Learning Visual Representation from Modality-Shared Contrastive Language-Image Pre-training

Haoxuan You[1]⋆, Luowei Zhou[2]⋆, Bin Xiao[2]⋆, Noel Codella[2]⋆,
Yu Cheng[3], Ruochen Xu[2], Shih-Fu Chang[1], and Lu Yuan[2]

[1] Columbia University, New York, USA
[2] Microsoft Cloud and AI, Redmond, USA
[3] Microsoft Research, Redmond, USA
{hy2612,sc250}@columbia.edu
{luozhou,bixi,ncodella,yu.cheng,ruox,luyuan}@microsoft.com

Abstract. Large-scale multi-modal contrastive pre-training has demonstrated great utility to learn transferable features for 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 suggests that transformers can support learning across multiple modalities and allow knowledge sharing. Inspired by this, we investigate a variety of Modality-Shared Contrastive Language-Image Pre-training (MS-CLIP) frameworks. More specifically, we question how many parameters of a transformer model can be shared across modalities during contrastive pre-training, and rigorously examine architectural design choices that position the proportion of parameters shared along a spectrum. In studied conditions, 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 modules further improve performance. Experimental results show that the proposed MS-CLIP approach outperforms vanilla CLIP by up to 13% relative in zero-shot ImageNet classification (pre-trained on YFCC-100M), while simultaneously supporting a reduction of parameters. In addition, our approach outperforms vanilla CLIP by 1.6 points in linear probing 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 transferring of common semantic structure (e.g., attention patterns) from language to vision. Code is available at https://github.com/Hxyou/MSCLIP.

1 Introduction

Contrastive Language-Image Pre-training (CLIP) has drawn much attention recently in the field of Computer Vision and Natural Language Processing [21,47],

---

⋆ Equal Contribution

where large-scale image-caption data are leveraged to learn generic vision representations from language supervision through contrastive loss. This allows the learning of open-set visual concepts and imbues the learned features 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 [47] uses a ViT [13] based image encoder, and a separate transformer [55] based language encoder. However, another work [38] recently discovered that transformer models pre-trained on language data could generalize well to visual tasks without altering the majority of parameters, suggesting patterns learned by one modality could transfer to another. These observations suggest that a unified encoder for CLIP may potentially be leveraged to promote learning commonly useful representations across modalities 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) Within each layer, which sub-module should be shared and which should not? (ii) In the CLIP training setting, which layers of the encoders for the two modalities should be shared, and which should be modality-specific? (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 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) [2], and output projection should be modality-specific. In contrast, all the remaining components can be shared across vision and text transformers, including the weights in self-attention and feed-forward modules. Addtionally, sharing all transformer layers even outperforms more complex strategies where we employ greedy selection of layers or use Neural Architecture Search (NAS) [12] to search for the optimal layer 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 Transformer block is replaced by modules that are specialized for each modality, respectively. This includes a set of lightweight cascaded residual convolutional neural networks (CNNs) for vision, and a Transformer layer for language. This early adaption allows the representation in each modality to abstract to a higher level before unified encoding, and introduces 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 MS-CLIP architectures on YFCC100M [54] and a subset of Laion-400M [48] with a similar size, and evaluate on 25 downstream datasets that encompass a broad variety of vision tasks. The experimental results demonstrate that MS-CLIP architectures, while having fewer parameters, can outperform original CLIP on the majority of tasks, including zero-shot recognition, zero-shot retrieval, and linear probing. Moreover, in order to better understand why MS-CLIP architectures work so well, we conduct studies on the learned embedding space, namely with a measurement on multi-modal feature fusion degree [5], 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. Section 2 covers related work. In Section 3, we introduce the shareable modules and modality-specific designs. In Section 4, we present a rigorous study varying amount of parameters shared across modalities and measure the impact of both modality-shared parameters and modality-specific modules to downstream performance and efficiency. And we comprehensively compare proposed MS-CLIP architectures against CLIP on 25 downstream datasets. Section 5 concludes.

