Distilled Dual-Encoder Model for Vision-Language Understanding

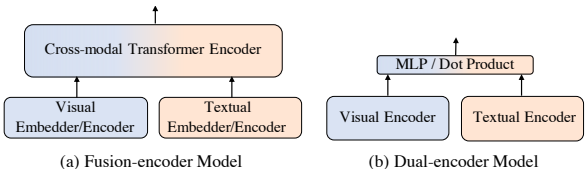
Zekun Wang[†*], **Wenhui Wang**[‡], **Haichao Zhu**[†], **Ming Liu**[†♮], **Bing Qin**[†♮], **Furu Wei**[‡]
[†]Harbin Institute of Technology, Harbin, China
[‡]Microsoft Research, Beijing, China
[♮]Peng Cheng Laboratory, Shenzhen, China
{zkwang,hczhu,mliu,qinb}@ir.hit.edu.cn
{wenwan,fuwei}@microsoft.com

Abstract

On **v**ision-**l**anguage **u**nderstanding (VLU) tasks, fusion-encoder vision-language models achieve superior results but sacrifice efficiency because of the simultaneous encoding of images and text. On the contrary, the dual-encoder model that separately encodes images and text has the advantage in efficiency, while failing on VLU tasks due to the lack of deep cross-modal interactions. To get the best of both worlds, we propose DIDE[1], a framework that distills the knowledge of the **fusion-encoder teacher** model into the **dual-encoder student** model. Since the cross-modal interaction is the key to the superior performance of teacher model but is absent in the student model, we encourage the student not only to mimic the predictions of teacher, but also to calculate the cross-modal attention distributions and align with the teacher. Experimental results demonstrate that DIDE is competitive with the fusion-encoder teacher model in performance (only a 1% drop) while enjoying $4\times$ faster inference. Further analyses reveal that the proposed cross-modal attention distillation is crucial to the success of our framework.

1 Introduction

Vision-language understanding (VLU) tasks (*e.g.*, visual reasoning (Suhr et al., 2019), visual entailment (Xie et al., 2019), visual question answering (Goyal et al., 2017)) require the model to understand the cross-modal interactions between images and text. Various fusion-encoder vision-language pretrained models (Tan and Bansal, 2019; Chen et al., 2020; Zhang et al., 2021; Kim et al., 2021; Dou et al., 2022; Alayrac et al., 2022) are proposed for VLU tasks. As shown in Figure 1(a), these models employ a Transformer (Vaswani et al., 2017) network as a cross-modal encoder to capture interactions between different modalities. Despite the remarkable performance, the heavy cross-modal encoder remains an efficiency bottleneck due to the simultaneous encoding of images and text, limiting the application in practical scenarios with massive images or text. Therefore, it is crucial to find an approach to accelerate inference for VLU.



(a) Fusion-encoder Model (b) Dual-encoder Model

Figure 1: Illustration of two architectures of vision-language models. (a) Fusion-encoder models simultaneously encode visual and textual inputs via modal-specific embedders/encoders and employ a cross-modal Transformer encoder to fuse representations. (b) Dual-encoder models encode images/text separately and adopt an extreme lightweight module (*e.g.*, MLP) for cross-modal interactions.

We turn to explore the dual-encoder vision-language model (shown in Figure 1(b)), which encodes images and text separately and then applies an extreme lightweight shallow module to model the interactions between modalities. The disentangled encoding paradigm enables off-line computing and caching visual or textual representations on demand, significantly lowering runtime latency. However, the shallow module is insufficient to handle complex VLU tasks, resulting in previous models (Radford et al., 2021; Jia et al., 2021) falling far behind fusion-encoder models (Kim et al., 2021). Can dual-encoder models obtain performance comparable to fusion-encoder models while preserving efficiency? In this work, we propose DIDE (a knowledge **Di**stillation framework for **D**ual-**E**ncoder models), where the dual-encoder model (student) is supervised by the fusion encoder models (teacher), as shown in Figure 2. Although soft label distillation (Hinton et al., 2015) is widely applied, our key observation is that *cross-modal*

[*]Contribution during internship at Microsoft Research.
[1]Our code and models will be publicly available at https://github.com/kugwzk/DiDE.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 8901–8913
December 7-11, 2022

attention distributions[2] are absent in the student, resulting in the inability to model complex cross-modal interactions. Thus, only distilling soft labels is not enough for the student to mimic the interactions of teacher deeply.

Considering that cross-modal interaction is critical for VLU, we introduce a plug-and-play objective *cross-modal attention distillation*, as fine-grained supervision to help the student better learn cross-modal interactions. Specifically, besides the soft label distillation, we compute the cross-modal attention of the student model and align it with the distribution in the teacher model during training. The training of DIDE consists of two-stage distillations. In the pre-training stage, the student learns a general initialization with distillation. In the fine-tuning stage, distillation helps the student learn more task-specific knowledge. Experimental results demonstrate that DIDE performs competitively with the fusion-encoder teacher model in various VLU tasks (retaining 96.9% to 99.9% performance) while having a $4\times$ inference speedup. Further analysis indicates that the proposed cross-modal attention distillation yields significant gains compared to distilling only with soft labels or other latent features of the teacher. Beyond VLU, DIDE also shows effectiveness in image-text retrieval.

