TransferCVLM: Transferring Cross-Modal Knowledge for Vision-Language Modeling

Dongha Choi[1], **Jung-jae Kim**[2] and **Hyunju Lee**[1*]
[1]GIST Artificial Intelligence Graduate School, Gwangju, South Korea
[2]Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
dongha528@gm.gist.ac.kr, jjkim@i2r.a-star.edu.sg, hyunjulee@gist.ac.kr

Abstract

Recent large vision-language multimodal models pre-trained with huge amount of image-text pairs show remarkable performances in downstream tasks. However, the multimodal pre-training has limitations in terms of resources and training time when it comes to obtaining new models that surpass existing models. To overcome these issues, we propose TransferCVLM, a method of efficient knowledge transfer that integrates pre-trained uni-modal models (and cross-modal fusion-encoder) into a combined vision-language model (CVLM), without pre-training the CVLM with large amount of multimodal data, and then for each task application, fine-tunes the CVLM and transfers the multimodal knowledge of a teacher vision-language model to the CVLM by using knowledge distillation techniques. We demonstrate that 1) the fine-tuned CVLM performs comparable to other vision-language models of similar size, that 2) the multimodal knowledge transfer consistently enhances the CVLM, and the knowledge-transferred CVLM composed of large-size unimodal models outperforms the teacher multimodal model in most of downstream tasks, and that 3) TransferCVLM can also be used for model compression when using small-size unimodal models. We estimate that the training of TransferCVLM takes only 6% of pre-training of other vision-language models. Our code is available at https://github.com/DMCB-GIST/TransferCVLM.

1 Introduction

Ever since neural networks and deep learning proved successful in addressing visual and textual comprehension problems, numerous tasks have surfaced with the goal of comprehending and processing both images and text simultaneously, such as visual question answering, image-text retrieval, and visual entailment (Goyal et al., 2017; Plummer et al., 2015; Xie et al., 2019). Developing models proficient in these tasks is vital given that many real-world problems involve multiple modalities.

A trend in deep learning is self-supervised learning, and models pre-trained in self-supervised manner are predominant in most domains. There have been several approaches to pre-training multimodal models for the multimodal tasks, either pre-training a randomly initialized model from scratch (Singh et al., 2022; Wang et al., 2023b), or combining two unimodal pre-trained models and further pre-train the combined model (Kim et al., 2021; Zeng et al., 2022; Dou et al., 2022; Xu et al., 2023). Both approaches require to pre-train the models with unlabeled multimodal data, typically pairs of text and image, to learn cross-modal dependencies, but the pre-training with a large amount of unlabeled multimodal data is hugely costly. Moreover, for the latter approach, when there are new better unimodal pre-trained models (e.g. GPT-3 < GPT-3.5 < GPT-4) (Brown et al., 2020), it should repeat the costly pre-training.

To reduce the repetitive cost of large-scale multimodal pre-training of the second approach, we propose a method to transfer cross-modal dependencies from an existing multimodal pre-trained model to a combination of two new unimodal pre-trained models for a given task. With the proposed knowledge transfer, when there are better unimodal pre-trained models, we do not have to redo pre-training the models with a large amount of unlabeled multimodal data.

We present TransferCVLM, a method of efficient knowledge transfer from an existing vision-language pre-trained model (VLM; e.g. FLAVA (Singh et al., 2022)) to a combination of two unimodal pre-trained models for vision and language (e.g. BERT (Devlin et al., 2019), ViT (Dosovitskiy et al., 2020)). For this paper, we consider the two popular settings that the vision pre-trained model is

*Hyunju Lee is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 16733–16746

encoder-only, and the language pre-trained model is either encoder-only or encoder-decoder. We first build a new multimodal model (CVLM) by concatenating the encoders of the two unimodal pre-trained models and connecting the two encoders with, if any, the fusion encoder (FE) of the existing multimodal model (VLM). We then fine-tune both multimodal models (VLM, CVLM) with labeled multimodal data of a given task and finally transfer the cross-modal dependencies of the existing model (VLM; teacher) to the new model (CVLM; student) via a knowledge distillation technique.

Our contributions can be summarized as follows:

- We propose a novel knowledge transfer approach of transferring cross-modal knowledge from an existing multimodal pre-trained model (VLM) to a new model (CVLM) composed of unimodal pre-trained models, without multimodal pre-training of the new model CVLM. This may reduce the burden of multimodal pre-training when adapting unimodal pre-trained models for multimodal downstream tasks.
- TransferCVLM with base-size unimodal models: We show that the CVLM fine-tuned with labeled task data performs the task comparably to the existing vision-language pre-trained model VLM of similar size.
- TransferCVLM with large-size unimodal models: We show that the proposed multimodal knowledge transfer consistently enhances the fine-tuned CVLM, and that the knowledge-transferred CVLM composed of large-size unimodal models outperforms the teacher model VLM in most of downstream tasks.
- TransferCVLM with small-size unimodal models: We show that TransferCVLM may be used for model compression when using small-size unimodal models.
- We demonstrate the effectiveness of TransferCVLM on multiple vision-language downstream tasks, including visual question answering, visual entailment, image-text retrieval, and image captioning.

