MS-CLIP: MODALITY-SHARED CONTRASTIVE LANGUAGE-IMAGE PRE-TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale multimodal contrastive pretraining has demonstrated great utility to support high performance in a range of downstream tasks by mapping multiple modalities into a shared embedding space. Typically, this has employed separate encoders for each modality. However, recent work suggest that transformers can support learning across multiple modalities and allow knowledge sharing. Inspired by this, we investigate how to build a modality-shared Contrastive Language-Image Pre-training framework (MS-CLIP). More specifically, we question how many parameters of a transformer model can be shared across modalities during contrastive pre-training, and rigorously study architectural design choices that position the proportion of parameters shared along a spectrum. We observe that a mostly unified encoder for vision and language signals outperforms all other variations that separate more parameters. Additionally, we find that light-weight modality-specific parallel adapter modules further improve performance. Experimental results show that the proposed MS-CLIP outperforms vanilla CLIP by 13% relatively in zero-shot ImageNet classification (pre-trained on YFCC100M), while simultaneously supporting a reduction of parameters. In addition, our approach outperforms vanilla CLIP by 1.6 points on a collection of 24 downstream vision tasks. Furthermore, we discover that sharing parameters leads to semantic concepts from different modalities being encoded more closely in the embedding space, facilitating the learning of common semantic structures (e.g., attention patterns) across modalities.

1 INTRODUCTION

Contrastive Language-Image Pre-training (CLIP) has drawn much attention recently in the field of Computer Vision and Natural Language Processing (Jia et al., 2021; Radford et al., 2021), where large-scale image-caption data are leveraged to learn generic vision and language representations through contrastive loss. This allows the learning of open-set visual concepts and imbues the learned visual feature with a robust capability to transfer to diverse vision tasks.

Prior work in this topic often employs separate language and image encoders, despite architectural similarities between the encoders for both modalities. For instance, the original CLIP work (Radford et al., 2021) uses a ViT (Dosovitskiy et al., 2020) based image encoder, and a separate transformer (Vaswani et al., 2017) based language encoder. However, Lu et al. (2021) recently discovered that transformer models pre-trained on language data could generalize well to visual tasks without altering the majority of parameters, suggesting useful patterns and structures may exist across modalities. In addition, shared architectures have been used to achieve state-of-art performance on a variety of vision-language tasks (Zellers et al., 2021; Li et al., 2019; Chen et al., 2019). These observations suggest that a unified encoder for CLIP may potentially be leveraged to realize performance and efficiency gains.

In this paper, we consequently investigate the feasibility of building a modality-shared CLIP (MS-CLIP) architecture, where parameters in vision encoder and text encoder can be shared. Through this framework, we seek answers to the following three questions: (i) In the CLIP training setting, which layers of the encoders for the two modalities should be shared, and which should be modality-specific? (ii) Within each layer, which sub-module should be shared and which should

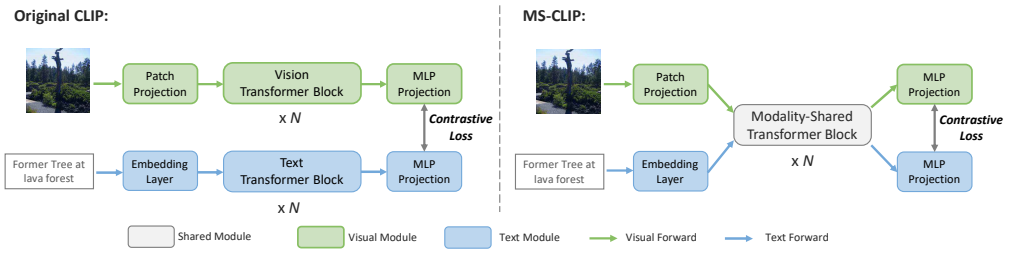

Figure 1: Overview of the original CLIP (left) and our proposed MS-CLIP (right).

not? (iii) Lastly, what is the impact to performance and efficiency when including lightweight modality-specific auxiliary modules to accommodate specializations in each modality?

In order to answer these questions, we first perform a comprehensive analysis on the impact of varying the degree of sharing of components across different layers. Our results show that in order to maximize performance, the input embedding, layer normalization (LN) (Ba et al., 2016), and output projection should be modality-specific. However, all the remaining components can be shared across vision and text transformers, including the weights in self-attention and feed-forward modules. Sharing all these layers even outperforms more complex strategies where we employ greedy selection of layers or use Neural Architecture Search (NAS) (Dong & Yang, 2019) to search for the optimal weight sharing policy.