Our contributions are summarized as follows:

- We propose DIDE, a knowledge distillation framework for the dual-encoder model to learn better cross-modal interactions of vision-language understanding from the fusion-encoder model.
- Our approach is plug-and-play with different vision-language tasks and can be applied on different model architectures.
- Experimental results show that our distilled model performs competitively with the teacher model and has a significant speedup. Further analysis indicates that our proposed cross-modal attention distillation is the key to success.

2 Related Work

2.1 Vision-Language Pre-Training

Language and vision pre-training advance the state of the art in downstream natural language processing tasks (Radford et al., 2018; Devlin et al., 2019; Dong et al., 2019; Liu et al., 2019; Bao et al., 2020; Lewis et al., 2020; Raffel et al., 2020; Conneau et al., 2020; Chi et al., 2021) and computer vision tasks (Dosovitskiy et al., 2021; Touvron et al., 2021; Bao et al., 2021). Vision-Language pre-training (Lu et al., 2019; Su et al., 2020; Gan et al., 2020; Li et al., 2021b; Wang et al., 2021a,b) has been shown to prevail in learning cross-modal representations. The model architectures fall into two lines: *fusion-encoder* and *dual-encoder* models. Fusion-encoder models jointly encode image-text pairs and employ a multi-layer cross-modal Transformer encoder to fuse the visual and textual representations. Previous models (Li et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020; Zhang et al., 2021) extract visual features through a pre-trained object detector (*e.g.*, Faster R-CNN (Ren et al., 2015)), which requires high-resolution input images and brings more computation costs. Huang et al. (2020); Li et al. (2021a); Dou et al. (2022) directly take image pixels or patches as input and encode visual features by CNN or Vision Transformer (Dosovitskiy et al., 2021). ViLT (Kim et al., 2021) directly applies a shared Transformer for joint encoding of image patches and textual token embeddings, achieving competitive performance with less overhead. The models exhibit a strong ability to model complex cross-modal interactions and achieve superior results on VLU tasks. However, the models rely on a cross-modal Transformer encoder to fuse visual and textual features simultaneously across layers, demanding a heavy computation budget and leading to a low inference speed.

On the contrary, dual-encoder models (Radford et al., 2021; Jia et al., 2021; Sun et al., 2021) encode images and text separately and take an MLP or dot product to model the interactions between the modalities. These models have the advantage of computational efficiency. The attention mechanism is computed only within tokens of the same modality. Moreover, thanks to the independent encoders, the visual or textual representations can be precomputed and cached off-line in the practical scenarios. However, the shallow module is not enough to handle complex cross-modal interactions, causing significant performance degradation on VLU tasks (Kim et al., 2021; Hendricks et al., 2021). To get the best of both worlds, we preserve the inference efficiency of dual-encoder model while achieving promising results on VLU

[2] visual-to-textual (blue) and textual-to-visual (orange) attention in Figure 2.

by knowledge distillation.

2.2 Knowledge Distillation

Knowledge distillation (KD; Hinton et al. (2015) aims to improve a student model by transferring knowledge from a teacher model. Transformer distillation is widely used in various domains (Jiao et al., 2020; Wang et al., 2020; Touvron et al., 2021; Fang et al., 2021). In this work, we focus on distillation under the cross-architecture setting (Hofstätter et al., 2020), where the architectures of the teacher and the student are different. Cao et al. (2020) decomposes the early layers of the Transformer and adopts a complete Transformer to guide the training for reading comprehension. In image-text retrieval, Miech et al. (2021) proposes distilling soft labels from a cross-attention model to a dual-encoder model with the reranking mechanism. But in VLU tasks, their method does not work, whereas our proposed cross-modal attention distillation is critical for success.

3 Method

Figure 2 gives an overview of our DIDE framework, a knowledge distillation approach for the dual-encoder model.

3.1 Model Overview

Input Representations We slice the input image $\boldsymbol{v} \in \mathbb{R}^{H\times W\times C}$ into patches $\boldsymbol{v}^p \in \mathbb{R}^{N\times(P^2C)}$, where $N = HW/P^2$ is the number of patches, (H, W) is the resolution of the input image, (P, P) is the resolution of each patch, and C is the number of channels. The input text $\boldsymbol{t}$ is tokenized into a sequence of M tokens. We prepend the special tokens `[I_CLS]` and `[T_CLS]` to the sequence of image patches and text tokens, respectively. We linearly project image patches $\boldsymbol{v}^p$ to obtain patch embeddings, and the final visual input embeddings $\boldsymbol{H}_0^v \in \mathbb{R}^{(N+1)\times D}$ are computed via:

$$\boldsymbol{H}_0^v = [\boldsymbol{v}_{\texttt{[I\_CLS]}}, \boldsymbol{V}\boldsymbol{v}_1^p, \ldots, \boldsymbol{V}\boldsymbol{v}_N^p] + \boldsymbol{V}_{pos} + \boldsymbol{V}_{type}$$

where $\boldsymbol{V} \in \mathbb{R}^{(P^2C)\times D}$ is linear projection, $\boldsymbol{V}_{pos} \in \mathbb{R}^{(N+1)\times D}$ is 1D positional embedding, $\boldsymbol{V}_{type} \in \mathbb{R}^D$ is visual type embedding. The textual input embeddings $\boldsymbol{H}_0^t \in \mathbb{R}^{(M+1)\times D}$ are obtained by summing word embeddings $\boldsymbol{W}$, the textual position embedding $\boldsymbol{T}_{pos}$ and textual type embedding $\boldsymbol{T}_{type}$:

$$\boldsymbol{H}_0^t = [\boldsymbol{w}_{\texttt{[T\_CLS]}}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_M] + \boldsymbol{T}_{pos} + \boldsymbol{T}_{type}$$