2   Related Work

Learning Visual Representation from Text: Our work is built on the recent success of learning visual representation from text supervision. VirTex [11] proposes to learn visual encoding through an image captioning objective. LocTex [34] introduces localized textual supervision to guide visual representation learning. Both studies are conducted on a relatively small scale. More recent work such as CLIP [47] and ALIGN [21] demonstrate that generic multi-modal pre-training could benefit from extremely large scale training (i.e., private datasets with hundreds of millions or billions of data pairs) and obtain strong zero-shot capability. They adopt a simple yet effective contrastive objective that attracts paired image and caption and repels unpaired ones. There have been several additional works following the line of CLIP/ALIGN [65]. Florence [64] and BASIC [45] scale up the dataset and training with various backbones. FILIP [63] focuses on generalizing the contrastive loss to local tokens for fine-grained supervision. DeCLIP [31], SLIP [40] and other recent works extend supervision signal from self-supervision, multi-view supervision, nearest-neighbor supervision, object detections [67], or external language knowledge [29]. Orthogonal to above mentioned works, this work focuses on the sharing of weights across vision and text modalities in large-scale contrastive pre-training.

Vision and Language Modeling: Another similar line of work is Vision-and-Language Pre-training (or VLP) [36,53,68,6,28,30,58,59,27,57], where both vi-

sion and language signals are fed into also a unified model to enable downstream multi-modal tasks. Moreover, [41] utilizes a set of shared tokens across different modalities to enable multi-modal fusion. But there are two main differences between VLPs and this work: First, in VLP approaches, the model input consists of both image and text modalities concurrently, where the model attends to both modalities at the same time (essentially conducting modality fusion). In CLIP and MS-CLIP, the Transformer's input is either image or text individually: each modality is processed in isolation, where the two modalities are never processed concurrently (except for computing the contrastive loss at the end). Secondly, VLP works focus on designing unified fusion modules to blend multi-modal input well and target at multi-modal tasks (*e.g.*, VQA, grounding), while the goal of our work is to allow parameter and knowledge sharing for uni-modal input and mainly serves visual-only downstream tasks.

Parameter-sharing Across Modalities: As humans reason over various modalities simultaneously, sharing modules for multi-modal processing has attracted increasing interests recently from the research community. [26] proposes to share the parameters of Transformers across both layers and modalities to save parameters. They focus on video-audio multi-modal downstream tasks and have an additional multi-modal Transformer for modality fusion. [37] proposes to train a fully shared Multi-modal Transformer on 12 vision-language datasets. [20] further introduces a shared Transformer decoder for multi-task multi-modal learning. The most relevant work to ours is VATT [1]. 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. In contrast, this work studies more than whether a model can be shared, but rather how various degrees of sharing and design nuances behave, and which of those design choices might be useful to improve performance.

3   Methods

3.1   Sharable Modules

Following [47], we use the Vision Transformer as the basic vision encoder (ViT-B/32 by default), and the transformer encoder as the basic text encoder, as shown in Fig. 1-1. The challenge is to merge these two architectures. To accomplish this, we adjust the hidden dimension of text transformer from 512 to 768 to match that in the vision transformer. The resulting additional baseline method is noted as CLIP (ViT-B/32, T768). After the adjustment, the resulting shared encoder uses 12 layers, with the vast majority of parameters able to be shared between two modalities, including 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.
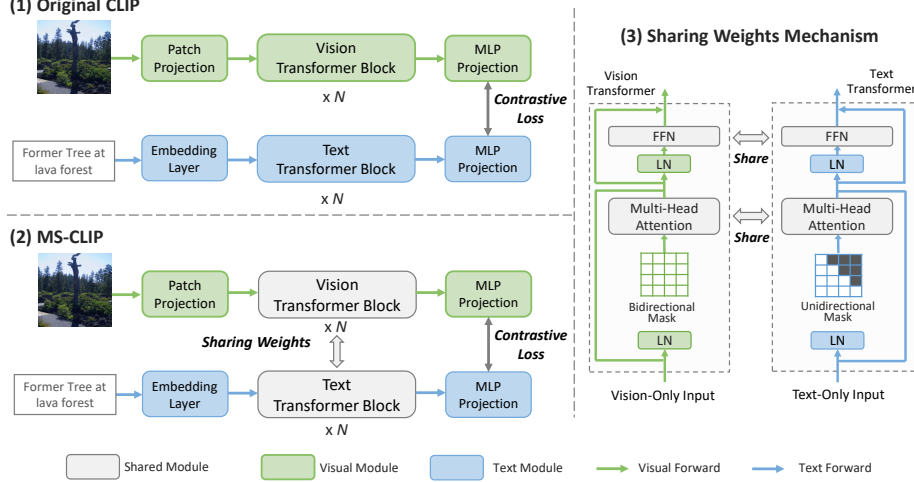
Fig. 1: Overview of the (1) vanilla CLIP and (2) our proposed baseline MS-CLIP, and (3) details of sharing mechanism MS-CLIP.