Finally, we explore whether introducing lightweight modality-specific components to the shared backbone may yield a better balance between cross-modality modeling and specializations within each modality. Studied designs include: (i) Early Specialization. The first layers in vision Transformer and text Transformer are replaced by extra modules that are specialized for each modality, respectively. This includes a set of lightweight cascaded residual convolutional neural networks (CNNs) for vision, and an additional Transformer layer for language. These early layers allow the representations in each modality to lift to higher level patterns before merging, and introduce shift invariance early in the visual branch. (ii) Efficient Parallel Branch. For the visual modality, we explore a lightweight multi-scale CNN network, parallel to the main modality-shared branch, and incorporate its multi-scale features to the main branch through depth-wise convolutional adaptors. This parallel branch enables augmenting the main branch with the benefits convolutions can instill from better modeling of spatial relationships.

We pre-train our MS-CLIP on the major public image-caption dataset YFCC100M (Thomee et al., 2016), and rigorously evaluate on 25 downstream datasets that encompass a broad variety of vision tasks. The experimental results demonstrate that MS-CLIP can out-perform original CLIP with fewer parameters on the majority of tasks, including zero-shot recognition, zero-shot retrieval, and linear probing. Moreover, in order to better understand the success of MS-CLIP, we conduct studies on the learned embedding space, namely with a measurement on multi-modal feature fusion degree (Cao et al., 2020) and quantitatively assess to what degree semantic structures (e.g., attention patterns) are shared across modalities.Our results reveal that sharing parameters can pull semantically-similar concepts from different modalities closer and facilitate the learning of common semantic structures (e.g., attention patterns).

The paper is subsequently organized as follows: in Section 2, we cover datasets and describe the shareable modules and modality-specific designs. In Section 3, we first present a rigorous study varying amount of parameters shared across modalities and measure the impact to downstream performance and efficiency. Then, we measure the impact of modality-specific designs to performance, and compare to model architectures with the adapters absent. Section 4 covers related work, and Section 5 concludes.

2 Methods

2.1 Sharable Modules

Following Radford et al. (2021), we use ViT-B/32 as the basic vision encoder, and the transformer encoder as the basic text encoder, as shown in Fig.1, left. We adjust the hidden dimension of text

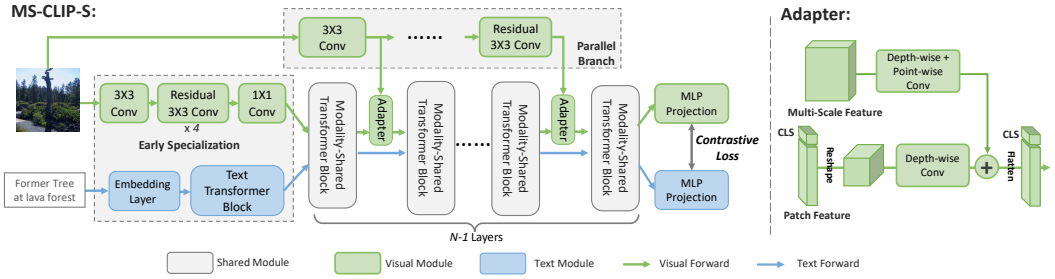

Figure 2: Overview of MS-CLIP-S.

transformer from 512 to 768 to match that in the vision transformer. The resulted additional baseline method is noted as CLIP (ViT-B/32, T768). After the adjustment, the vast majority of parameters between the two encoders can be shared, such as the attention modules, feedforward modules, and LayerNorm (LN) layers. Modules that cannot be shared include the input embedding layer (where the vision encoder deploys a projection layer to embed image patches, while the text encoder encodes word tokens), and the output projection layer. Both encoders have 12 transformer layers. We dub this Naïve modality sharing model MS-CLIP (see Fig. 1, right).

2.2 Modality-Specific Auxiliary Module

In this section we describe the two variations of lightweight modality-specific auxiliary modules used in our study.

Early Specialization In the field of multi-modal learning, it is found beneficial to employ different specialized feature extractors for different modalities and unify them together with the same module in latter layers (Castrejon et al., 2016; Hu & Singh, 2021). Motivated by above, we begin the modality-specific design with making only the first layer specialized for visual and text, leaving other layers shared. Concretely, on vision side, we employ a series of convolutional networks with residual connection as our specialization layer, in which the feature resolution is down-sampled and the channel dimension is increased. The detailed configuration is shown in Tab.1, inspired by a (Xiao et al., 2021). We further add residual connections between convolutional layers, which is empirically more stable for large-scale training. On the language side, we reuse the de-facto Transformer layer for language modeling