We take $\boldsymbol{H}_0^v, \boldsymbol{H}_0^t$ as visual and textual inputs for the teacher and student models.

Fusion-Encoder Model (Teacher) concatenates the input representations $\boldsymbol{H}_0^v$ and $\boldsymbol{H}_0^t$ as $\boldsymbol{H}_0^{vl} = [\boldsymbol{H}_0^v; \boldsymbol{H}_0^t]$, and feeds into a L-layer cross-modal Transformer encoder to obtain contextual representations $\boldsymbol{H}_L^{vl}$. The cross-modal Transformer encoder fuses representations of different modalities via the multi-head attention mechanism. Specifically, for each head a, the whole attention distribution $\boldsymbol{A}_a^{vl}$ is computed via:

$$\boldsymbol{A}_a^{vl} = \text{softmax}(\frac{\boldsymbol{Q}_a^{vl}\boldsymbol{K}_a^{vl\intercal}}{\sqrt{d_k}})$$

where queries $\boldsymbol{Q}_a^{vl}$ and keys $\boldsymbol{K}_a^{vl}$ are obtained by linearly projecting the hidden states using parameters $\boldsymbol{W}_{l,a}^Q, \boldsymbol{W}_{l,a}^K \in \mathbb{R}^{D\times d_k}$, respectively. d_k is the size of the attention head. The output vectors of `[I_CLS]` and `[T_CLS]` are fed into the task-specific layer to obtain predictions.

Dual-Encoder Model (Student) encodes $\boldsymbol{H}_0^v$ and $\boldsymbol{H}_0^t$ are separately via visual and textual Transformer encoders: $\boldsymbol{H}_l^v$ and $\boldsymbol{H}_l^t$. The output vectors of `[I_CLS]` and `[T_CLS]` are used as final representations of the images and text. Then, a shallow module f is applied to fuse the two representations. For vision-language understanding, we adopt an MLP as the module f. For image-text retrieval, we use the dot product function to obtain similarity scores of image-text pairs.

3.2 Distillation Objectives

Cross-Modal Attention Distillation The Transformer captures fine-grained interactions between tokens, mainly benefiting from the multi-head attention mechanism (Hao et al., 2021). In the dual-encoder model, tokens only compute attention with those within the same modality, named *uni-modal attention*: visual-to-visual $\boldsymbol{A}^{v2v} \in \mathbb{R}^{N\times N}$ and textual-to-textual $\boldsymbol{A}^{t2t} \in \mathbb{R}^{M\times M}$. As shown in Figure 2, compared to the whole attention $\boldsymbol{A}^{vl} \in \mathbb{R}^{(N+M)\times(N+M)}$ of the fusion-encoder model, we observe that the attention of computing intermodality tokens (named *cross-modal attention*[3]) is absent in the student, including visual-to-textual $\boldsymbol{A}^{v2t} \in \mathbb{R}^{N\times M}$ and textual-to-visual $\boldsymbol{A}^{t2v} \in \mathbb{R}^{M\times N}$. Thus, the student lacks the ability to capture cross-modal interactions, which is critical for vision-language tasks.

Taking into account the weakness of the dual-encoder model, we propose the *cross-modal at-*

[3]We name the uni-modal attention and cross-modal attention from the perspective of dual-encoder models.