We performed an experimental analysis to examine the impact of various degrees of weight sharing across modalities (see Sec. 4.3: *On Modality-Shared Components*). In summary, the observations of that study are as follows: (1) LNs need to be modality-specific while the rest can be modality-shared; (2) Sharing all layers is better than a subset. Subsequently, a model sharing the attention and feedforward modules, while keeping the LNs modality specific, across all 12 layers, is regarded as the baseline of our model family. We dub this Naïve modality sharing model MS-CLIP (see Fig. 1-2 and 1-3).

3.2   Modality-Specific Auxiliary Module Architecture

In this section we describe modifications introducing two lightweight modality-specific auxiliary modules, shown in Fig. 2. We name the full model with both modality-specific designs as MS-CLIP-S, where "S" indicates "Specialized branches".

Early Specialization: The first modality-specific design specializes only the first layer for visual and text modalities, leaving other layers shared. Concretely, on vision side, we employ a series of convolutional networks with residual connections 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, with ViT-B/32 as the visual encoder , inspired by [62]. For other visual encoders, such as ViT-B/16, the configuration only differs in the strides of convolutions (see Supplement). 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: For image representations, multi-scale information has been demonstrated to be valuable [4,52]. Vision Transformers [13], however,
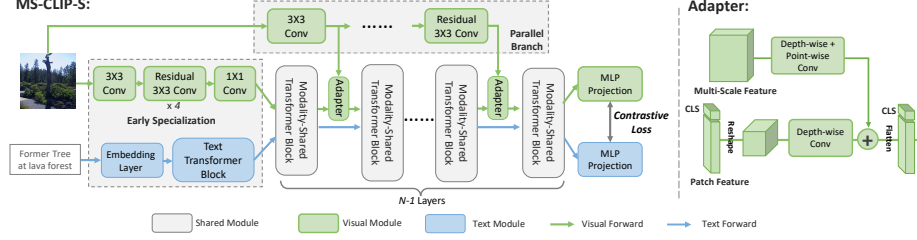
Fig. 2: Overview of MS-CLIP-S. Based on MS-CLIP, Early Specialization is further introduced at the beginning of the network. Simultaneously, an Efficient Parallel Branch is integrated to provide modality-specific multi-scale features to main modality-shared branch. See Section 3.2 text for more details.

Table 1: Setting of Early Specialization. $N\times N$ signifies the 2D kernel size of CNNs.

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	4.5M	

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		

typically operate on a fixed scale. In recent works that introduce multi-scale information into ViT [33,60], the patch size is gradually reduced and the dimension of the channel is increased, 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 [16], we propose to have an auxiliary parallel vision branch alongside the shared Transformer, which consists of one convolution layer and four residual convolution layers, to decrease the resolution and increase the channel dimension (see Fig. 2). In contrast with the plain residual convolutions in Early Specialization, here we utilize the bottleneck design in ResNet [18] to be parameter efficient. The main function of the parallel branch is to supplement the shared branch with multi-scale features for image information. Therefore, we employ one adapter after each parallel layer to integrate features from varying scales into layers of the shared Transformer. For further 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 $(H, H^{'})$ is the adapter's input and output, respectively. bn and ln denote batch normalization and layer normalization. Note that the CLS token is not fused with parallel branch and

remains unchanged. The outputs of 5 parallel layers are fused with every other shared Transformer layers. The detailed configuration when ViT-B/32 being visual encoder is provided in Tab. 2. For other visual encoders, such as ViT-B/16, only the kernel size and stride differs, and we attach the configuration in Supplementary.

4   Experiments

Section 4.1 introduces the pre-training and evaluation setup. Sections 4.2 and 4.3 details the primary experimental results and related ablations. Section 4.4 presents experiments where the pretraining data is changed. Finally, Section 4.5 presents experiments to better elucidate why MS-CLIP works.

4.1   Setup

Training Details: Similar to the original CLIP paper [47], 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 [35]. 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.

Pre-training Dataset: By default, we use YFCC100M [54] for pre-training. Following the filtering process in [47], we only keep image-text pairs where captions are in English. This leaves us around 22 million data pairs. All our results are reported on this data version, including the vanilla CLIP [47]. Subsequently, we also pre-train both our model and vanilla CLIP on a subset of the more recent dataset: LAION-400M [48]. More details can be found in Sec. 4.4.

