tsung-Yi, L., Weicheng, K., & Yin, C. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations* (2022).
22. Jacob, D., Ming-Wei, C., Kenton, L., & Kristina, T. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*. abs/1810.04805, 4171–4186.
23. Wonjae, K., Bokyung, S. & Ildoo, K. Vilt: Vision-and-language transformer without convolution or region supervision. *International Conference on Machine Learning* **139**, 5583–5594 (2021).
24. Fan, Y., Heng, F., Peng, C., Erik, B., & Haibin, L. Clustered object detection in aerial images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019). abs/1904.08008, 8310–8319.
25. Joseph, K. J., Salman, K., Fahad Shahbaz, K., & Vineeth N, B. (2021) Towards open world object detection. In *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. abs/2103.02603, 5830–5840.
26. Mathilde, C. *et al.* Unsupervised learning of visual features by contrasting cluster assignments. *Conference on Neural Information Processing Systems* **33**, 9912–9924 (2020).
27. Jang Hyun, C., Utkarsh, M., Kavita, B., & Bharath, H. (2021) PiCIE: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. abs/2103.17070, 16794–16804.
28. Shaoyu, C., Jiemin, F., Qian, Z., Wenyu, L., & Xinggang, W. Hierarchical aggregation for 3D instance segmentation. In *IEEE International Conference on Computer Vision*. abs/2108.02350, 15447–15456 (2021).
29. Emanuele, B., Ryan, C., Naoaki, O., & Desmond, E. Multimodal pretraining unmasked: Unifying the vision and language BERTs. arXiv preprint arXiv:2011.15124.
30 Michael, H., Matthew, P. & Derek, D. dbscan: Fast density-based clustering with R. *J. Stat. Softw.* **91**(1), 10–300 (2019).
31. Soravit, C., Piyush, S., Nan, D., & Radu, S. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2021). abs/2102.08981, 3558–3568.
32. Vicente, O., Girish, K., & Tamara L., B. Im2Text: Describing images using 1 million captioned photographs. In *Conference on Neural Information Processing Systems*, 1143–1151 (2011).
33. Tsung-Yi, L. *et al.* Microsoft coco: Common objects in context. *Lect. Notes Comput. Sci.* **8693**, 740–755 (2014).
34. Ranjay, K. *et al.* Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **123**(1), 32–73 (2017).
35. Christoph, S., Richard, V., Romain, B., Robert, K., Clayton, M., Aarush, K., Theo, C., Jenia, J., & Aran, K. (2021) LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. CoRR, abs/2111.02114.
36. Karan, D., Gaurav, K., Zubin Trivadi, A., & Justin, J. (2021) RedCaps: Web-curated image-text data created by the people, for the people. *Computing Research Repository*, abs/2111.11431.
37. Alexey, D., Lucas, B., Alexander, K., Dirk, W., Xiaohua, Z., Thomas, U., Mostafa, D., Matthias, M., Georg, H., Sylvain, G., Jakob, U., & Neil, H. (2020) An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, abs/2010.11929.
38. Ilya, L., & Frank, H. (2019) Decoupled weight decay regularization. In *International Conference on Learning Representations*.
39. Bryan, A. P. *et al.* Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IEEE International Conference on Computer Vision* **123**(1), 74–93 (2016).
40. Andrej, K. & Li, F. Deep visual-semantic alignments for generating image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(4), 664–676 (2017).