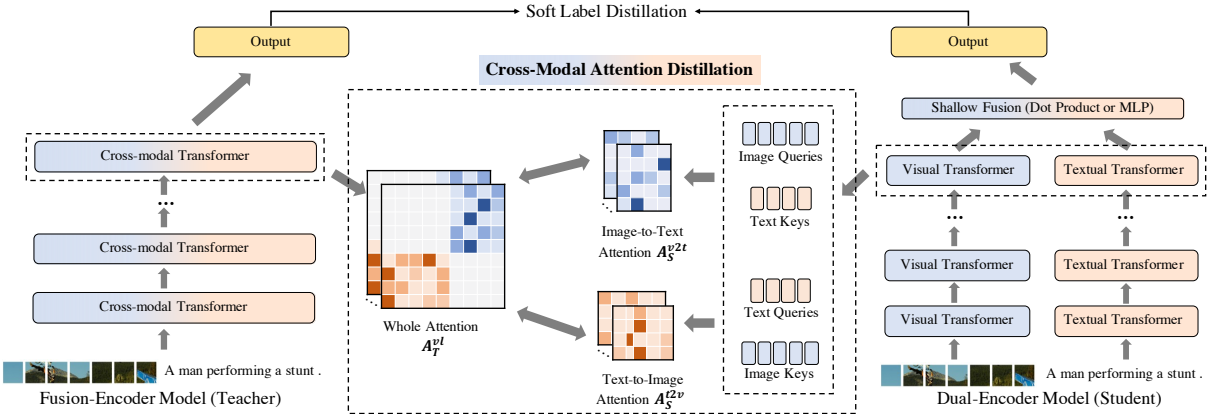


Figure 2: Overview of our framework DIDE, best viewed in color. Besides soft labels, we introduce cross-modal attention distillation to guide the model training. The visual-to-textual attention $\boldsymbol{A}^{v2t}$ (blue) and the textual-to-visual $\boldsymbol{A}^{t2v}$ (orange) of the dual-encoder model (student) are aligned to the fusion-encoder model (teacher). Other part of the attention distributions (grey) are omitted.

tention distillation objective. Specifically, during training, the student calculates the cross-modal part of the attention distribution and mimics it with the teacher. The cross-modal attention distributions of the student $\boldsymbol{A}_S^{v2t}$, $\boldsymbol{A}_S^{t2v}$ are computed as follows:

$$\boldsymbol{A}_S^{v2t} = \text{softmax}(\frac{\boldsymbol{Q}_S^v \boldsymbol{K}_S^{t\intercal}}{\sqrt{d_k}})$$
$$\boldsymbol{A}_S^{t2v} = \text{softmax}(\frac{\boldsymbol{Q}_S^t \boldsymbol{K}_S^{v\intercal}}{\sqrt{d_k}})$$

where $\boldsymbol{Q}_S^v, \boldsymbol{K}_S^v$ are visual queries and keys of the attention module in the student. $\boldsymbol{Q}_S^t, \boldsymbol{K}_S^t$ are queries and keys for textual features. To better align the student, we calculate the cross-modal attention $\boldsymbol{A}_T^{v2t}, \boldsymbol{A}_T^{t2v}$ of the teacher in a way similar to above, instead of directly splitting the whole attention $\boldsymbol{A}_T^{vl}$. We use the cross-modal attention distillation loss to minimize the KL-divergence of the cross-modal attention distribution:

$$\mathcal{L}_{CA} = D_{KL}(\boldsymbol{A}_S^{v2t} \parallel \boldsymbol{A}_T^{v2t}) + D_{KL}(\boldsymbol{A}_S^{t2v} \parallel \boldsymbol{A}_T^{t2v})$$

We empirically find that only distilling between the last layer of teacher and student is more effective (detailed in Section 4.4).

Soft Label Distillation In addition to cross-modal attention distillation, we also apply the soft label distillation loss to align the predictions between the teacher and the student:

$$\mathcal{L}_{SL} = D_{KL}(\boldsymbol{z}_S \parallel \boldsymbol{z}_T)$$

where $\boldsymbol{z}_S, \boldsymbol{z}_T$ are the output logits of the student and the teacher, respectively.

3.3 Two-Stage Distillation Training

DIDE applies the distillation objectives via the prevalent two-stage training paradigm: pre-training and then fine-tuning.

3.3.1 Pre-Training Distillation

We consider three typical pre-training tasks: cross-modal contrastive learning, image-text matching, and masked language modeling.

Cross-Modal Contrastive Learning (CMC) We introduce an InfoNCE contrastive loss (van den Oord et al., 2018) with in-batch negative sampling to optimize the shared space of visual and textual representations. Specifically, the image-to-text contrastive loss is computed as:

$$\mathcal{L}_{NCE}^{i2t} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(I_i, T_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(I_i, T_j)/\tau)}$$

where τ is a trainable temperature parameter, I_i and T_i are representations of the i-th image-text pair in the batch. We use the dot product as the $\text{sim}(\cdot, \cdot)$ function. Similarly, the text-to-image contrastive loss is computed as follows:

$$\mathcal{L}_{NCE}^{t2i} = -\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(T_i, I_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(T_i, I_j)/\tau)}$$

For the soft label distillation, the fusion-encoder model requires joint encoding of each image-text pair, which results in quadratic time complexity to obtain outputs. To reduce the training computation, we omit the soft label loss while only applying the cross-modal attention distillation on N matched pairs with gold labels:

$$\mathcal{L}^{CMC} = \mathcal{L}_{NCE}^{i2t} + \mathcal{L}_{NCE}^{t2i} + \mathcal{L}_{CA}^{CMC} \quad (1)$$

Datasets	NLVR2	SNLI-VE	VQA	Flickr30K
#Images	119K	31K	204K	32K
#Texts	100K	565K	1.1M	160K

Table 1: Statistics of downstream VL datasets.

