Compressing Visual-linguistic Model via Knowledge Distillation

Zhiyuan Fang[♠] , Jianfeng Wang[♡], Xiaowei Hu[♡], Lijuan Wang[♡], Yezhou Yang[♠], Zicheng Liu[♡]
[♠]Arizona State University, [♡]Microsoft Corporation

Abstract

Despite exciting progress in pre-training for visual-linguistic (VL) representations, very few aspire to a small VL model. In this paper, we study knowledge distillation (KD) to effectively compress a transformer based large VL model into a small VL model. The major challenge arises from the inconsistent regional visual tokens extracted from different detectors of Teacher and Student, resulting in the misalignment of hidden representations and attention distributions. To address the problem, we retrain and adapt the Teacher by using the same region proposals from Student's detector while the features are from Teacher's own object detector. With aligned network inputs, the adapted Teacher is capable of transferring the knowledge through the intermediate representations. Specifically, we use the mean square error loss to mimic the attention distribution inside the transformer block, and present a token-wise noise contrastive loss to align the hidden state by contrasting with negative representations stored in a sample queue. To this end, we show that our proposed distillation significantly improves the performance of small VL models on image captioning and visual question answering tasks. It reaches 120.8 *in CIDEr score on COCO captioning, an improvement of* 5.1 *over its non-distilled counterpart; and an accuracy of* 69.8 *on VQA 2.0, a* 0.8 *gain from the baseline. Our extensive experiments and ablations confirm the effectiveness of VL distillation in both pre-training and fine-tuning stages.*

1. Introduction

There have been exciting progress in visual linguistic (VL) pre-training to learn omni-representation models [44, 60, 9, 64, 83, 42] which could benefit a number of downstream tasks (*i.e.* , image captioning, VQA, image retrieval, etc.). The success can largely be attributed to the self-attention-based [68] transformer architecture, *e.g.* , BERT [14], which is effective in learning from image-text pairs at scale. So far, much of the work has focused on large models that suffer from high latency and large memory footprints at the time of inference, which limits their deployment to resource constrained edge devices for real-world applications.

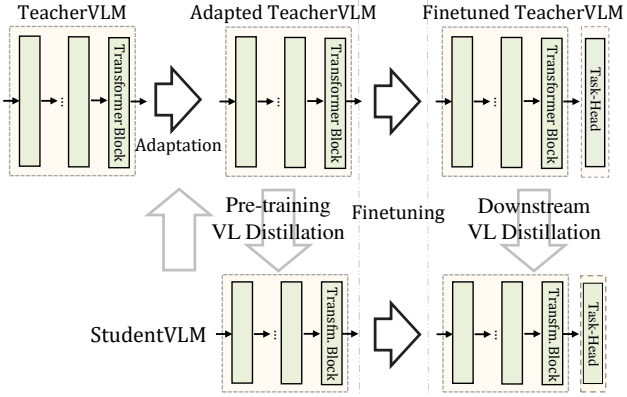

Figure 1: Overview of our proposed VL distillation schema. The VL model typically contains a region feature extraction module and a multi-modal transformer module. To have an aligned input, we adapt the Teacher VL model based on the region proposals from Student's region feature extractor. The VL distillation is then performed in both the pre-training stage and the fine-tuning stage.

As one of the effective techniques to compress large models, knowledge distillation (KD) [25, 6] was proposed by injecting the knowledge from a strong Teacher model into a smaller Student model without losing too much generalization power. Typically, the knowledge is transferred though mimicking the output logit [25, 57, 17], reducing the divergence of feature maps [80, 27, 78], or learning the intermediate layer representations [36, 1], *etc*.

In recent years, KD has been proven effective in compressing language models. For instance, Kim *et al.* [35] adopt KD for sequential model compression. In the transformer based language model, DistillBERT [57] reduces the size of the BERT-base model by 40% using a cosine embedding loss on the basis of hidden embedding in the transformer block, and a soft-target probability loss. TinyBERT [34], MobileBERT [63] and MiniLM [72] further highlight the importance of minimizing the self-attention

https://asu-active-perception-group.github.io/DistillVLM

distributions across Teacher and Student networks. In particular, [10] visually shows that attention maps in BERT capture substantial linguistic knowledge and syntactic relations that provide critical information during the distillation [34].

Heretofore, these advances have not been carried over to VL model compression. We identify the major challenges that prevent us from applying these techniques directly to VL distillation: Most existing VLP works [83, 42] use pre-trained object detector (*e.g.*, Faster-RCNN [54]) to extract regional features as visual tokens then feed them into the multi-modal transformer network for VL pre-training. A smaller VL model usually uses a lightweight detector for faster inference (*e.g.*, EfficientNet [65] based detector is adopted in [71] as visual feature extractor) that may be different from Teacher's detector. The object proposals from the two different detectors are usually very different, and there is no easy way to obtain the semantic correspondence between the two sets of object proposals. It is therefore unable to align the attention distributions or hidden embeddings between Student and Teacher.

To address the aforementioned challenges, we propose a set of strategies to enable distillation of VL models. First, instead of using object proposals from two different detectors, we use the same set of object proposals, obtained from Student's lightweight detector for the visual token extraction of both Teacher and Student (as shown in Figure 2). This ensures the semantic correspondence between the Teacher and Student's visual tokens. Second, we use a loss term to have the Student to mimic the Teacher's self-attention distribution at the last transformer layer. Third, We further distill the knowledge from the outputs of the transformer layers (*i.e.*, the hidden embeddings). We find that simply learning from the layer-wise Teacher embedding does not provide adequate supervision for the distillation. Hence, we use a noise contrastive loss to align the token embeddings by contrasting them with randomly sampled negative embeddings that are held in a sample queue. Figure 1 gives an overview of our proposed VL distillation schema, where VL distillation is applied for both the pre-training and fine-tuning stages. In order to examine the effectiveness of our VL distillation, we choose the same compact transformer architecture used in [72, 71], and the lightweight object detector as in [71], but leverages knowledge distillation techniques to facilitate the training of the small VL model (dubbed as DistillVLM). We show that our DistillVLM achieves a comparable performance to a large VL model, and clearly outperforms its non-distilled counterpart [71].