41. Ning, X., Farley, L., Derek, D., & Asim, K. (2019) Visual entailment: A novel task for fine-grained image understanding. *Computing Research Repository*, abs/1901.06706.
42. Dong,, L., Yang,, N., Wang,, W., Wei,, F., Liu,, X., Wang,, Y., Gao,, J., Zhou,, M., & Hon,, H. (2019) Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, 32, 13042–13054.
43. Chenliang, L., Haiyang, X., Junfeng, T., Wei, W., Ming, Y., Bin, B., Jiabo, Y., Hehong, C., Guohai, X., Zheng, C., Ji, Z., Songfang, H., Fei, H., & Jingren, Z. (2022) mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In *Conference on Empirical Methods in Natural Language Processing*, 7241–7259.
44. Shuai, S. *et al.* Objects365: A large-scale. High-quality dataset for object detection. *IEEE International Conference on Computer Vision* **1**, 8429–8438 (2019).
45. Yanghao, L., Hanzi, M., Ross, G. & Kaiming, H. Exploring plain vision transformer backbones for object detection. *Lect. Notes Comput. Sci.* **13669**, 280–296 (2022).
46. Navaneeth, B., Bharat, S., Rama, C. & Larry, S. Soft-Nms—improving object detection with one line of code. *IEEE International Conference on Computer Vision* **2017**(1), 5562–5570 (2017).
47. Jianwei, Y., Chunyuan, L., Xiyang, D., Lu, Y., & Jianfeng, G. Focal modulation networks. In *Conference on Neural Information Processing Systems* (2022).
48. Zhuofan, Z., Guanglu, S., & Yu, L. DETRs with collaborative hybrid assignments training. dblp, abs/2211.12860 (2023).
49. Haoyu, W., Sizheng, G., Jin, Y., Zhongying, D., Junlong, C., Tianbin, L., Jianpin, C., Yanzhou, S., Ziyan, H., Yiqing, S., Bin, F., Shaoting, Z., Junjun, H., & Yu, Q. SAM-Med3D. CoRR, abs/2310.15161 (2023).
50. Fabian, I., Paul, F. J., Simon, A. A. K., Jens, P. & Klaus, H. M. Nnu-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **18**(2), 203–211 (2021).
51. Andriy, M. 3D MRI brain tumor segmentation using autoencoder regularization. *Lect. Notes Comput. Sci.* **11384**, 311–320 (2019).
52. Wenxuan, W. *et al.* TransBTS: Multimodal brain tumor segmentation using transformer. *Med. Image Comput. Comput.-Assist. Interv.* **12901**, 109–119 (2021).
53. Ali, H., Vishwesh, N., Yucheng, T., Dong, Y., Holger, R., & Daguang, X. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes)*, 12962, 272–284 (2021).
54. Olga, R. *et al.* ImageNet large scale visual recognition challenge. *Int. J. Comput. Vision* **115**(3), 211–252 (2015).
55. Zhicheng, H., Zhaoyang, Z., Yupan, H., Bei, L., Dongmei, F., & Jianlong, F. Seeing out of the box: end-to-end pre-training for vision-language representation learning. In *Proceedings - IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. abs/2104.03135, 12976–12985 (2021).
56. Alex Jinpeng, W., Kevin Qinghong, L., David Junhao, Z., Stan Weixian, L., & Mike Zheng, S. Too large; Data reduction for vision-language pre-training. CoRR, abs/2305.20087: 3124–3134 (2023).
57. Enrico, F., Pietro, A., Adriana, R., Jakob, V., & Michal, D. (2023) Improved baselines for vision-language pre-training. CoRR, abs/2305.08675.
58. Erich, S., Jörg, S., Martin, E., Hans Peter, K. & Xiaowei, X. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42**, 19 (2017).
59. Ze, L., Yutong, L., Yue, C., Han, H., Yixuan, W., Zheng, Z., Stephen, L., & Baining, G. (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE International Conference on Computer Vision*, abs/2103.14030, 9992–10002.