Evaluation Datasets: In total, we adopt 25 public datasets for evaluation by either zero-shot learning or linear probing: ImageNet [10], Food-101 [3], CIFAR-10 [24], CIFAR-100 [24], SUN397 [61], Stanford Cars [23], FGVC Aircraft [39], Pascal Voc 2007 Classification [14], Describable Texture (DTD) [8], Oxford-IIIT Pets [44], Caltech-101 [15], Oxford Flowers 102 [42], MNIST [25], Facial Emotion Recognition (FER) [43], STL-10 [9], GTSRB [51], PatchCamelyon [56], UCF101 [50], Hateful Memes [22], Country211 [47], EuroSAT [19], Kitti-distance [17], Rendered-SST2 [49], Resisc45 [7], MSCOCO [32]. These datasets cover various categories, 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 [47]. For zero-shot recognition, we

Table 3: Experimental results of zero-shot image classification (ZS*), linear probing and zero-shot image-text retrieval (ITR*) across 25 datasets.

Eval.	Datasets		CLIP (ViT-B/32)	MS-CLIP-S (ViT-B/32)	Δ	CLIP (ViT-B/16)	MS-CLIP-S (ViT-B/16)	Δ
Linear Probing	Food-101		71.3	**76.0**	+4.7	80.1	81.5	+1.4
	SUN397		68.1	**71.7**	+3.6	72.3	**73.2**	+0.9
	Stanford Cars		21.8	**27.5**	+5.7	27.6	**32.0**	+4.4
	FGVC Aircraft		31.8	**32.9**	+1.1	33.6	**38.4**	+4.8
	Pascal Voc 2007		84.4	**86.1**	+1.7	85.6	**86.7**	+1.1
	DTD		64.1	**69.4**	+5.3	67.6	**71.9**	+4.3
	Oxford-IIIT Pets		61.1	**62.1**	+1.0	63.0	**63.7**	+0.7
	Caltech-101		**82.8**	81.6	−1.2	83.6	**83.8**	+0.2
	Oxford Flowers 102		90.7	**93.8**	+3.1	94.0	**95.2**	+1.2
	MNIST		96.5	**97.2**	+0.7	**96.9**	96.7	−0.2
	FER		**54.9**	53.6	−1.3	55.3	**56.2**	+0.9
	STL-10		**95.4**	95.1	−0.3	**96.9**	96.7	−0.2
	GTSRB		67.1	**69.9**	+2.8	72.5	**78.3**	+5.8
	PatchCamelyon		78.3	**81.3**	+3.0	**82**	80.4	−1.6
	UCF101		72.8	**74.6**	+1.8	74.6	**75.3**	+0.7
	CIFAR-10		**91.0**	87.2	−3.8	**91.1**	89.8	−1.3
	CIFAR-100		**71.9**	66.7	−5.2	**72.6**	71.5	−1.1
	Hateful Memes		50.6	**52.4**	+1.8	**51.6**	50.2	−1.4
	ImageNet		58.5	**63.7**	+5.1	64.7	66.7	+2.0
	Country211		19.9	**21.9**	+2.0	23.5	**23.6**	+0.1
	EuroSAT		**94.4**	93.5	−0.9	**94.6**	94.3	−0.3
	Kitti-distance		39.7	**45.1**	+5.4	35.7	**40.2**	+4.5
	Rendered-SST2		55.2	**56.0**	+0.8	56.8	**56.9**	+0.1
	Resisc45		83.3	**85.1**	+1.8	85.6	**86.5**	+0.9
	Avg.		66.9	**68.5**	+1.6	69.2	**70.4**	+1.2
ZS*	ImageNet		32.2	**36.7**	+4.5	36.9	**39.0**	+2.1
ITR* MSCOCO	I2T	R@1	24.4	**28.5**	+4.1	27.5	**29.9**	+2.4
		R@5	48.5	**54.1**	+5.6	51.9	**56.8**	+4.9
	T2I	R@1	14.8	**19.4**	+4.6	17.7	**20.4**	+2.7
		R@5	34.9	**40.8**	+5.9	38.7	**42.9**	+4.2

report zero-shot accuracy on the ImageNet [10] 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 [32]

4.2   Experimental Results

Compared Models: We conduct experiments on proposed MS-CLIP-S and vanilla CLIP [47]. Both ViT-B/32 and ViT-B/16 are adopted as visual encoders. As stated in Sec 4.1, we strictly follow the implementation in [47].