Efficient Parallel Branch In image representation, multi-scale information has always been essential (Cai et al., 2016; Szegedy et al., 2015). Earlier work in vanilla vision Transformer (Dosovitskiy et al., 2020), however, operate on a fixed scale. In recent works that introduce multi-scale into ViT (Liu et al., 2021a; Wu et al., 2021), they gradually reduce the patch size and increase the dimension of channel stage by stage. Nevertheless, directly sharing weights between multi-scale ViT and the language Transformer is non-trivial, due to the discrepancy in their channel dimensions. Motivated by Feichtenhofer et al. (2019), we propose to have an auxiliary parallel branch alongside the shared vision Transformer. It consists of one convolution layer and four residual convolution layers, to lower the resolution and widen the channel. Different from plain residual convolution in Early Specialization, here we utilize the bottleneck design in ResNet (He et al., 2016) to save parameters. The main function of parallel branch is to supplement the main branch with multi-scale feature when an image is taken as the input. Therefore, we also employ one adapter after each parallel layer to integrate feature in different scales into different layer of shared Transformer. For efficiency, we adopt depth-wise convolutions (DWConv) and point-wise convolution (PWConv) in adapters to adjust the feature map size and depth. The adapter can be formulated as:

$$\begin{aligned} H_p^{'} &= bn(\text{PWConv}(\text{DWConv}(H_p))) \\ H^{'} &= ln(bn(\text{DWConv}(H)) + H_p^{'}) \end{aligned} \tag{1}$$

where H_p is the multi-scale feature in parallel branch and $H^{'}$ is the adapter's output. bn and ln denote batch normalization and layer normalization. It's noted the *CLS* token is not fused with parallel branch and keeps unchanged. The detailed configuration is provided in Tab.2.

We name the full model with both modality-specific designs as MS-CLIP-S, where "S" indicates supreme (see Fig. 2).

Table 1: Setting of Early Specialization, N*N means 2D kernel size of convs.

Module	Dim	Resolution
3*3 Conv	3→48	224→112
Residual 3*3 Conv	48→96	112→56
Residual 3*3 Conv	96→192	56→28
Residual 3*3 Conv	192→384	28→14
Residual 3*3 Conv	384→768	14→7
1*1 Conv	768→768	7→7
Total # Parameters	4.1M	

Table 2: Setting of Efficient Parallel Branch. Fusion Layer means fusing with which modality-shared layer.

Parallel Module	Adapter Module	Fusion Layer	Resolution
3*3 Conv	16*16 DWConv	2	224→112
Bottleneck 3*3 Conv	8*8 DWConv	4	112→56
Bottleneck 3*3 Conv	4*4 DWConv	6	56→28
Bottleneck 3*3 Conv	2*2 DWConv	8	28→14
Bottleneck 3*3 Conv	1*1 DWConv	10	14→7
Total # Parameters	3.9M		

3 EXPERIMENTS

We start this section by introducing the pre-training and evaluation setup. Then we systematically explore how varying the degree of sharing weights across modalities impacts performance, using the models mentioned in Sec. 2.1 for initial investigation. Further, we validate whether lightweight modality-specific components introduced in Sec. 2.2 could yield a better balance between knowledge sharing and specializations. Comprehensive zero-shot and linear probing evaluations are conducted on a variety of downstream datasets. We conclude this section with probing studies and qualitative results.

3.1 SETUP

Training Details: Similar to the original CLIP paper (Radford et al., 2021), we maintain separate attention masks for image and text: vision transformer allows upper layers to attend to all tokens from lower layers with a bi-directional mask, while the mask in text transformer is auto-regressive. The optimizer is AdamW (Loshchilov & Hutter, 2017). The learning rate is decayed from 1.6e-3 to 1.6e-4, with a cosine scheduler and a warm up at first 5 epochs. We train our models on 16 NVIDIA V100 GPUs with the batch size per GPU set to be 256. For MS-CLIP and MS-CLIP-S, the weight decay for non-shared parameters and shared parameters are separately set to 0.05 and 0.2. We found that a higher weight decay for shared parameters works better, simply because shared parameters are updated twice in each iteration, and a higher weight decay can mitigate over-fitting.

Pretraining Dataset: We use YFCC100M (Thomee et al., 2016) as the pre-training dataset. Following the filtering process in (Radford et al., 2021), we only keep image-text pairs where caption is in English. This leaves us around 22 million data pairs[1].