To summarize our contributions:

- For the first time, we propose VL distillation, a technique that leverages knowledge distillation to facilitate training of smaller VL models.
- Compared to non-distilled VL model pre-training, VL distillation offers a significant boosting in performance for VL tasks such as image captioning and visual question answering: DistillVLM achieves 120.8 in CIDEr score on COCO captioning [43] and 69.8 in accuracy on VQA [20] tasks, which are 5.1 more or 0.8 higher than the VL pre-training baselines.
- We provide extensive ablations of DistillVLM , and systematically analyze the effect of various KD strategies. This provides insights for future research on VL model distillation.

2. Related Work

Visual-linguistic Pre-training. Following the prominent progress in the transformer-based [68] pre-training in natural language [14, 51, 38, 5, 11, 52], visual-linguistic pre-training models, either for image+text [44, 64, 9, 42, 26, 81, 41, 18, 40, 45] or for video+text [61, 40, 46, 84, 39]. These representations have achieved great success when transferred to a number of downstream V+L tasks, *e.g.*, image/video captioning [3, 79, 70, 76, 26, 15], VQA [21, 4, 19], textual grounding [56, 16, 24, 82], *etc*. Most existing VL models are designed in a two-step fashion: a pre-trained object detector is used to encode the image as set of regional features (as offline visual tokens) followed by pre-training on a large scale visual-linguistic corpus using tasks like masked language modeling, image-text matching or masked region modeling losses. In particular, Zhang *et al.* [81] demonstrate the significant role of visual features in VL pre-training and looks for more effective visual representations from a larger object detector. Li *et al.* [42] shows that a larger transformer VL model can learn better from larger VL corpus. However, the marginal costs are greater than the marginal benefits. Recently, Wang *et al.* [71] propose a small VL model called MiniVLM that uses a lightweight visual feature extractor and smaller transformer to reduce the model size by 73% and maintain good accuracy on VL tasks. Nevertheless, the cost of pre-training on MiniVLM is associated with sub-optimal efficiency: it requires a large amount of training data (14M) to learn a good representation. Thus, it is worth exploring a more efficient way to train small VL models. There are other lines of VL pre-training works in which grid features [29, 32] are extracted from the convolutional layers without the proposal computation. [53, 50, 13] learn visual representation from scratch using Convolutional Neural Network as image encoder with a transformer for VL pre-training on a large amount of image-text pairs. The notion of VL distillation is not limited to just the two-stage VL models, it can potentially benefit other types of transformer based VL models as well.

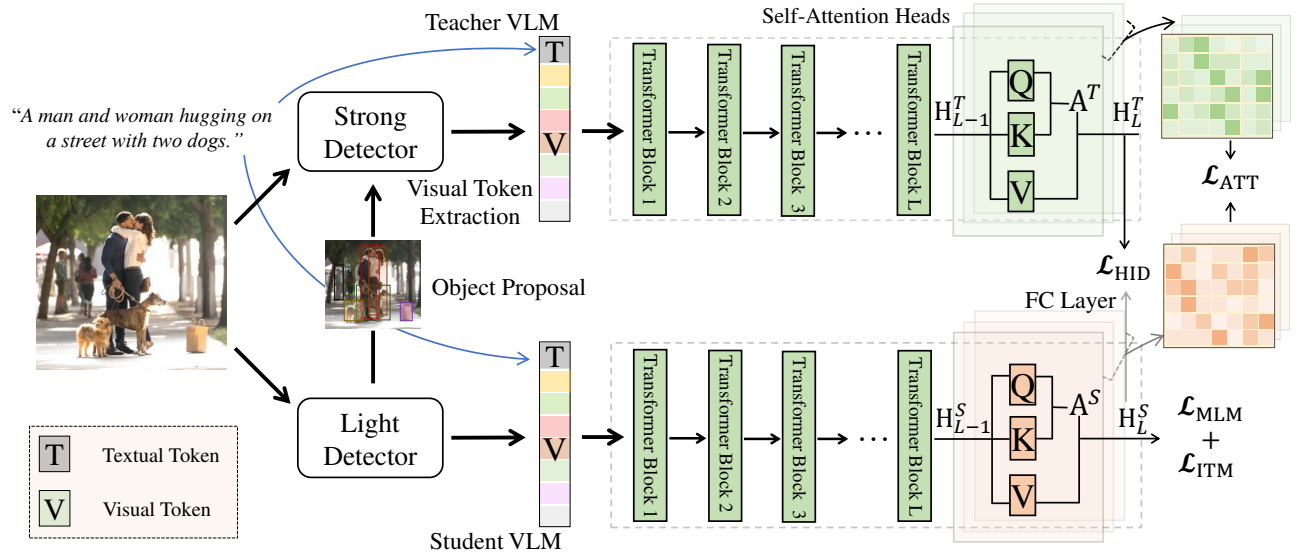

Figure 2: Illustration of our proposed DistillVLM architecture. The lightweight detector extracts the region features, and the region proposals are injected into the strong detector so that the region features are aligned between Teacher and Student. The Teacher transformer network is adapted with the new input before distillation. The Student VLM is distilled based on the hidden embedding matching and attention distribution alignment.