Zero-Shot ImageNet Classification: The experimental results are reported in the row of ZS* in Tab. 3. By comparing the four columns, we find that, based on ViT-B/32 (ViT-B/16), MS-CLIP-S can outperform CLIP by 4.5 (2.1) percentage points, or 13.9% (5.6%) relative, in zero-shot recognition accuracy on ImageNet.

Linear Probing: To fully compare our model with vanilla CLIP, the results of linear probing on 24 various datasets are shown in Tab. 3. Overall, with ViT-

Table 4: Experimental results of sharing different components in Transformer layer. First two rows are baselines without sharing. LN1 denotes the LN before Attn. LN2 denotes the LN before FFN.

Text Width	# Params	Shared Module	Non-Shared Module	Zero-shot Acc(%)
512	151M	-	Attn, FFN, LN1, LN2	32.15
768	209M	-	Attn, FFN, LN1, LN2	31.85
768	126M	Attn, FFN, LN1, LN2	-	28.40
768	126M	Attn, FFN, LN1	LN2	27.57
768	126M	Attn, FFN, LN2	LN1	32.16
768	126M	Attn, FFN	LN1, LN2	**32.99**

B/32 (ViT-B/16) as backbone, MS-CLIP-S outperforms vanilla CLIP on 18 (17) out of 24 tasks, and the average improvement on 24 tasks is 1.62 (1.16) points.

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 test set, which has 5,000 images. The comparison between MS-CLIP-S and vanilla CLIP, both pre-trained on YFCC, is shown in the last 4 rows of Tab. 3. With both ViT-B/32 and ViT-B/16, our MS-CLIP-S outperforms vanilla CLIP by a large margin across the board.

4.3   Ablation Study

For the following ablation analysis, we use ViT-B/32, and report zero-shot accuracy on ImageNet validation set.

On Modality-Shared Components: We systematically study the impact of varying the degree of sharing of components across different layers, and make the following observations:

1. **LNs need to be modality-specific.** We examine the shareable modules within each Transformer layer, excluding the input and output projection layers, which cannot be shared. As shown in Tab. 4, the first model variant shares all components (across all layers for simplicity), including two LN layers and transformation weights in the self-attention and feedforward modules, which results in worse performance (28.4%) compared to CLIP (ViT-B/32) (32.15%) and CLIP (ViT-B/32, T768)(31.85%). Then we examine making the two LN layers modality-specific, which yields better performance in both zero-shot accuracy (32.99%) and parameter efficiency. Note that the number of parameters in LNs is negligible compared with the transformation weights. Our observation echos the finding in FPT [38] that only tuning LNs in a mostly-frozen pre-trained language model yields satisfactory performance on vision tasks.

Table 5: 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	126M	132M	139M	153M	167M	181M	195M	209M	174M

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 the 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. 5 suggests that sharing all 12 layers performs the best while requiring the least number of parameters. This design of sharing all layers is what we refer to as MS-CLIP. Additionally, inspired by recent work on Neural Architecture Search (NAS) [66,12], we train a model that learns a policy to control which layer to (not) share via Gumbel Softmax [12]. Despite its sophistication, it still underperforms MS-CLIP.

On Modality-Specific Designs: We conduct experiments with the following settings: (1) CLIP (ViT-B/32): The same as [47], this uses ViT-B32 as the visual encoder, and Text Transformer with width set to 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. 4.3: 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. 3. (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.

The result is summarized in Tab. 6. By comparing the 2nd row and the 3rd row, we find that directly increasing the capacity of the text transformer yields worse results. 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) (as also shown in previous ablation on modality-shared modules). Comparing 4th and 5th row against the 1st row, we notice that early specialization can contribute to 2.1% improvement with only a 4M parameters increase and auxiliary parallel branch on vision has a 1.1% boost. The full model in 6th row further advances to 36.66%, a 4.5% absolute gain over the baseline CLIP (ViT-B/32).

4.4   Pre-training Data Quality

To verify that our proposed model can generalize to pre-training datasets of various quality, we pre-train both vanilla CLIP (ViT-B/32) and MS-CLIP-S (ViT-

Table 6: Ablation on Modality-Sepcific Designs.