Evaluation Datasets: In total, we choose 25 public datasets for evaluation: ImageNet (Deng et al., 2009), Food-101 (Bossard et al., 2014), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009), SUN397 (Xiao et al., 2010), Stanford Cars (Krause et al., 2013), FGVC Aircraft (Maji et al., 2013), Pascal Voc 2007 Classification (Everingham et al.), Describable Texture (dtd) (Cimpoi et al., 2014), Oxford-IIIT Pets (Parkhi et al., 2012), Caltech-101 (Fei-Fei et al., 2004), Oxford Flowers 102 (Nilsback & Zisserman, 2008), MNIST (LeCun et al., 1998), Facial Emotion Recognition (Pantic et al., 2005), STL-10 (Coates et al., 2011), GTSRB (Stallkamp et al., 2012), PatchCamelyon (Veeling et al., 2018), UCF101 (Soomro et al., 2012), Hateful Memes (Kiela et al., 2020), Country211 (Radford et al., 2021), EuroSAT (Helber et al., 2019), Kitti-distance (Geiger et al., 2012), Rendered-SST2 (Socher et al., 2013), Resisc45 (Cheng et al., 2017), MSCOCO (Lin

[1]Note that this is more than the 15 million from (Radford et al., 2021) as we use a slightly different English dictionary to exclude non-English words. All our results are reported on this data version, including the baseline (Radford et al., 2021).

Table 3: Experimental results of sharing different components in Transformer layer. LN1 denotes the LN before Attn. LN2 denotes the LN before FFN.

Text Width	# Params	Shared Module	Non-Shared Module	Zero-shot Acc(%)
512	150M	-	Attn, FFN, LN1, LN2	32.15
768	209M	-	Attn, FFN, LN1, LN2	31.85
768	125M	Attn, FFN, LN1, LN2	-	28.40
768	125M	Attn, FFN, LN1	LN2	27.57
768	125M	Attn, FFN, LN2	LN1	32.16
768	125M	Attn, FFN	LN1, LN2	**32.99**

Table 4: Results of sharing different layers in Transformer.

Share Last X layers	12	11	10	8	6	4	2	0	NAS-Search
Zero-shot Acc(%)	32.99	31.25	32.21	32.39	32.85	30.91	nan	31.85	30.97
# Parameters	125M	132M	139M	153M	167M	181M	195M	209M	174M

et al., 2014). These datasets cover various visual scenarios, including generic objects, memes, scenes and etc. We perform linear probing with logistic regression on top of extracted image features, exactly following the protocol in the original CLIP paper (Radford et al., 2021). For zero-shot recognition, we report zero-shot accuracy on ImageNet (Deng et al., 2009) validation set. Following CLIP, we use an ensemble of multiple prompts to extract text features as category features. For zero-shot image-text retrieval, we report recall on MSCOCO (Lin et al., 2014)

3.2 Initial Investigation on MS-CLIP

For validation purposes, we report zero-shot accuracy on ImageNet validation set in our initial study.

1. **LNs need to be modality-specific.** We mainly examine the shareable modules within each Transformer layer, as the input and output projection layers could not be shared. As shown in Tab.3, the first model variant shares all components, including two LN layers and transformation weights in self-attention module and feedforward module, which results in worse performance compared to CLIP (ViT-B/32) and CLIP (ViT-B/32, T768). Then we make the two LN layers modality-specific, which yields better performance and even surpasses the non-shared version in both zero-shot accuracy and parameter efficiency. Noted that the number of parameters in LNs is almost negligible compared with the transformation weights. The sharing is applied in all 12 layers for simplicity. Our observation echos the finding in FPT (Lu et al., 2021) that only tuning LNs in a mostly-frozen pretrained language model yield satisfactory performance on vision tasks.

2. **Less is more: Sharing all layers is better than some.** We further study which layer should be modality-specific and which should be modality-shared. We conduct experiments on sharing last N layers where N is ranging from 12 to 0. $N = 12$ indicates all layers are shared and $N = 0$ indicates the non-shared baseline CLIP (ViT-B/32, T768). Tab. 4 suggests that sharing all 12 layers performs the best while requires the least number of parameters. This sharing-all model is defined as MS-CLIP earlier. Additionally, inspired by recent work on Neural Architecture Search (NAS) (Zheng et al., 2021; Dong & Yang, 2019), we train a model that learns a policy to control which layer to (not) share via Gumbel Softmax (Dong & Yang, 2019). Despite its sophistication, it still underperforms MS-CLIP.

3. **Shared model exhibits higher multi-modal fusion degree.** To probe the multi-modal fusion degree, following (Cao et al., 2020), we measure the Normalized Mutual Information (NMI) between visual features and text features at each layer. For each image-caption pair, we use K-means algorithm (K=2) to group all feature vectors from the forward pass of visual input and text input into 2 clusters. Then, NMI is applied to measure the difference between the generated clusters and ground-truth clusters. The higher the NMI score is, the easier the visual features and text features can be separated, and the lower the multi-modal fusion degree is.

Table 5: Layer-wise NMI scores of models.