Knowledge Distillation has been applied to model compression task across different domains with its main goal being to transfer the "*knowledge*" $f(x_i)$ of sample (x_i, y_i) from a strong Teacher network (T) to the Student network (S) by minimizing the divergence between them:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{\text{S}}(x_i, y_i) + \mathcal{L}_{\text{KD}}\Big(f^S(x_i), f^T(x_i)\Big)\Big), \quad (1)$$

where $\mathcal{L}_{\text{S}}(\cdot)$ refers to the original supervision signal(s) on the Student. In practice, this term can possibly be replaced by the exclusive use of L_{KD}. Depending on the type of knowledge transferred, $\mathcal{L}_{\text{KD}}$ can derive from soft cross-entropy, mean squared error (MSE) function or *KL*-divergence. For example, [25, 6] transfer the learned knowledge by mimicking the mass function of the output probability across classes, or by minimizing the divergence of intermediate features [78, 36, 28, 77, 73]. [67, 67, 17] propose contrastive distillation for visual representation learning. In addition, remarkable advances have been made in knowledge distillation for language model compression (*i.e.*, BERT [14]), and these works show that mimicking the distribution of self-attention and intermediate representations of transformer blocks increases performances [57, 33, 63, 75] for downstream tasks. In particular, in the transformer-based language model distillation, DistillBERT [57] proposes to train the small BERT by mimicking the Teacher's output probability of masked language prediction and the embedding features. TinyBERT [33] and MobileBERT [63] leverage the layer-wise attention distributions for distillation with MSE function. [72] suggests distilling on the last transformer layer and bringing extra flexibility for training. [62, 8] also use the contrastive distillation in transformer based language model compression. [17, 62] propose using a sample queue to store history embeddings and show that contrasting with more negative samples is beneficial for knowledge distillation.

3. Visual-linguistic Knowledge Distillation

Compared to knowledge distillation in language models, VL knowledge distillation requires knowledge transferring from Teacher to Student in both modalities. We present DistillVLM for the task of visual-linguistic distillation (the overall architecture is illustrated in Figure 2), together with the detailed strategies for our model training.

3.1. Visual Token Alignment

VL pre-training methods such as OSCAR [42] take as input an image-text pair in the format of Word-Tag-Image triple $(\boldsymbol{w}, \boldsymbol{q}, \boldsymbol{v})$, where $\boldsymbol{w}$ and $\boldsymbol{q}$ denote the sequence of caption embedding and the word embedding of detected object tags (in texts). To obtain the visual tokens $\boldsymbol{v}$ and object tags, a set of image regional vectors are extracted from an object detector. A Faster R-CNN [55] detector pre-trained on Visual Genome [37] is used to extract the visual feature vector of each region, which is concatenated with its regional position coordinates to form a positional-sensitive region feature vector. This vector is then fed into a linear projection to ensure that the final vector $\boldsymbol{v}$ has the same dimension as the caption/tag embedding. The VL pre-training can be seen as a semantic alignment process between the

image regions and the textual units. It is worth mentioning that, the top regions to be extracted from the image is dependent on their associated confidence score output by the detector [71], which leads to some over-sampled and noisy visual tokens. Typically, the order of visual tokens is specified in descending order using the confidence score. As an alternative to Faster-RCNN, MiniVLM [71] uses a lightweight detector (*i.e.* , TEE) in which the backbone is replaced with EfficientNet [65] and a BiFPN [66] module is added to generate multi-scale features. These strategies obviously accelerate inference process, but inevitably also lead to different visual tokens between the Teacher and Student networks during distillation. For this reason, the direct application of the distillation loss to attention matrices or hidden representations leads to an invalid transfer of knowledge. Hence, we extract and align the Teacher/Student's visual tokens by using the same set of detected bounding boxes recognized by the lightweight detector, and keep the same token orders based on their confidence scores (as in Figure 2). Both the Teacher and Student VLM use the same object tags from the lightweight detector during the distillation. Having the Teacher use the visual tokens extracted by proposals from the lightweight detector may result in small performance drop. In practice, we address this issue by fine-tuning/re-training the Teacher VLM using the new visual tokens (Teacher adaptation).

3.2. Attention Distribution Distillation

One critical component of the transformer block is the multi-head self-attention module [68]. which enables contextualized information to be captured from an input sequence. A multi-head attention module outputs a set of attended values:

$$\texttt{Attention}\,(\mathbf{Q},\mathbf{K},\mathbf{V}) = \texttt{softmax}\,(\frac{\mathbf{Q}\cdot\mathbf{K}}{\sqrt{d_k}})\cdot\mathbf{V}, \quad (2)$$

where $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ denote query, key and value that are retrieved after three independent linear transformations on the hidden embedding $\mathbf{H}_i$ from i-th transformer block, and d_k is the dimension of key as a scaling factor. The dot-product between key and query after the softmax normalization is the attention matrix:

$$\mathbf{A} = \texttt{softmax}\,(\mathbf{Q}\cdot\mathbf{K}/\sqrt{d_k}). \quad (3)$$

Each transformer block consists of a set of consecutive linear transformations, which include one multi-head attention module, a two-layer feed forward network, followed by a normalization layer, and finally a residual connection.

Previous attempts in language model distillation [33, 63] have demonstrated the importance of transferring self-attention matrices, that are believed to contain latent linguistic information, *e.g.* , syntactic and co-reference relation of input tokens [10, 31]. [72] shows that using just the last transformer block's attention map yields equivalent results, allowing the Teacher and Student to have a different number of layers. In the case of the VL pre-training task, Cao *et al.* [7] show that certain attention matrices of the pre-trained VL models contain extensive intra-modal and cross-modal co-reference relations. These visual-linguistic knowledge is implicitly encoded, but shows a very promising potential for VL distillation. We formulate the distillation loss of the attention distribution by minimizing the divergence between the self-attention matrices of the last layer of the Teacher and the Student:

$$\mathcal{L}_{\text{ATT}} = \frac{1}{T\cdot H}\sum_{i=1}^{T}\sum_{j=1}^{H}\texttt{MSE}(\mathbf{A}^S_{i,j},\mathbf{A}^T_{i,j}), \quad (4)$$

where T, H denote the number of tokens and attention heads in a transformer. $\mathbf{A}_{i,j}$ is the normalized attention for i-th token at j-th head. We further study the effects of the distillation over the attention distribution in ablations.