Content courtesy of Springer Nature, terms of use apply. Rights reserved

60. Pengpeng, Z., Jinkuan, Z., Jingkuan, S., & Lianli, G. (2022) Progressive tree-structured prototype network for end-to-end image captioning. In *ACM International Conference on Multimedia*, 5210–5218
61. Haonan, Z. *et al.* SPT: Spatial pyramid transformer for image captioning. *IEEE Trans. Circuits Syst. Video Technol.* **99**, 1–1 (2023).
62. Zeng, P., Zhang, H., Song, J., & Gao, L. (2022). S2 transformer for image captioning. In *IJCAI*, 1608–1614.
63. Haonan, Z. *et al.* Learning visual question answering on controlled semantic noisy labels. *Pattern Recognit.* **138**, 109339–109339 (2023).
64. Hao, L., Jingkuan, S., Lianli, G., Pengpeng, Z., Haonan, Z., & Gongfu, L. (2022) A differentiable semantic metric approximation in probabilistic embedding for cross-modal retrieval. In *Conference on Neural Information Processing Systems*.

Author contributions

Conceptualization: S.F., and Z.M.. Methodology: S.F., and H.C.N. Software and Validation: all authors. Writing (original draft): S.F., and H.C.N. Writing (review and editing): Y.W.S., and Z.M.

Funding

This work is supported by Huzhou college, Linyi university, Guangxi University and the National Natural Science Foundation of China (No. 62266004).

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.F. or Z.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024

Content courtesy of Springer Nature, terms of use apply. Rights reserved

Terms and Conditions

Springer Nature journal content, brought to you courtesy of Springer Nature Customer Service Center GmbH (“Springer Nature”).
Springer Nature supports a reasonable amount of sharing of research papers by authors, subscribers and authorised users (“Users”), for small-scale personal, non-commercial use provided that all copyright, trade and service marks and other proprietary notices are maintained. By accessing, sharing, receiving or otherwise using the Springer Nature journal content you agree to these terms of use (“Terms”). For these purposes, Springer Nature considers academic use (by researchers and students) to be non-commercial.
These Terms are supplementary and will apply in addition to any applicable website terms and conditions, a relevant site licence or a personal subscription. These Terms will prevail over any conflict or ambiguity with regards to the relevant terms, a site licence or a personal subscription (to the extent of the conflict or ambiguity only). For Creative Commons-licensed articles, the terms of the Creative Commons license used will apply.
We collect and use personal data to provide access to the Springer Nature journal content. We may also use these personal data internally within ResearchGate and Springer Nature and as agreed share it, in an anonymised way, for purposes of tracking, analysis and reporting. We will not otherwise disclose your personal data outside the ResearchGate or the Springer Nature group of companies unless we have your permission as detailed in the Privacy Policy.
While Users may use the Springer Nature journal content for small scale, personal non-commercial use, it is important to note that Users may not:

1. use such content for the purpose of providing other users with access on a regular or large scale basis or as a means to circumvent access control;
2. use such content where to do so would be considered a criminal or statutory offence in any jurisdiction, or gives rise to civil liability, or is otherwise unlawful;
3. falsely or misleadingly imply or suggest endorsement, approval , sponsorship, or association unless explicitly agreed to by Springer Nature in writing;
4. use bots or other automated methods to access the content or redirect messages
5. override any security feature or exclusionary protocol; or
6. share the content in order to create substitute for Springer Nature products or services or a systematic database of Springer Nature journal content.

In line with the restriction against commercial use, Springer Nature does not permit the creation of a product or service that creates revenue, royalties, rent or income from our content or its inclusion as part of a paid for service or for other commercial gain. Springer Nature journal content cannot be used for inter-library loans and librarians may not upload Springer Nature journal content on a large scale into their, or any other, institutional repository.
These terms of use are reviewed regularly and may be amended at any time. Springer Nature is not obligated to publish any information or content on this website and may remove it or features or functionality at our sole discretion, at any time with or without notice. Springer Nature may revoke this licence to you at any time and remove access to any copies of the Springer Nature journal content which have been saved.
To the fullest extent permitted by law, Springer Nature makes no warranties, representations or guarantees to Users, either express or implied with respect to the Springer nature journal content and all parties disclaim and waive any implied warranties or warranties imposed by law, including merchantability or fitness for any particular purpose.
Please note that these rights do not automatically extend to content, data or other material published by Springer Nature that may be licensed from third parties.
If you would like to use or distribute our Springer Nature journal content to a wider audience or on a regular basis or in any other manner not expressly permitted by these Terms, please contact Springer Nature at

onlineservice@springernature.com