Image-Text Matching (ITM) The goal of image-text matching is to predict whether the input image and text are matched. We employ cross-modal attention distillation loss over the input pairs and soft-label loss for training:

$$\mathcal{L}^{ITM} = \mathcal{L}^{ITM}_{CA} + \mathcal{L}^{ITM}_{SL}$$

Masked Language Modeling (MLM) Masked language modeling aims to recover the masked tokens from the other unmasked tokens. Similarly to ITM, the student needs to mimic both cross-modal attention distributions and soft labels from the teacher:

$$\mathcal{L}^{MLM} = \mathcal{L}^{MLM}_{CA} + \mathcal{L}^{MLM}_{SL}$$

3.3.2 Fine-Tuning Distillation

Vision-Language Understanding For VLU tasks, the student is fine-tuned with cross-modal attention and soft label distillation objectives:

$$\mathcal{L}^{VLU} = \mathcal{L}^{VLU}_{CA} + \mathcal{L}^{VLU}_{SL}$$

Image-Text Retrieval For image-text retrieval, the student is fine-tuned on the image-text retrieval task with the same objective as the CMC task (Equation 1).

4 Experiments

4.1 Datasets

We use four commonly used datasets for pre-training: COCO (Lin et al., 2014), Conceptual Captions (Sharma et al., 2018), SBU Captions (Ordonez et al., 2011) and Visual Genome (Krishna et al., 2017), with in total 4M images. We experiment on three vision-language understanding datasets and one image-text retrieval fine-tuning dataset. Table 1 shows the statistics of datasets.

Natural Language for Visual Reasoning The NLVR2 dataset (Suhr et al., 2019) is a visual reasoning task that aims to determine whether a textual statement describes a pair of images. Following previous work (Chen et al., 2020; Kim et al., 2021), we construct two pairs of image-text, each consisting of the image and a textual statement. The representations of the two pairs are fed into a classifier layer to obtain the final prediction.

Visual Entailment The SNLI-VE (Xie et al., 2019) is a three-way classification dataset, aiming to predict the relationship between an image and a text hypothesis: *entailment*, *natural*, and *contradiction*.

Visual Question Answering The task requires the model to answer questions based on the input image. We evaluate on the widely used VQAv2 (Goyal et al., 2017) dataset. Following Anderson et al. (2018), we formulate the problem as a classification task with 3,129 answer candidates.

Image-Text Retrieval The task consists of two subtasks: image retrieval and text retrieval. We experiment on the Flickr30K (Plummer et al., 2015) with the standard split (Karpathy and Fei-Fei, 2015).

4.2 Implementation Details

In the main experiments, we use ViLT (Kim et al., 2021) as our teacher due to its simplicity and effective performance. The visual and textual Transformers of the student model DIDE consist of 12-layer blocks with 768 hidden size and 12 attention heads. The intermediate size of feed-forward networks is 3072. Following Kim et al. (2021), the images are resized to 384×640 resolution and the patch size is 32×32. The maximum length of the text sequence is set to 40. We optimize DIDE with Adam (Kingma and Ba, 2015) using a batch size of 1024 for a total of 200K steps on 16 Nvidia V100 GPU cards. Note that our computation is less than the previous dual-encoder and fusion-encoder models. Refer to Appendix A for more details.

For the inference stage, we cache visual representations[4] for two reasons: (1) the averaged length of the visual tokens is longer than the textual tokens (240 *vs.* 40). (2) As shown in Table 1, an input image is combined with multiple text sentences. We reuse the cached visual representations with different text inputs to lower the inference latency.

4.3 Results

Vision-Language Understanding We compare DIDE with three types of vision-language pretrained models: (1) Dual-encoder models. CLIP (Radford et al., 2021) is pre-trained with image-text contrastive loss on 400M image-text pairs, significantly more than our pre-training data.

[4] Acturally, we can cache the visual or textual features according to the situation to improve efficiency.

Models	NLVR2		SNLI-VE		VQA	Inference
	dev	test-P	val	test	test-dev	Speedup
Fusion-encoder models without using object region features						
PixelBERT-R50	71.7	72.4	-	-	71.4	0.2×
ViLT (Teacher)	75.7	76.1	76.6	76.4	71.3	1.0×
Dual-encoder models without using object region features						
CLIP†	50.9	51.1	68.4	68.6	50.2	4.1×
SLIP†	50.9	51.1	70.9	71.0	55.9	4.1×
DeCLIP†	50.9	51.1	69.9	70.2	59.6	4.1×
DIDE(Ours)	**75.3**	**75.6**	**76.5**	**76.3**	**69.2**	4.0×
Using object region features from the object detector						
VisualBERT	67.4	67.0	-	-	70.8	$\ll$ 1.0×
LXMERT	74.9	74.5	-	-	72.4	$\ll$ 1.0×
UNITER-Base	75.9	75.8	78.6	78.3	72.7	$\ll$ 1.0×

Table 2: Results on vision-language understanding tasks. The results are averaged over 4 runs. We report vqa-score on VQA, accuracy for NLVR2 and SNLI-VE. † is our reimplementation of fine-tuning, which is the same as DIDE. We evaluate the inference speed of dual-encoder models and ViLT on the NLVR2 dataset with the same hyper-parameters. The inference speedup of other models is taken from Kim et al. (2021).

Models	NLVR2	SNLI-VE	VQA
Online inference time			
ViLT (Teacher)	150.3s	189.4s	1103.9s
DIDE	37.6s (4.0×)	49.7s (3.8×)	299.6s (3.7×)
Offline cache time			
DIDE	42.5s	10.6s	307.2s

Table 3: Averaged inference and cache time (in seconds) of our model and teacher model ViLT on three VLU datasets. The inference time and cache time are evaluated on a P100 GPU with a batch size of 32.