3.3. Hidden Representation Distillation

Similar to previous works [33, 63], we also use the hidden representations for the Teacher and Student alignment during distillation. In particular, previous efforts formulate the task as minimizing the divergence of the hidden embedding ($\mathbf{H}\in\mathbb{R}^{T\times d}$) of every Transformer block, whose objective is as follows:

$$\mathcal{L}_{\text{HID-MSE}} = \frac{1}{T\cdot L}\sum_{i=1}^{T}\sum_{j=1}^{L}\texttt{MSE}(\mathbf{H}^S_{i,j}\mathbf{W}_h,\mathbf{H}^T_{i,j}), \quad (5)$$

and L stands for the number of transformer blocks. $\mathbf{W}_h$ is a learnable linear transformation that maps the Student hidden embedding into the identical dimension of Teacher embedding. However, there are limitations for such layer-to-layer alignment method. For example, TinyBERT must employ a uniform-function mapping to selectively choose a subset of the layers for learning, and MobileBERT requires the Teacher and Student to have identical number of layers. Since visual tokens are noisy during the VL distillation, this also leads to an increased difficulty in alignment. Sun *et al.* [62] propose CoDIR, which takes advantage of the noise contrastive estimation (NCE) loss to align the Teacher & Student's hidden representations by contrasting the target instance ($\mathbf{h}^S$) with more random instances as negative samples and aligning with its positive sample ($\mathbf{h}^T$), $\mathbf{h}\in\mathbb{R}^{d_T}$. Following [22, 17, 62], we employ a pre-defined instance queue $[\mathbf{h}^T_0,\mathbf{h}^T_1\cdots\mathbf{h}^T_K]$ to store K random sampled embeddings and one positive embedding from the Teacher network. And the objective of NCE is as:

$$\mathcal{L}_{\text{HID}} = -\log\frac{\exp(\mathbf{h}^S_i\cdot\mathbf{h}^T_i/\tau)}{\sum_{j=0}^{K}\exp(\mathbf{h}^S_i\cdot\mathbf{h}_j/\tau)}, \quad (6)$$

where τ denotes the temperature hyper-parameter, $\langle\cdot\rangle$ is the cosine similarity function. There are different ways to retrieve hidden representations $\mathbf{h}$, *e.g.*, [62] uses mean-pooled token representations as layer-wise summarized embedding. We find that applying the NCE loss to token-wise embedding leads to better distillation results, as discussed in Section 4.3. A linear mapping is introduced for the identical dimension transformation: $\phi : \mathbb{R}^{d_S} \rightarrow \mathbb{R}^{d_T}$ (d_S, d_T denote the hidden embedding dimension for Student and Teacher networks). To update the instance queue, we en-queue the Teacher-derived representation of the current batch ($\mathbf{h}^T$) and de-queue the earliest stored samples after the iteration. The introduction of the queue design enables batch-size independent distillation and allows the comparison with more contrastive samples with limited computational resources. In ablations, we discuss the effect of enlarging queue size and other distillation methods. In contrast to [22, 62], we store representations from the pre-trained and frozen Teacher network in the sample queue, which remain constant during training. This frees us from the use of momentum encoder like in [17].

3.4. Classification Distillation

The losses mentioned above allow the task-agnostic distillation during the pre-training stage. In addition, in the fine-tuning stage, we carry out knowledge distillation that benefits certain VL downstream tasks. Specifically, most VL downstream tasks are classification based tasks with labels, *e.g.*, image captioning or VQA tasks. Continuing the distillation at the downstream alleviates the domain gap brought by different pre-training VL corpus. As in [25], we minimize the softmax prediction of Student and Teacher networks and the loss is measured by the cross-entropy:

$$\mathcal{L}_{\text{CLS}} = \texttt{CE}(\mathbf{z}^S/\tau_d, \mathbf{z}^T/\tau_d), \tag{7}$$

where τ_d refers to the temperature parameter, and we simply maintain it as a constant $\mathbb{1}$. $\mathbf{z}^S/\mathbf{z}^T$ are the soft label outputs from Student/Teacher network.

3.5. Training

For the training, we keep the original VL pre-training objective losses ($\mathcal{L}_{\text{VLP}}$) [44] which consist of: masked language modeling loss ($\mathcal{L}_{\text{MLM}}$), where 15% of the textual tokens are masked and replaced with a special token `[MASK]` and the VL model is expected to classify these tokens; Image-text (contrastive) matching (ITM) loss ($\mathcal{L}_{\text{ITM}}$) where the model is expected to predict whether the image-text pair matches. Our final total loss on distillation at the pre-training stage is the combination of the above:

$$\mathcal{L} = \mathcal{L}_{\text{VLP}} + \alpha\mathcal{L}_{\text{ATT}} + \beta\mathcal{L}_{\text{HID}}, \tag{8}$$

where α are β are the weights of the loss terms. We find that $\mathcal{L}_{\text{CLS}}$ does not obviously contribute to the pre-training stage so we simply apply it at the fine-tuning distillation stage as:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{CLS}} + \alpha\mathcal{L}_{\text{ATT}} + \beta\mathcal{L}_{\text{HID}}, \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ is the original classification task in the specfic downstream-task. We study the effects of different learning losses in our ablations.