Layer	0	1	2	3	4	5	6	7	8	9	10	11	Avg.
CLIP (ViT-B/32, T768)	0.586	0.387	0.265	0.252	0.255	0.241	0.239	0.243	0.235	0.23	0.227	0.185	0.278
MS-CLIP (B/32)	0.589	0.332	0.235	0.211	0.2	0.21	0.2	0.202	0.214	0.197	0.192	0.173	0.246
w/ Early Specialization	0.471	0.348	0.215	0.21	0.218	0.221	0.22	0.213	0.19	0.183	0.179	0.161	**0.235**
MS-CLIP-S (B/32)	0.519	0.536	0.243	0.216	0.199	0.221	0.19	0.247	0.216	0.215	0.224	0.217	0.270

Table 6: Experimental results of zero-shot recognition on ImageNet validation.

Module Name	# Parameters	Zero-shot Acc(%)
CLIP (ViT-B/32)	150M	32.15
CLIP (ViT-B/32, T768)	209M	31.85
MS-CLIP (B/32)	125M	32.99
w/ Early Specialization	129M	35.18
w/ Parallel Branch	129M	34.18
MS-CLIP-S (B/32)	133M	**36.66**

NMI scores are then used to probe the multi-modal fusion degree of the shared model (MS-CLIP (B/32)) vs. non-shared model (CLIP (ViT-B/32, T768)). Here we choose CLIP (ViT-B/32, T768) instead of CLIP (ViT-B/32) in that the feature dimensions of two modalities have to be the same for clustering. NMI scores of all 12 layers and the average are listed in the first two rows of Tab.5. Shared model has lower NMI scores than original CLIP on almost all the layers and the average, indicating a higher degree of multi-modal fusion.

3.3 Experimental Results

Compared Models: We conduct comprehensive experiments with following settings. (1) CLIP (ViT-B/32): The same as Radford et al. (2021), this uses ViT-B32 as visual encoder and Text Transformer as text encoder with width to be 512. (2) CLIP (ViT-B/32, T768): This model sets the width of Text Transformer as 768 to unify the dimension of both encoders. (3) MS-CLIP (B/32): Compared with CLIP (ViT-B/32, T768), this model utilizes the modality-shared transformer blocks to substitute non-shared transformer blocks in visual and text encoders. We use the best setting found in Sec. 3.2: sharing all except for two layer normalizations. (4) MS-CLIP (B/32) + Early Specialization: Based on (3), we specialize the first layer of shared visual&text encoders following Sec. 2. (5) MS-CLIP (B/32) + Parallel Branch: Based on (3), we add a parallel branch to shared visual encoder. (6) MS-CLIP-S (B/32): Based on (3), we apply both early specialization and parallel branch to our shared visual&text encoders.

Zero-Shot ImageNet: The experimental results are reported in Tab.6. In the first row, we reproduce the CLIP (ViT-B/32) pre-trained on YFCC, following the officially released code. On YFCC, Radford et al. (2021) only reported the result of CLIP (ResNet50), which is 31.3% on zero-shot recognition of ImageNet. It proves that our re-implementation can basically re-produce the results reported. By comparing 1-st row and last row, we find MS-CLIP-S (B/32) can outperform CLIP (ViT-B/32) by 4.5% absolutely and 13.9% relatively in zero-shot recognition accuracy on ImageNet, with less parameters.

Ablation Study: In Tab.6, we further analyze the effect of components in MS-CLIP. By comparing 2-nd row and 3-rd row, it is found that directly increasing the text transformer's capacity is useless and even a bit harmful. That is also mentioned in Radford et al. (2021). Then comparing 3-rd row and 4-th row, we find that sharing parameters in vision and text transformer improves the performance and even can outperform CLIP (ViT-B/32) by 0.8%. It demonstrates that sharing the parameters enables the visual and text information to benefit and complement each other. Then we evaluate the proposed auxiliary modality-specific modules one by one. The comparison between 5-th row and 4-th row tells that early specialization can bring 2.1% improvement with only 4M parameters increased. On the other hand, from 6-th row and 5-th row, we realize that auxiliary parallel

Table 7: Results of zero-shot image-text retrieval.

	MSCOCO Val.				MSCOCO Test.			
	I2T		T2I		I2T		T2I	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Vanilla CLIP	23.9	48.8	14.6	34.4	16.3	37.8	14.9	35.7
MS-CLIP-S	**28.9**	**55.1**	**18.7**	**39.8**	**20.0**	**42.5**	**19.3**	**41.7**

branch on vision can also improve by 1.1%. Those two auxiliary modules can work together to further boost the accuracy to 36.66%.