SLIP (Mu et al., 2021) and DeCLIP (Li et al., 2022) are improvements of CLIP. SLIP introduces self-supervised contrastive loss to CLIP. DeCLIP further leverages widespread supervision among the image-text pairs. For a fair comparison, we fine-tune them with the same range of hyperparameters as DIDE. (2) Fusion-encoder models without the object detector, such as our teacher ViLT (Kim et al., 2021). (3) Fusion-encoder models with the object detector, such as UNITER (Chen et al., 2020), that need a pretrained object detector to extract image region features, bringing more computational overhead.

Table 2 presents the performance of the VLU tasks. The results are averaged over four random seeds. DIDE achieves competitive performance compared to the ViLT teacher (retaining 99.3% in NLVR2, 99.9% in SNLI-VE, and 96.9% in VQA) while enjoying a 4 times speedup. DIDE significantly outperforms previous dual-encoder baselines (CLIP and its variants) by a large margin. It is worth mentioning that dual-encoder baselines only achieve chance-level accuracy on the complex visual reasoning dataset NLVR2, while our DIDE obtains promising results (from 50.9 to 75.3 points). To the best of our knowledge, it is the first demonstration that the dual-encoder model can obtain promising performance on the NLVR2 dataset. Compared to other fusion-encoder models (with or without the pretrained object detector), DIDE also obtains comparable or even better results in VLU tasks. This indicates that our approach can achieve a better efficiency-performance trade-off of the VLU.

Inference Speed We evaluate the latency of the student DIDE and the teacher ViLT in the batch inference setting, which is more favorable in low-latency scenarios (Zhang et al., 2019). For DIDE, we pre-compute visual representations offline and cache them. Table 3 shows the averaged inference time measured on the test split of the datasets. On all tasks, we obtain nearly 4 times online inference speedup. Even considering the offline cache time, DIDE is still faster than ViLT (1.9×~3.0×). Meanwhile, the storage cost is much cheaper than the cost of computing on GPUs (Cao et al., 2020). Thus, the dual-encoder model is more practical for production environments with massive inputs.

Image-Text Retrieval To explore the generalization of our DIDE framework beyond VLU, we conduct experiments on image-text retrieval. Table 4 reports the results of the Flickr30K dataset. Our model achieves substantial speedup with competitive performance compared to the teacher model

Models	Image Retrieval			Text Retrieval			Inference Speedup
	R@1	R@5	R@10	R@1	R@5	R@10	
ViLT (Teacher)	64.4	88.7	93.8	**83.5**	**96.7**	98.6	40398s
DIDE	**68.2**	**89.8**	**94.2**	83.2	**96.7**	**98.8**	16.1s (2509.2×)
−Cross-modal attention	66.6	89.2	93.4	81.6	95.6	98.4	-

Table 4: Retrieval results on the Flickr30K dataset. "−Cross-modal attention" is ablation trained without cross-modal attention distillation. The inference speed of our model and ViLT is evaluated under the same setup.

Models	Pre-training		Fine-tuning		NLVR2		SNLI-VE		VQA	Avg Δ
	STD	KD	STD	KD	dev	test-P	val	test	test-dev	
DIDE	✗	✓	✗	✓	**75.56**	**75.26**	**76.53**	**76.33**	**69.05**	-
Ablations	-	-	✓	✗	50.85	51.07	72.10	71.83	64.94	-12.40
	-	-	✗	✓	67.63	68.14	75.37	74.95	67.06	-3.94
	✓	✗	✓	✗	69.77	70.71	75.26	74.99	66.25	-3.15
	✓	✗	✗	✓	73.06	74.11	76.11	75.87	67.10	-1.30
	✗	✓	✓	✗	70.98	71.40	75.30	75.21	66.84	-2.60

Table 5: Ablation results on vision-language understanding tasks. "STD" denotes training with original ground truth labels. "KD" denotes the models trained using our distillation objectives.

Methods	NLVR2		SNLI-VE		VQA
	dev	test-P	val	test	test-dev
DIDE	**67.6**	**68.2**	**75.4**	**75.0**	**67.1**
- Soft Label	66.8	67.9	74.2	74.7	66.9
- Cross-model Attn	50.9	51.1	73.6	73.5	66.5
+ Hidden States	56.5	56.0	71.5	71.3	62.6
+ Uni-Modal Attn	64.8	65.6	74.8	74.8	66.6
+ Whole Attn	64.7	66.0	74.9	74.7	66.6

Table 6: Effects of using different knowledge distillation objectives. "Attn" is short for attention distributions. "Whole Attn" is the combination of "Uni-modal Attn" and "Cross-modal Attn".

ViLT. The student model DIDE even outperforms the ViLT teacher in image retrieval. Furthermore, removal of cross-modal attention distillation substantially harms performance on all metrics, showing that cross-modal attention distillation is also effective in image-text retrieval.

4.4 Analysis

Effects of Distillation in Training Stages We investigate the effect of applying our proposed distillation in the pre-training and fine-tuning stages. We compare with the baseline trained without distillation objectives, as the standard training.