4. Experiments

In this section, we conduct extensive experiments on VL distillation both in pre-training and fine-tuning stages. To evaluate the effectiveness of our proposed distillation schema, we provide results and ablations for the image captioning and VQA tasks.

4.1. Datasets

Following [42], we construct our VL pre-training dataset by combining multiple existing VL datasets. Specifically, we use Conceptual Captions (CC) [59], SBU captions [47], training splits of Flicker30k [49], GQA [30], COCO Captions [43], and VQA-2.0 [20], yielding 4 million unique images, and 7 million image-text pairs (VL-7M). Both our Teacher model and DistillVLM are pre-trained on VL-7M and are then transferred to downstream VL tasks: image captioning on COCO Captions and visual question answering on VQA-2.0. We follow Karpathy's split[1] and have $\sim$11k images for training, and 5k/5k images for validation/testing. For the VQA task, we conduct downstream fine-tuning and testing on VQA-2.0 dataset, which consists of 83k images/444k questions for training, 41k images/214k questions for validation. For a fair comparison with previous works, we report results on `test-std` and `test-dev` splits via the online evaluation server[2], and compare ablation results using `test-dev` split.

4.2. Implementation Details

Visual Representation. Earlier VL pre-training (VLP) works mostly use Faster R-CNN [3, 54] or even advanced architecture [74, 83] for visual region representation extraction. To obtain visual tokens with more semantics, the object detector for VLP is usually pre-trained on Visual Genome Dataset [37], which contains $1,600$ object and 500 attribute categories. Following MiniVLM [71], we also adopt the EfficientNet [66] based lightweight object detector (TEE) for visual feature extraction. TEE reduces 90% of total inference time and has 91% fewer parameters (86.9M for R101-F *vs.* 7.5M for TEE). Same as MiniVLM, we also pre-train the TEE detector on Object365 [58] and Visual Genome [37] datasets before the visual representation extraction. We use R101 [23] based Faster-RCNN and TEE detected proposals for

[1] https://github.com/karpathy/neuraltalk2
[2] https://visualqa.org/challenge.html

Method	# Param	# *I-T* Pairs	Visual Feat.	P. D.	F. D.	COCO Captioning				VQA	
						B@4	M	C	S	test-std	test-dev
UVLP [83]	111.7M	3M	ResNeXt101	✗	✗	36.5	28.4	116.9	21.2	70.7	–
OSCAR$_B$ [42]	111.7M	7M	R101-F	✗	✗	36.5	30.3	123.7	23.1	73.4	73.2
MiniVLM [71]	34.5M	7M	TEE	✗	✗	34.3	28.1	116.7	21.3	-	-
MiniVLM [71]	34.5M	14M	TEE	✗	✗	35.6	28.6	119.8	21.6	69.4	69.1
DistillVLM	34.5M	7M	TEE	✗	✗	34.0	28.0	115.7	21.1	69.0	68.8
				✗	✓	34.5	28.2	117.1	21.5	69.2	69.0
				✓	✗	35.2	28.6	120.1	21.9	69.7	69.6
				✓	✓	35.6	28.7	120.8	22.1	69.8	69.6

Table 1: DistillVLM distills from stronger VL model (as Teacher), and retains high accuracy on COCO captioning task under different evaluating metrics, regardless of the effect brought by the lightweight visual feature extractor (TEE *v.s.* R101-F). Our model shows competitive results comparing to MiniVLM [71], even only half of the image-text pairs (# *I-T* Pairs) are available for pre-training. The VL distillation strategy brings consistent improvement in both the pre-training stage (P.D.) and fine-tuning stage (F.D.). All captioning methods are shown with cross-entropy optimization.

Teacher's regional visual representation extraction. This guarantees the semantic correspondence of the input tokens between Teacher and Student. Prevailing VL pre-training method like [42] shows that applying object tags in VL pre-training contributes to the performances. During distillation, we use consistent object tags detected by TEE for both the Teacher and Student networks. The lengths for object tags and visual tokens are 15 and 50, respectively.

VL Pre-training&Distillation. We use a compact transformer architecture for the VLP and VL distillation. In particularly, we follow [72, 71] and adopt a 12-layer transformer with 12 attention heads and 384 hidden dimension. For the Teacher model, we use Oscar$_b$ [42], a 12-layer transformer with 12 attention heads and 768 hidden size, pre-trained on the VL-7M corpus for 1M steps (100 epochs), with learning rate $5e^{-5}$ and batch size 768, using AdamW optimizer.[3] Overall, our compact transformer uses the same architecture as MiniVLM [71], and it has 34.5M learnable parameters and is 70% less than Oscar$_b$. For VL distillation, we first adapt the Teacher VLM by re-training it using the new visual tokens. Then, we keep the Teacher model frozen without further updating throughout the VL distillation. In contrast to [83, 42], weights in DistillVLM are randomly initialized without inheriting weights from BERT [14]. We adopt a learning rate at $2e^{-4}$ with batch size 768 for pre-training/distillation. We report and compare the effect of VL distillation with previous VLP baselines in Table 1. We set $\tau = \tau_d = 1$ and $\alpha = 10$, $\beta = 10$. Similar results are observed when using different values. We set the queue size to $4,096$ and further study the effect of different hyper-parameters in ablations.

Transferring to Downstream Tasks. In order to validate the efficacy of our proposed VL distillation schema, we transfer the pre-trained model to VL downstream tasks. Image captioning and VQA task can be formulated as a typical classification task, which enables direct task-specific distillation and comparisons in the downstream. We mainly examine them in this work, while the VL distillation is not task-specific and can be extended to other VL tasks as well. We conduct downstream distillation by using the output logit from downstream fine-tuned Teacher as soft-labels ($\mathcal{L}_{\text{TASK}}$). More details on distillations and ablations for the downstream tasks can be found at Appendix.