Zero-shot Image-Text Retrieval: We evaluate our MS-CLIP-S on two sub-tasks: image-to-text retrieval and text-to-image retrieval under zero-shot setting. The dataset we used is MSCOCO validation set and test set, where each has 5,000 images. The comparison between MS-CLIP and vanilla CLIP, both pre-trained on YFCC, is shown in Tab. 7.

Table 8: Linear probing results on 24 datasets

Datasets	CLIP (ViT-B32)	MS-CLIP-S (B32)	Δ
Food-101	71.3	**76.0**	+ 4.7
SUN397	68.1	**71.7**	+ 3.6
Stanford Cars	21.8	**27.5**	+ 5.7
FGVC Aircraft	31.8	**32.9**	+ 1.1
Pascal Voc 2007	84.4	**86.1**	+ 1.7
Describable Texture (dtd)	64.1	**69.4**	+ 5.3
Oxford-IIIT Pets	61.1	**62.1**	+ 1.0
Caltech-101	**82.8**	81.6	− 1.2
Oxford Flowers 102	90.7	**93.8**	+ 3.1
MNIST	96.5	**97.2**	+ 0.7
Facial Emotion Recognition	**54.9**	53.6	− 1.3
STL-10	**95.4**	95.1	− 0.3
GTSRB	67.1	**69.9**	+ 2.8
PatchCamelyon	78.3	**81.3**	+ 3.0
UCF101	72.8	**74.6**	+ 1.8
CIFAR-10	**91.0**	87.2	− 3.8
CIFAR-100	**71.9**	66.7	− 5.2
Hateful Memes	50.6	**52.4**	+ 1.8
ImageNet	58.5	**63.7**	+ 5.1
Country211	19.9	**21.9**	+ 2.0
EuroSAT	**94.4**	93.5	− 0.9
Kitti-distance	39.7	**45.1**	+ 5.4
Rendered-SST2	55.2	**56.0**	+ 0.8
Resisc45	83.3	**85.1**	+ 1.8
Avg.	66.9	**68.5**	+ 1.6

Linear Probing: Since we already conduct ablation study under zero-shot recognition, in linear probing, we only compare the CLIP (ViT-B/32) and MS-CLIP-S (B/32). All the results are listed in Tab. 8. Overall, MS-CLIP-S (B/32) outperforms CLIP (ViT-B/32) on 18 out of 24 tasks. The average improvement of 24 tasks in total is 1.62%. The reason behind the improvement of visual encoder might be that, the integration of modality-shared module and modality-specific module enables the visual encoder to benefit from useful language information.

3.4 Further Analysis

NMI Score In Sec. 3.2, we already explain how to measure NMI score and reports the NMI scores of CLIP (ViT-B/32, T768) and MS-CLIP (B/32). We further measure the NMI scores of MS-CLIP (B/32) + Early Specialization and MS-CLIP-S (B/32). The result shows that introducing early specialization can further improve the multi-modal fusion degree. But adding parallel branch leads to a decrease of multi-modal fusion degree. That might be due to the integration of modality-specific multi-scale visual features. From the Tab. 6, adding parallel branch indeed improves the transferable representation, which means NMI score may not be a direct indicator of representation quality. In following subsection, we introduce another metric to analyze the knowledge learnt in MS-CLIPs.

Multi-modal Common Semantic Structure To understand why modality-shared Transformer blocks and proposed auxiliary modality-specific modules can improve the representation, we dig deeper into the what our modules have learnt after training. Our hypothesis is that MS-CLIPs should better capture the common semantic structures existing inside concepts in different modalities. To quantitatively measure it, we probe the attention weights during inference and measure the similarity between attentions in visual and attentions in text. To be more specific, the dataset we use is Flick30K-Entity (Plummer et al., 2015), where there are multiple objects in each image grounded to corresponding concepts in caption. Given an image, assume there are grounded objects (visual

Table 9: Common Semantic Structure distance

Layer	0	1	2	3	4	5	6	7	8	9	10	11	Avg.
CLIP (ViT-B/32)	0.18	0.203	0.227	0.186	0.178	0.164	0.118	0.103	0.106	0.109	0.105	0.074	0.143
MS-CLIP (B/32)	0.175	0.128	0.153	0.132	0.136	0.136	0.106	0.119	0.092	0.106	0.083	0.058	0.113
+ Early Specialization	-	0.107	0.142	0.16	0.12	0.12	0.103	0.103	0.096	0.111	0.11	0.058	0.111
MS-CLIP-S (B/32)	-	0.085	0.162	0.105	0.102	0.103	0.105	0.114	0.093	0.094	0.093	0.061	**0.101**