Table 5 shows the evaluation results. Without the pre-training initialization, the models are directly initialized by the weights of pretrained ViLT. We can observe that without the pre-training stage, the performance significantly drops arcoss tasks. Under this setting, without the proposed distillation training, the model obtains a chance-level performance on NLVR2, similar to previous work (Kim et al., 2021; Shen et al., 2021). But our distillation method substantially improves the results, reducing the gap from -12.40 to -3.94. This indicates that our distillation method is better than standard training, even without pre-training. Furthermore, in the pre-training stage, the model still benefited from the distillation objectives compared with the standard training. Another interesting observation is that fine-tuning distillation brings more gains compared to pre-training distillation. This suggests that it is crucial for dual-encoder models to learn more ability of the task-specific cross-modal interactions. Overall, performing our proposed method in both pre-training and fine-tuning stages delivers the best performance across VLU tasks.

Effects of Different Distilled Knowledge We investigate the effects of different knowledge used in our framework. The compared ablations include training without soft label distillation (− soft label) or cross-modal attention distillation (− cross-modal attn). For more clarity, the dual-encoder student models are directly initialized by the pre-trained ViLT and fine-tuned with varying distillation objectives.

Table 6 illustrates the results of the VLU tasks. We find that both distillation objectives contribute to the success of the DIDE framework, while the proposed cross-modal attention distillation is more critical than soft label distillation. The student only trained with the soft label distillation objective only achieves random performance on NLVR2. We further incorporate other intermediate representations

Methods	NLVR2		SNLI-VE		VQA
	dev	test-P	val	test	test-dev
Last Layer (Ours)	**67.6**	**68.2**	**75.4**	**75.0**	**67.1**
Top-Layers Layerwise	66.0	67.3	75.2	74.8	66.8
Bottom-Layers Layerwise	63.5	63.0	75.1	74.7	66.6
All-Layers Layerwise	67.0	67.4	75.2	74.8	66.8

Table 7: Effects of different layer mapping strategies for our distillation method.

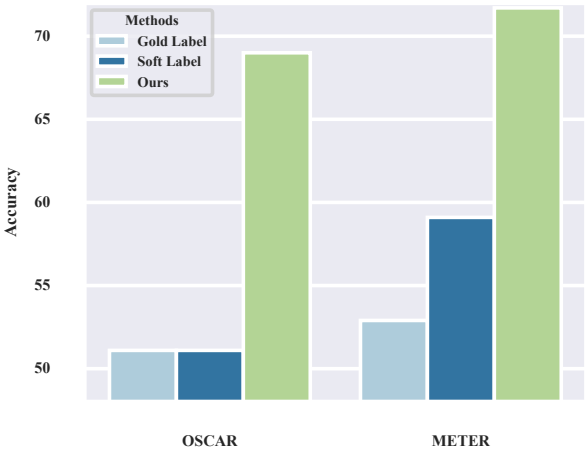


Figure 3: NLVR2 results of models initialized by OSCAR or METER. "Ours" means that besides the soft label distillation, the dual-encoder model also trained with our proposed cross-modal attention distillation.

of the teacher model, except for cross-modal attention. We observe that using attention distributions brings more gains across three tasks compared to the hidden states. Furthermore, we also explore which part of the attention distribution is more critical, *cross-modal attention* or *uni-modal attention*. As shown in Table 6, mimicking the teacher's cross-modal attention distributions achieves more improvements. Furthermore, we find that only the use of cross-modal attention distributions performs better than using the whole attention distributions (cross-modal + uni-modal). These observations validate our motivation that cross-modal interactions are more crucial for VLU tasks.

Effects of Different Layer Mapping Strategies for Distillation. To validate the effectiveness of distilling the knowledge of the teacher last layer for the student, we compare it with the layer-wise mapping strategy, including all layers, the upper part of layers, and the bottom part of layers. As shown in Table 7, last-layer strategy obtains better results. Furthermore, our strategy requires less computation than the layerwise methods. Thus, distillation in the last layer is more practical.

Effects on other VLP models. To evaluate the generalization of DIDE, in addition to ViLT (Kim et al., 2021), we also conduct experiments on NLVR2 with two other fusion-encoder vision-language pretrained models, OSCAR (Li et al., 2020) and METER (Dou et al., 2022). The architectures of two models are different from ViLT: Oscar applies the pretrained object detector to extract visual features. METER uses the visual encoder of CLIP to encode image patches and also applies RoBERTa (Liu et al., 2019) as textual encoder. The visual and textual representations are then fused by a multi-layer Transformer network.

We directly adopt the pretrained models to initialize the dual-encoder model and take the fine-tuned models as the teacher. We compare our method with the baselines supervised only by the gold labels of the original dataset and the soft label from the teacher model. Figure 3 illustrates the performance of the methods. For the models initialized by OSCAR and METER, we observe that only fine-tuning with gold labels or soft labels can not benefit the performance of the dual-encoder model on the complex visual reasoning task NLVR2. On the contrary, our method can improve performance by a large margin on both models, which is consistent with the results on ViLT. Our proposed cross-modal attention distillation shows effectiveness on different architectures of vision-language models.