Image Captioning. We evaluate our model by transferring it to the image captioning task. We fine-tune our model by randomly masking out 15% of the caption tokens and impose a classification task to predict the masked token id using cross-entropy loss. Similar to [14], we trim and pad textual sentences to the length of 20. At inference, we recursively feed in `[MASK]` tokens and predict out captions one after the other with the beam search size at 1. The performance of captioning models is evaluated via BLEU@4 [48], METEOR [12], CIDEr [69] and SPICE [2] metrics. We perform the parameter search in a limited range: learning rate $\{2e^{-5}, 5e^{-6}\}$ and epochs $\{20, 30, 40\}$.

VQA. For the VQA task, the model must select the correct answer from the multi-options list given an image and textual question. We conduct fine-tuning on the VQA-2.0 dataset [20] and report the accuracy on `test-std` and `test-dev` splits. Following [3], we train the VQA model as a $3,129$-way classification task. We perform a light combinatorial parameter search on VQA task within a limited range: learning rate $\{1e^{-5}, 5e^{-5}\}$ and epochs $\{20, 40\}$.

4.3. Results and Analysis

Table 1 summarizes the results of DistillVLM using Oscar$_b$ as the Teacher model. We list VLP baselines with larger transformer architectures and stronger visual representations in the top lines. In particularly, DistillVLM without VL distillation achieves 34.0 BLEU@4 and 115.7 CIDEr scores with TEE visual representa-

[3] https://github.com/microsoft/Oscar

$\mathcal{L}_{\text{VLP}}$	$\mathcal{L}_{\text{ATT}}$	$\mathcal{L}_{\text{HID}}$	COCO Captioning				VQA
			B@4	M	C	S	test-dev
✓	✗	✗	33.0	27.3	110.6	20.4	68.5
✓	✓	✗	32.9	27.5	111.8	20.6	68.9
✓	✗	✓	34.0	27.8	114.4	21.1	69.2
✗	✓	✓	33.9	27.8	114.7	21.1	69.2
✓	✓	✓	34.6	27.9	115.6	21.3	69.4

Table 2: Detailed distillation effects based on attention matrices ($\mathcal{L}_{\text{ATT}}$), hidden hidden embedding ($\mathcal{L}_{\text{HID}}$), compared with VL pre-training losses ($\mathcal{L}_{\text{VLP}}$) at pre-training stage. Results are reported after 20 epochs of pre-training/distillation on 7M Image-Text pairs, then fine-tuned at the downstream (with cross-entropy optimization only).

tions using VLP [42] (masked language prediction and image-text matching losses). This is slightly lower than the performance reported by MiniVLM [71] pre-trained on VL-7M: 116.7 CIDEr score *vs*. our reproduced 115.7, which might be caused by the sub-optimal hyper-parameters. The apparent performance gaps between larger and smaller VLP models indicate the importance of visual representations so that the VL distillation is desired on small VL architectures. Notably, when equipped with downstream distillation, it performs better on COCO captioning dataset, 1.4 more on CIDEr, and 0.5 more on BLEU@4 scores. Downstream distillation on VQA task show marginal improvement: 69.2 *vs*. 69.0. We conjecture that this is mainly because the classification distillation on YES/NO or counting type of question does not provide better guidance, that the answers in the VQA task are mostly irrelevant/mutually exclusive. However, VL distillation in the pre-training stage increases the performances of DistillVLM on both captioning and VQA tasks consistently across all metrics: $\Delta = 1.2\%$ at B@4, 4.4 at CIDEr and 0.7 higher on VQA test-std split. Compared to its non-distilled counterpart MiniVLM [71], DistillVLM shows better results with only half the size of VL-corpus. To this end, the combination of the VL distillation in both pre-training and fine-tuning stage achieves the best results of DistillVLM, which shows comparable performances with Oscar_b: 120.8 *vs*. 123.7 with 70% fewer parameters. To learn more about DistillVLM, we conduct ablations on different designing options and examine the advantages of distillation at different epochs and data usage at Section 4.3.

Distillation over Different Losses. Table 2 presents the individual contribution of each distillation loss (attention matrices, hidden embedding) on the basis of the VL pre-training. The experiments for VL Pre-training/Distillation are trained for 20 epochs using identical hyper-parameters as before. From the table, we have the following observations: First, the non-distilled baseline alone reaches 110.6 CIDEr score for image captioning and 67.2 accuracy on VQA benchmark (shown in the first line of Table 2). By mimicking the distribution of attention, minor improvements are made, that is 1.2 for CIDEr and 0.4 for VQA scores respectively. Similarly, we observe the same trend when combining VLP with hidden embedding distillation. Compared with the VLP baseline, hidden embedding distillation significantly improves the performance under all criteria, demonstrating the efficacy of the alignment schema. In the end, the combination of all the loss terms gives the best performance, confirming that our proposed attention and hidden embedding distillation losses are complementary to each other. We find that using distillation objective alone also produces satisfactory performance, showing that knowledge transfer from distillation is to some extent equivalent to VL pre-training loss.

Methods	COCO Captioning				VQA
	B@4	M	C	S	test-dev
VL Pre-training [42]	33.0	27.3	110.6	20.4	68.5
Textual Distill	34.1	27.7	114.3	20.9	69.0
MSE + Layerwise	34.2	27.8	114.8	21.1	69.2
MSE + Last-layer*	33.3	27.6	112.4	20.7	68.5
MSE + Last-layer	34.3	27.8	115.3	21.2	69.4
NCE + Last-layer*	34.3	27.9	115.4	21.2	69.3
NCE + Last-layer	34.6	27.9	115.6	21.3	69.4

Table 3: Ablation of DistillVLM using different distillation strategies, *i.e.*, layer-to-layer distillation or last-layer distillation, using mean-square-error distance (MSE) or noise-contrastive (NCE) loss. Textual Distill represents applying the distillation only to the textual tokens without using visual tokens. Captioning results are reported after 20 epochs of training/distillation on VL-7M with cross-entropy optimization. * is the result using mean-pooled token embedding for distillation.