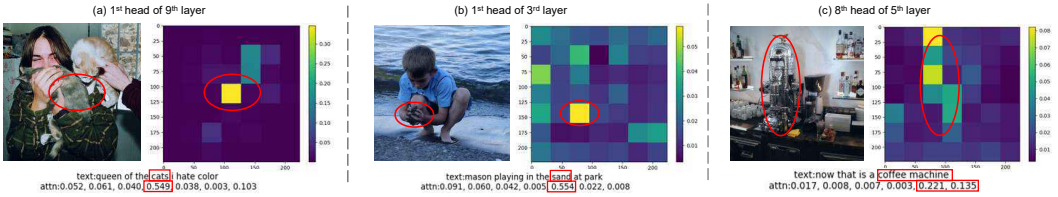

Figure 4: Visualized attention maps of shared attention head.

concepts) $\{vc_1, vc_2, ..., vc_n\}$ and corresponding grounded text concepts $\{tc_1, tc_2, ..., tc_n\}$, in which tc_i refers to vc_i. In the h-th head of l-th attention layer, we take the raw visual attention map M^{lh} and raw text attention map K^{lh}. In order to get the relationship between concepts, we map the text concept tc_i to its last token t_i, and map the visual concept vc_i to its center patch v_i. Through this mapping, we can treat the attention value between tc_i and tc_j as K^{lh}_{ij}, and attention value between vc_i and vc_j as M^{lh}_{ij}. Then for each concept pairs $\{i, j\}$ in both vision and text, we normalize the attention value over starting concept i with softmax function, and average the normalized attention values over all heads in that attention layer. Further, we compute the $l1$ distance between attention values of the same concept pair in different modalities. Finally, we sum the $l1$ distances of all the concept pairs and treat it as the Common Semantic Structure (CSC) distance of that attention layer. A lower CSC distance means more common attention patterns learnt in Transformer across two modalities. The whole process can be formulated as:

$$dis^l_{ij} = |\sum_{h=1}^{H} \frac{1}{H} softmax_i(M^{lh}_{ij}) - \sum_{h=1}^{H} \frac{1}{H} softmax_i(K^{lh}_{ij})| \tag{2}$$

$$CSC^l = dis^l = \sum_{i=1}^{n} \sum_{j=1}^{n} (dis^l_{ij}) \tag{3}$$

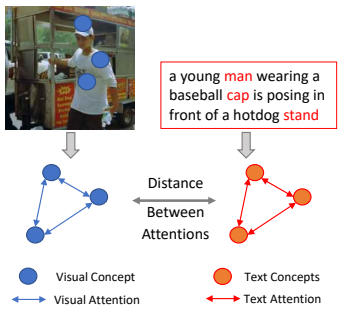

Figure 3: Diagram of computing Common Semantic Structure distance

The layer-wise CSC distance of CLIP (ViT-B/32), MS-CLIP (B/32), MS-CLIP (B/32) + Early Specialization and MS-CLIP-S (B/32) are reported in Tab. 9. It is worth noting we use 10k image-caption pairs from Flick30k-Entity to compute, which is large enough for getting a stable CSC distance. Since the first layer of MS-CLIP (B/32) + Early Specialization and MS-CLIP-S (B/32) doesn't contain attention module in vision branch, we average the last 11 layers' CSC distance to evaluate it. We can find that both the modality-shared Transformer blocks and proposed auxiliary modality-specific modules can lower the CSC distance and learn more semantic structure similarity of vision and text. It is natural that sharing parameters can enforce the attention to learn more common information. As for proposed modality-specific modules, we suspect that those well designed models can account for the discrepancy of separate modalities and make the remaining shared modules focus more on the common patterns.

Visualization of Shared Attention Head In order to intuitively understand how shared attention module works, we visualize the visual attention patterns and text attention patterns of the same shared attention head during inference. More precisely, for vision, we visualize the attention weights between *CLS* token and all patches. For text, we visualize attention weights between *EOS* token and

all other tokens. The reason is that both *CLS* token and *EOS* token will be used as output global feature. The model we use is MS-CLIP-S (B/32). We surprisingly find some heads being able to highlight the same concepts from different modalities. Some samples are visualized in Fig. 3. Take Fig. 3(a) as an example. Given the image and caption respectively as input, the 1st head of 9-th attention layer gives the highest attention value to the region of "cat" in image and token "cats" in text. It validates that the attention heads in MS-CLIP can learn the co-reference between concepts across vision and language.