5 Conclusion

On vision-language understanding tasks, fusion-encoder models obtain superior performance while sacrifice efficiency. In contrast, dual-encoder models have the advantage of efficiency, but previous models are insufficient to handle complex vision-language understanding. In this work, to obtain the efficiency-performance trade-off, we propose DIDE, a knowledge distillation framework for dual-encoder models to improve their performance on VLU tasks while retaining their efficiency. The key of DIDE is that we employ the cross-modal attention of a fusion encoder model as fine-grained supervision to guide the dual-encoder model for learning complex cross-modal interactions. Experimental results on several vision-language understanding tasks show that our DIDE achieves competitive performance with a four-time speedup over the fusion-encoder teacher model. Further analyses verify that distillation with cross-modal attention is critical for dual-encoder models.

Limitations

DIDE is pretrained on the publicly accessible resources consisting of 4M image-text pairs. We do not have enough computational resources to explore the situation with larger data, as used in CLIP (Radford et al., 2021). It is also interesting to combine our approach with other model acceleration methods summarized in Xu et al. (2021) to further toward Green AI (Schwartz et al., 2020).

Acknowledgements

We thank anonymous reviewers for their insightful feedback that helped improve the paper. We acknowledge Haoyang Wen, Yang Xu, and Yiheng Xu for helpful discussions. Zekun Wang, Haichao Zhu, Ming Liu, and Bing Qin are supported by the Science and Technology Innovation 2030 - "New Generation Artificial Intelligence" Major Project (2018AA0101901), the National Science Foundation of China (61976073, 62276083), Shenzhen Foundational Research Funding (JCYJ20200109113441941), the Project of State Key Laboratory of Communication Content Cognition (A02101), the Major Key Project of PCL (PCL2021A06). Ming Liu is the corresponding author.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086. Computer Vision Foundation / IEEE Computer Society.

Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT pre-training of image transformers. *CoRR*, abs/2106.08254.

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, and Hsiao-Wuen Hon. 2020. UniLMv2: Pseudo-masked language models for unified language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 642–652. PMLR.

Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020. Deformer: Decomposing pre-trained transformers for faster question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4487–4497. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Zewen Chi, Shaohan Huang, Li Dong, Shuming Ma, Saksham Singhal, Payal Bajaj, Xia Song, and Furu Wei. 2021. XLM-E: cross-lingual language model pre-training via ELECTRA. *CoRR*, abs/2106.16138.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural*

Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 13042–13054.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. 2022. An empirical study of training end-to-end vision-and-language transformers.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 1408–1418. IEEE.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12963–12971. AAAI Press.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *Trans. Assoc. Comput. Linguistics*, 9:570–585.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. 2020. Improving efficient neural ranking models with cross-architecture knowledge distillation. *CoRR*, abs/2010.02666.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3128–3137. IEEE Computer Society.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,*

ACL 2020, Online, July 5-10, 2020, pages 7871–7880. Association for Computational Linguistics.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021a. Align before fuse: Vision and language representation learning with momentum distillation. *CoRR*, abs/2107.07651.

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

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021b. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2592–2607. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2022. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. In *International Conference on Learning Representations*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2021. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9826–9836. Computer Vision Foundation / IEEE.

Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. 2021. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM*, 63(12):54–63.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 2556–2565. Association for Computational Linguistics.

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

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6418–6428. Association for Computational Linguistics.

Siqi Sun, Yen-Chun Chen, Linjie Li, Shuohang Wang, Yuwei Fang, and Jingjing Liu. 2021. Lightningdot: Pre-training visual-semantic embeddings for real-time image-text retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 982–997. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 10347–10357. PMLR.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

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

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *CoRR*, abs/1901.06706.

Jingjing Xu, Wangchunshu Zhou, Zhiyi Fu, Hao Zhou, and Lei Li. 2021. A survey on green deep learning. *CoRR*, abs/2111.05193.

Chengliang Zhang, Minchen Yu, Wei Wang, and Feng Yan. 2019. Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving. In *2019 USENIX Annual Technical Conference, USENIX ATC 2019, Renton, WA, USA, July 10-12, 2019*, pages 1049–1062. USENIX Association.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5579–5588. Computer Vision Foundation / IEEE.

A Details of Hyperparameters

For pre-training, visual and textual encoders of DIDE are initialized by the weights of the teacher model. We use Adam (Kingma and Ba, 2015) with $\beta_1 = 0.9, \beta_2 = 0.999$ for optimization. The learning rate is set to 1e-4, with the warm-up ration of 0.1, and linear decay. The weight decay is set to 0.01. For the ITM task, we replace the matched image with the probability of 0.5 to construct negative examples following previous work (Chen et al., 2020; Li et al., 2020; Kim et al., 2021), For the MLM task, we use 15% masking probability as in BERT (Devlin et al., 2019). For the downstream fine-tuning, we follow most of the hyperparameters in Kim et al. (2021). We fine-tune the model for 10 epochs with a batch size of 256 for VQA and

SNLI-VE. For NLVR2, we train the model for 20 epochs with a batch size of 128. For Flickr30k, the model is trained for 20 epochs with a batch size of 1024. We apply RandAugment (Cubuk et al., 2020) without color inversion and cutout.