Different Distillation Strategies. Table 3 shows the results of distillation using different strategies, *i.e.*, layer-to-layer distillation *vs*. last-layer distillation, and MSE loss *vs*. NCE loss. We first study the effect of our proposed visual token alignment by applying attention distribution and hidden embedding distillation loss only to the textual token part: *e.g.*, using the "textual-to-textual" attention sub-matrices and their corresponded textual token embedding. The second line in Table 3 is the result of textual distillation, which shows a slight improvement over the VLP baseline. Following previous language distillation works [33], we also conduct the layer-to-layer attention and hidden distillation between Teacher and Student, and observe inferior performances than the last-layer strategy. Beyond that, the layer-to-layer method can also be severely limited by their architectural structures [72] (*e.g.*, different number of layers and attention heads). "NCE + Last-layer" represents the results of DistillVLM using our proposed contrastive objective function that uses negative samples for

# *Neg.*	COCO Captioning				VQA
	B@4	M	C	S	test-dev
1	33.3	27.6	112.5	20.7	68.5
128	33.6	27.7	112.7	20.9	68.9
512	33.7	27.8	113.3	21.0	68.8
1,024	34.1	27.9	114.7	21.2	69.1
4,096	34.3	27.9	115.4	21.2	69.3

Table 4: Effect of the number of negative samples for noise contrastive estimation loss. A larger queue size incrementally contributes to the distillation performance. When queue size approaches 1, the NCE loss is approximately the MSE loss with an only positive anchor from the Teacher. All the experiments are trained for 20 epochs using sample queues in different sizes on VL-7M, and then transferred to downstream.

the alignment learning. We find that contrastive learning leads to slightly better results than MSE loss. To this end, we study the differences in using token-wise embedding and the mean-pooled layer-wise embedding for contrastive learning and observe that learning with token-wise embedding gives much better results, which is a different observation from [62]. However, applying the mean-pooled embedding with NCE loss mitigates this issue and gives on par results with token-wise NCE method (see last two lines in Table 3). We further provide the ablations of VL distillation for the downstream tasks in the appendix. In Table 4, we study the effect of using more negative samples in NCE loss. We observe that increasing the size of the sample queue can steadily contribute to the performances of VL models. Especially, when we only use one negative sample, the model reaches 112.5 CIDEr score, which aligns with the MSE results (112.4 CIDEr score) at Table 3. When increased to 4,096, the model performs best across all metrics. While continuing to use more negative samples may produce better results, we just set the size of the sample queue as 4,096 in our experiments. Note that our queue stores the random sample representations from Teacher VLM, which remain consistent throughout the distillation process. This also implies the feasibility of leveraging in-batch samples for contrastive learning, while the queue design relieves the model from batch-size requirements and allows the use of more negative samples.

Data-efficient VL Distillation. One critical aspect of VL distillation for real-world application is its ability to efficiently train smaller VL model with limited cost, *i.e.*, with a smaller VL corpus (data scarcity) and less converging epochs (training efficiency). To further assess whether VL distillation can cope with these challenges, we perform VL distillation at the pre-training stage when trained with 1, 5, 10, 20, 50, 100 epochs and compare their results with VLP. In addition, as pointed by [45] that specific partial VL data might contribute more to performances, we propose to conduct VL distillation/pre-training using evenly sampled partial data (1%, 5%, 10%, 20%, 50% and 100% of VL-7M). These also help to verify whether DistillVLM benefits from more converging epochs and more VL data. Several conclusions can be drawn from the above results. First, VL distillation brings a consistent CIDEr gain across different training epochs. Non-distilled VL pre-training method achieves only 99.8 CIDEr score with 1 epoch of training, while DistillVLM reaches 103.1 (see Figure 3). Notably, CIDEr score of DistillVLM increases steadily with more training epochs. When it comes to using different percentages of VL data, we also see a similar trend. In the most extreme case, with only 1% of VL-7M corpus available, VL pre-training produces 89.1 CIDEr score, 4.1 lower than the VL distillation. With more image-text pairs, VL distillation obviously gives even better results: 8.6 higher for 10% and 5.0% higher for 100%. This shows that regardless of the amount of data available, VL distillation provides more effective and informative supervision than the normal pre-training strategy.

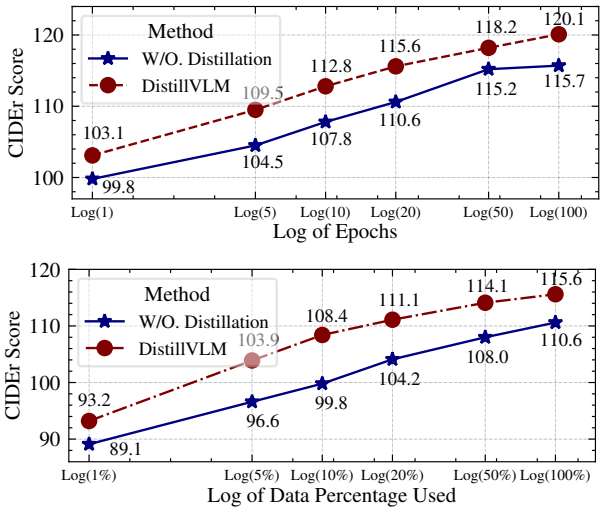

Figure 3: Top: Captioning CIDEr score gain from VL distillation under different epochs (1, 10, 20, 50, 100) in pre-training/distillation on VL-7M; Bottom: Using 1%, 10%, 20%, 50% and 100% of VL-7M image-text pairs with 20 epochs of pre-training/distillation.