Module Name	# Parameters	Zero-shot Acc(%)
CLIP (ViT-B/32)	151M	32.15
CLIP (ViT-B/32, T768)	209M	31.85
MS-CLIP (B/32)	126M	32.99
$\cdots$ w/ Early Specialization	129M	35.18
$\cdots$ w/ Parallel Branch	129M	34.18
MS-CLIP-S (B/32)	132M	**36.66**

Table 7: Results of models pre-trained on Laion-20M: zero-shot image classification, linear probing and zero-shot image-text retrieval (ITR*).

	ImageNet Zero-shot Acc(%)	MSCOCO Test ITR* I2T		T2I		Linear Probing on 24 datasets Average	#Wins
		R@1	R@5	R@1	R@5		
CLIP	35.5	24.7	48.1	16.2	35.8	70.5	5
MS-CLIP-S	**40.2**	**31.2**	**57.4**	**20.6**	**43.6**	**73.3**	**19**
Δ	+4.7	+6.5	+9.3	+4.4	+7.8	+2.8	+14

B/32) on a subset of the recently released public Laion-400M dataset [48]. This proof-of-concept subset contains 20M randomly-sampled image-caption pairs from Laion-400M, similar to the size of filtered YFCC. We name it Laion20M. The complete experimental results are shown in Tab. 7, where our model outperforms vanilla CLIP substantially. Since in the building of Laion-400M, a pre-trained CLIP is used to filter out the noisy image-text pairs, the dataset is believed to have higher data quality. This can also be proved by comparing vanilla CLIP's results in Laion and YFCC. Comparing Tab. 7 and Tab. 3 side by side, we find that the improvement brought by proposed MS-CLIP-S pre-trained on Laion-20M is generally higher than on YFCC (22M). It might imply that our method can benefit more when pre-training data quality is higher. Detailed performance of linear probing on 24 datasets is added in Supplementary.

4.5   Further Analysis

There are likely multiple reasons that explain observed improvements in performance. Firstly, sharing the majority of parameters across vision and language can implicitly encourage the model to focus on the common pattern across two modalities and alleviate overfitting of trivial vision (*e.g.*, illumination) or language cues (*e.g.* stop words). Additionally, the auxiliary modality-specific modules, *Early Specialization* and *Parallel Branch*, provide vision-specific multi-scale

Table 8: 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 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
··· w/ 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**

features and language-specific features to complement the shared modules. To have an in-depth understanding, we perform the following further analysis:

NMI Score: Shared model exhibits higher multi-modal fusion degree. To probe the multi-modal fusion degree, following [5], 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.

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. The measurement is performed on randomly sampled 50k image-caption pairs from YFCC100M dataset. NMI scores of all 12 layers and the average are listed in the first two rows of Tab. 8. 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.

Following the same procedure as above, we further report the NMI scores of MS-CLIP (B/32) + Early Specialization and MS-CLIP-S (B/32) (see Tab. 8). The result shows that sharing parameters and introducing early specialization can improve the multi-modal fusion degree, which coincides with our hypothesis mentioned above. However, adding parallel branch leads to a lower fusion score. This is somewhat conflicting with what we see in Tab. 6, where adding parallel branch enhances the learned representation. In the following subsection, we explore other metrics to further probe into what contributes to this behavior.

Multi-modal Common Semantic Structure: The Integration of Modality-Shared and Modality-Specific modules learns better common patterns. One of the hy-

potheses on why MS-CLIP architectures perform better is that they better capture the common semantic structures inherent to concepts from different modalities.
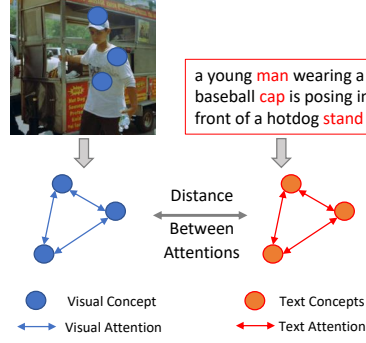

Fig. 3: Diagram of computing Common Semantic Structure distance.