4 Related Work

4.1 Vision and Language Modelling

This work is built on the recent success of learning visual representation from text supervision. VirTex (Desai & Johnson, 2021) proposes to learn visual encoder from image captioning objectives. LocTex (Liu et al., 2021b) introduces localized textual supervision to guide visual representation learning. Both studies are conducted on a relatively small scale. A more recent work CLIP (Radford et al., 2021) demonstrates that generic multimodal pre-training could benefit from extremely large scale training (i.e., a private dataset with 400 million image-caption pairs) and obtain strong zero-shot capability. It adopts a simple but effective contrastive objective that attracts paired image and caption and repels unpaired ones. ALIGN (Jia et al., 2021) has a similar model design except for using EfficeintNet (Tan & Le, 2019) as their visual encoder, and is pre-trained on an even larger dataset. Our work focuses on the shareability of transformers in vision and text in large-scale contrastive pre-training and are orthogonal to above mentioned works. Another line of work similar to ours is Vision-and-Language Pre-training (or VLP) (Lu et al., 2019; Tan & Bansal, 2019; Zhou et al., 2020; Chen et al., 2019; Li et al., 2019; 2020; Wang et al., 2021a;b), where both vision and language signal are fed into also a unified model to enable downstream multimodal tasks. Our work focuses on learning uni-modal representation instead (i.e., learning visual representation from text supervision) and serves visual-only downstream tasks. Our model could potentially extend to handle multimodal scenarios (Shen et al., 2021) and compare against the VLP counterparts, but is out-of-scope of this paper.

4.2 Parameter-sharing Across Modalities

Humans reason over various modalities simultaneously. Sharing modules for multi-modal processing has attracted increasing interests recently from the community. Lee et al. (2020) proposes to share the parameters of Transformers across both layers and modalities to extremely save parameters. They focuses on video-audio multi-modal downstream task and has an additional multi-modal Transformer for modality fusion. (Hu & Singh, 2021) introduces a shared Transformer decoder for multi-task multi-modal learning. In terms of multimodal fusion, (Nagrani et al., 2021) utilizes a set of shared tokens across different modalities to enable multimodal fusion. The most relevant work to ours is VATT (Akbari et al., 2021). VATT introduces a modality-agnostic transformer that can process video, text, and audio input and is pre-trained on a contrastive objective. The proposed model naively reuses the entire network for all modalities and yields results worse than the non-shared counterpart. We study more than *whether* we can have a shared model, but *how* different degrees of sharing and design nuances behave and *when* we can achieve better performance than non-sharing.

5 Conclusion

We propose MS-CLIP, a modality-shared contrastive language-image pre-training approach, where most parameters in vision and text encoders are shared. To explore how many parameters of a transformer model can be shared across modalities, we carefully investigate various architectural design choices through plenty of experiments. In addition, we propose two modality-specific auxiliary designs: Early Specialization and Auxiliary Parallel Branch. Experiments on both zero-shot and linear probing demonstrate the superior of MS-CLIP over CLIP in both effectiveness and parameter efficiency. Finally, we analyze the reasons behind and realize that sharing parameters can map two modalities into a closer embedding space and promote the common semantic structure learning across modalities.

REFERENCES

Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pp. 354–370. Springer, 2016.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pp. 565–580. Springer, 2020.

Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Learning aligned cross-modal representations from weakly aligned data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2940–2949, 2016.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11162–11173, 2021.

Xuanyi Dong and Yi Yang. Searching for a robust neural architecture in four gpu hours. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1761–1770, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6202–6211, 2019.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3354–3361. IEEE, 2012.

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

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *arXiv preprint arXiv:2005.04790*, 2020.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter efficient multimodal transformers for video representation learning. *arXiv preprint arXiv:2012.04124*, 2020.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021a.

Zhijian Liu, Simon Stent, Jie Li, John Gideon, and Song Han. Loctex: Learning data-efficient visual representations from localized textual supervision. *arXiv preprint arXiv:2108.11950*, 2021b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *arXiv preprint arXiv:2107.00135*, 2021.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pp. 722–729. IEEE, 2008.

Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *2005 IEEE international conference on multimedia and Expo*, pp. 5–pp. IEEE, 2005.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural networks*, 32:323–332, 2012.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pp. 6105–6114. PMLR, 2019.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant CNNs for digital pathology. June 2018.

Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021a.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021b.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. Cvt: Introducing convolutions to vision transformers. *arXiv preprint arXiv:2103.15808*, 2021.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pp. 3485–3492. IEEE, 2010.

Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *arXiv preprint arXiv:2106.14881*, 2021.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*, 2021.

Xiawu Zheng, Rongrong Ji, Yuhang Chen, Qiang Wang, Baochang Zhang, Jie Chen, Qixiang Ye, Feiyue Huang, and Yonghong Tian. Migo-nas: Towards fast and generalizable neural architecture search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):2936–2952, 2021.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13041–13049, 2020.