5. Conclusion

We have proposed the first VL distillation, which leverages the knowledge distillation technique to compress large visual-linguistic models. Our experiments confirmed the validity of VL distillation from several aspects: Compared to the non-distilled VL pre-training method, VL distillation not only brings better performances, it is also more data efficient. Our extensive ablations also verified that our VL distillation strategies are simple yet effective.

Acknowledgements: Z. Fang and Y. Yang are partially supported by the US National Science Foundation project #1750082.

References

[1] S. Ahn, S. X. Hu, A. Damianou, N. D. Lawrence, and Z. Dai. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9163–9171, 2019.

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[5] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[6] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[7] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, and J. Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pages 565–580. Springer, 2020.

[8] L. Chen, Z. Gan, D. Wang, J. Liu, R. Henao, and L. Carin. Wasserstein contrastive representation distillation. *arXiv preprint arXiv:2012.08674*, 2020.

[9] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. Uniter: Learning universal image-text representations. *European Conference on Computer Vision*, 2020.

[10] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning. What does bert look at? an analysis of bert's attention. *In Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.

[11] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020.

[12] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[13] K. Desai and J. Johnson. Virtex: Learning visual representations from textual annotations. *arXiv preprint arXiv:2006.06666*, 2020.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *North American Chapter of the Association for Computational Linguistics*, 2019.

[15] Z. Fang, T. Gokhale, P. Banerjee, C. Baral, and Y. Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *Conference on Empirical Methods in Natural Language Processing*, 2020.

[16] Z. Fang, S. Kong, C. Fowlkes, and Y. Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6378–6388, 2019.

[17] Z. Fang, J. Wang, L. Wang, L. Zhang, Y. Yang, and Z. Liu. Seed: Self-supervised distillation for visual representation. *The International Conference on Learning Representations*, 2021.

[18] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

[19] T. Gokhale, P. Banerjee, C. Baral, and Y. Yang. Vqa-lol: Visual question answering under the lens of logic. In *European conference on computer vision*, pages 379–396. Springer, 2020.

[20] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] L. A. Hendricks, R. Hu, T. Darrell, and Z. Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279, 2018.

[25] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop (2015)*, 2015.

[26] X. Hu, X. Yin, K. Lin, L. Wang, L. Zhang, J. Gao, and Z. Liu. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[27] Z. Huang and N. Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[28] Z. Huang and N. Wang. Like what you like: Knowledge distill via neuron selectivity transfer. *arXiv preprint arXiv:1707.01219*, 2017.

[29] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020.

[30] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question

answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[31] G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[32] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen. In defense of grid features for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10267–10276, 2020.

[33] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*, 2019.

[34] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online, Nov. 2020. Association for Computational Linguistics.

[35] Y. Kim and A. M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, Nov. 2016. Association for Computational Linguistics.

[36] A. Koratana, D. Kang, P. Bailis, and M. Zaharia. Lit: Learned intermediate representation training for model compression. In *International Conference on Machine Learning*, pages 3509–3518. PMLR, 2019.

[37] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[38] K. Lagler, M. Schindelegger, J. Böhm, H. Krásná, and T. Nilsson. Gpt2: Empirical slant delay model for radio space geodetic techniques. *Geophysical research letters*, 40(6):1069–1073, 2013.

[39] J. Lei, L. Li, L. Zhou, Z. Gan, T. L. Berg, M. Bansal, and J. Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *arXiv preprint arXiv:2102.06183*, 2021.

[40] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

[41] L. Li, Z. Gan, and J. Liu. A closer look at the robustness of vision-and-language pre-trained models. *arXiv preprint arXiv:2012.08673*, 2020.

[42] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[44] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[45] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[46] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020.

[47] V. Ordonez, G. Kulkarni, and T. Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24:1143–1151, 2011.

[48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[49] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015.

[50] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[51] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.

[52] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019.

[53] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*, 2021.

[54] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[55] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[56] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[57] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[58] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun. Objects365: A large-scale, high-quality dataset

for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8430–8439, 2019.

[59] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[60] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *The International Conference on Learning Representations*, 2020.

[61] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473, 2019.

[62] S. Sun, Z. Gan, Y. Cheng, Y. Fang, S. Wang, and J. Liu. Contrastive distillation on intermediate representations for language model compression. *arXiv preprint arXiv:2009.14167*, 2020.

[63] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou. MobileBERT: a compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2158–2170, Online, July 2020. Association for Computational Linguistics.

[64] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *Conference on Empirical Methods in Natural Language Processing*, 2019.

[65] M. Tan and Q. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[66] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[67] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.

[68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (Neurips)*, 2017.

[69] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[70] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[71] J. Wang, X. Hu, P. Zhang, X. Li, L. Wang, L. Zhang, J. Gao, and Z. Liu. Minivlm: A smaller and faster vision-language model. *arXiv preprint arXiv:2012.06946*, 2020.

[72] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[73] Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020.

[74] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[75] C. Xu, W. Zhou, T. Ge, F. Wei, and M. Zhou. Bert-of-theseus: Compressing bert by progressive module replacing. *arXiv preprint arXiv:2002.02925*, 2020.

[76] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[77] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

[78] J. Yim, D. Joo, J. Bae, and J. Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4133–4141, 2017.

[79] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[80] S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *The International Conference on Learning Representations*, 2017.

[81] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Making visual representations matter in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

[82] F. Zhiyuan, K. Shu, Y. Tianshu, and Y. Yezhou. Weakly supervised attention learning for textual phrases grounding. *CVPR Workshop on Vision and Language*, 2018.

[83] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.

[84] L. Zhu and Y. Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8746–8755, 2020.