To justify this hypothesis, we propose to measure the similarity between attention weights on visual concepts and the corresponding language concepts (see Fig. 3). The measurement is performed on a surrogate dataset named Flick30K-Entities [46], where object regions in each image are grounded to their corresponding phrases in a caption, in the form of bounding boxes. Given an image, assuming there are grounded object regions (visual concepts) $\{vc_1, vc_2, ..., vc_n\}$ and corresponding object words (language concepts) $\{tc_1, tc_2, ..., tc_n\}$, where tc_i is associated with vc_i semantically. In the h-th head of l-th attention layer, we denote the raw visual attention matrix as M^{lh} and the raw text attention matrix as K^{lh}. We then regard the attention value between tc_i and tc_j as K_{ij}^{lh}, and attention value between vc_i and vc_j as M_{ij}^{lh}. We extract the attention values from concept i to all other concepts (i.e., j) and normalize for visual attention and language attention, respectively (denoted as "attention vectors"). The final attention vectors are averaged over all heads in that attention layer. We compute the attention vectors for all concept pairs i. Finally, we measure the $l1$ distance between the visual attention vector and the language attention vector and sum them up over 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 learned across modalities. The whole process can be formulated as:

$$dis_{ij}^l = |\sum_{h=1}^{H} \frac{1}{H} softmax_i(M_{ij}^{lh}) - \sum_{h=1}^{H} \frac{1}{H} softmax_i(K_{ij}^{lh})| \quad (2)$$

$$CSC^l = dis^l = \sum_{i=1}^{n} \sum_{j=1}^{n} (dis_{ij}^l). \quad (3)$$

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. The first layers of MS-CLIP (B/32) + Early Specialization and MS-CLIP-S (B/32) are omitted as their vision branch do not contain any attention weights. The average score is computed based on the CSC distance on their last 11 layers. We find that our proposed modules lower the CSC distance and learn more modality-agnostic representation. Unsurprisingly, sharing parameters can enforce the attention to learn more common information. At the same
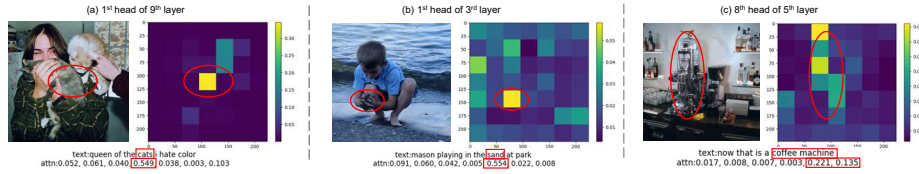
Fig. 4: Visualized attention maps of shared attention head (Please zoom in to see the caption).

time, it might reduce the overfitting brought by training separately. As for our proposed modality-specific modules, we suspect that these well designed modules can account for the discrepancy of individual modalities, especially by the vision-specific multi-scale feature, and thus facilitate the learning of the common patterns with the share component.

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 at the final layer from the *CLS* token to all its input tokens. For text, we perform the same except on the *EOS* token. Note that both *CLS* token and *EOS* token are treated as the feature representation. Results on MS-CLIP-S (B/32) are shown in Fig. 4. Interestingly, some heads are able to attend on the same concepts from different modalities. We take Fig. 4(a) as an example. Given the image and caption respectively as input, the 1st head of 9th attention layer gives the highest attention value to the region of "cat" in image and token "cats" in text. It suggests the learning of co-reference across modalities.

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/layers can be shared across modalities, we carefully investigate various architectural design choices through extensive experiments. In addition, we propose two modality-specific auxiliary designs: Early Specialization and Auxiliary Parallel Branch. Experiments on both zero-shot recognition and linear probing demonstrate the superiority of MS-CLIP architectures over the vanilla CLIP in both effectiveness and parameter efficiency. Finally, further analysis into the proposed architecture shows that sharing parameters can help map the two modalities into a closer embedding space and promote learning a common semantic structure.

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. arXiv preprint arXiv:2104.11178 (2021)
2. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision (2014)
4. Cai, Z., Fan, Q., Feris, R.S., Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection. In: European conference on computer vision. pp. 354–370. Springer (2016)
5. Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.C., Liu, J.: Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In: European Conference on Computer Vision. pp. 565–580. Springer (2020)
6. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019)
7. Cheng, G., Han, J., Lu, X.: Remote sensing image scene classification: Benchmark and state of the art. Proceedings of the IEEE **105**(10), 1865–1883 (2017)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)
9. Coates, A., Ng, A., Lee, H.: 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)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Desai, K., Johnson, J.: Virtex: Learning visual representations from textual annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11162–11173 (2021)
12. Dong, X., Yang, Y.: 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)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
14. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html (2007)
15. Fei-Fei, L., Fergus, R., Perona, P.: 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)
16. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)

17. Geiger, A., Lenz, P., Urtasun, R.: 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)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Helber, P., Bischke, B., Dengel, A., Borth, D.: 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)
20. Hu, R., Singh, A.: Unit: Multimodal multitask learning with a unified transformer. arXiv preprint arXiv:2102.10772 (2021)
21. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. arXiv preprint arXiv:2102.05918 (2021)
22. Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P., Testuggine, D.: The hateful memes challenge: Detecting hate speech in multimodal memes. arXiv preprint arXiv:2005.04790 (2020)
23. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 554–561 (2013)
24. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
25. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
26. Lee, S., Yu, Y., Kim, G., Breuel, T., Kautz, J., Song, Y.: Parameter efficient multimodal transformers for video representation learning. arXiv preprint arXiv:2012.04124 (2020)
27. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv preprint arXiv:2201.12086 (2022)
28. 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 arXiv:1908.03557 (2019)
29. Li, M., Xu, R., Wang, S., Zhou, L., Lin, X., Zhu, C., Zeng, M., Ji, H., Chang, S.F.: Clip-event: Connecting text and images with event structures. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16420–16429 (2022)
30. Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., Wang, H.: Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. arXiv preprint arXiv:2012.15409 (2020)
31. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021)
32. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:2103.14030 (2021)

34. Liu, Z., Stent, S., Li, J., Gideon, J., Han, S.: Loctex: Learning data-efficient visual representations from localized textual supervision. arXiv preprint arXiv:2108.11950 (2021)
35. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
36. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv preprint arXiv:1908.02265 (2019)
37. Lu, J., Goswami, V., Rohrbach, M., Parikh, D., Lee, S.: 12-in-1: Multi-task vision and language representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10437–10446 (2020)
38. Lu, K., Grover, A., Abbeel, P., Mordatch, I.: Pretrained transformers as universal computation engines. arXiv preprint arXiv:2103.05247 (2021)
39. Maji, S., Rahtu, E., Kannala, J., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. arXiv preprint arXiv:1306.5151 (2013)
40. Mu, N., Kirillov, A., Wagner, D., Xie, S.: Slip: Self-supervision meets language-image pre-training. arXiv preprint arXiv:2112.12750 (2021)
41. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. arXiv preprint arXiv:2107.00135 (2021)
42. Nilsback, M.E., Zisserman, A.: 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)
43. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and Expo. pp. 5–pp. IEEE (2005)
44. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.: Cats and dogs. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3498–3505. IEEE (2012)
45. Pham, H., Dai, Z., Ghiasi, G., Liu, H., Yu, A.W., Luong, M.T., Tan, M., Le, Q.V.: Combined scaling for zero-shot transfer learning. arXiv preprint arXiv:2111.10050 (2021)
46. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: 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)
47. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)
48. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
49. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: 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)
50. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
51. Stallkamp, J., Schlipsing, M., Salmen, J., Igel, C.: Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. Neural networks **32**, 323–332 (2012)

52. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1–9 (2015)
53. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
54. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Communications of the ACM **59**(2), 64–73 (2016)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
56. Veeling, B.S., Linmans, J., Winkens, J., Cohen, T., Welling, M.: Rotation equivariant CNNs for digital pathology (Jun 2018)
57. Wang, J., Hu, X., Gan, Z., Yang, Z., Dai, X., Liu, Z., Lu, Y., Wang, L.: Ufo: A unified transformer for vision-language representation learning. arXiv preprint arXiv:2111.10023 (2021)
58. Wang, W., Bao, H., Dong, L., Wei, F.: Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. arXiv preprint arXiv:2111.02358 (2021)
59. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
60. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808 (2021)
61. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: 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)
62. Xiao, T., Singh, M., Mintun, E., Darrell, T., Dollár, P., Girshick, R.: Early convolutions help transformers see better. arXiv preprint arXiv:2106.14881 (2021)
63. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
64. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
65. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., Beyer, L.: Lit: Zero-shot transfer with locked-image text tuning. arXiv preprint arXiv:2111.07991 (2021)
66. Zheng, X., Ji, R., Chen, Y., Wang, Q., Zhang, B., Chen, J., Ye, Q., Huang, F., Tian, Y.: Migo-nas: Towards fast and generalizable neural architecture search. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(9), 2936–2952 (2021)
67. Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L.H., Zhou, L., Dai, X., Yuan, L., Li, Y., et al.: Regionclip: Region-based language-image pretraining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16793–16803 (2022)
68. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13041–13049 (2020)