2 Methodology

We introduce the implementation of TransferCVLM first to build an encoder-only model in Sections 2.1 (CVLM) and 2.2 (knowledge transfer) and then an encoder-decoder model in Section 2.3. We also explain details of experiments about adapting TransferCVLM to build a small-size model in Section 2.4. Figure 1 depicts the overview of TransferCVLM for encoder-only model.

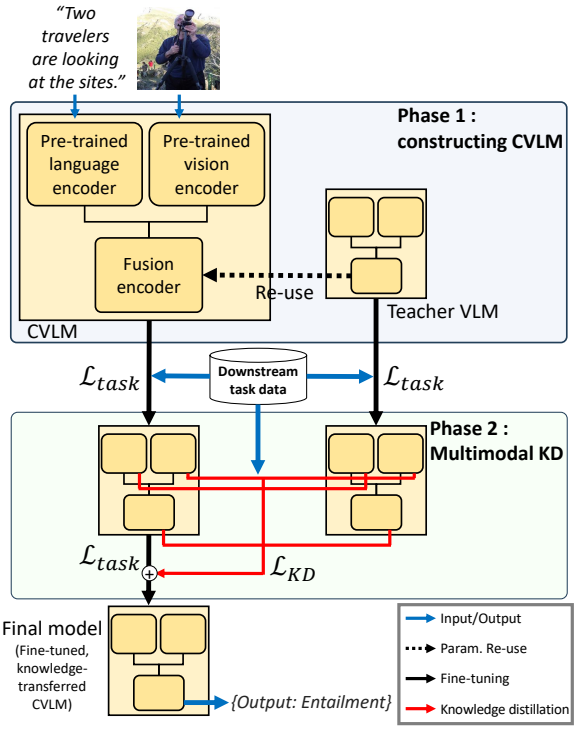

Figure 1: An overview of the TransferCVLM applied for encoder-only vision model and encoder-only language model, with an example of SNLI-VE task data.

2.1 Phase 1: Construction of Combined Vision-Language Model (CVLM)

In this step, we combine vision and language unimodal pre-trained models with the fusion encoder module of a pre-trained multimodal model to build a new multimodal model called Combined Vision-Language Model (CVLM). For an image-text pair dataset $X = \{(x_i^1, x_t^1), \ldots (x_i^m, x_t^m)\}$, the unimodal representations of the j-th image (x_i^j) and text (x_t^j) are encoded via the unimodal encoders. The two computed representations are then fed to a pre-trained fusion encoder E_F, which is taken out from a multimodal pre-trained model. Before that, we added a learnable linear projection to match the dimensions of the fusion encoder input (d_F) and the dimensions of the unimodal representations, which can be varied depending on model selection. Formally, the projected unimodal representations

of the j-th image (x_i^j) and text (x_t^j) are as follows:

$$\begin{cases} H_V^j = \text{projection}_V(E_V(x_i^j)) \\ H_L^j = \text{projection}_L(E_L(x_t^j)), \end{cases} \quad (1)$$

where E_V and E_L are vision and language pre-trained encoders, respectively, and $H_V^j \in \mathbb{R}^{l_V \times d_F}, H_L^j \in \mathbb{R}^{l_L \times d_F}$. We assume that both encoders are transformer-based, and thus the outputs of the encoders have their respective sequence lengths l_V and l_L determined by the model input settings.

Finally, these two projected representations are combined with the multimodal classification token $t_{[CLS]}$ and fed to the fusion encoder module:

$$H_F^j = E_F(\mathbf{C}(t_{[CLS]}, H_V^j, H_L^j)), \quad (2)$$

where $H_F^j \in \mathbb{R}^{(1+l_V+l_L)\times d_F}$. $\mathbf{C}(\cdot)$ denotes the combination scheme according to the fusion method (e.g., concatenation, cross-attention). Note that we use the final output representation of the first token $t_{[CLS]}$, also known as classification embedding, as the input of the task-specific head for downstream discriminative tasks (visual entailment, visual question answering, and image-text retrieval).

In practice, multimodal pre-trained models have slightly different structures from each other. For instance, both FLAVA (Singh et al., 2022) and X-VLM (Zeng et al., 2022) consist of two unimodal encoders and one fusion encoder, but their fusion encoders are different such that FLAVA's fusion encoder is a simple self-attention encoder, while X-VLM's fusion encoder incorporates 'asymmetric' cross-attention (Hendricks et al., 2021), receiving the vision encoder output as key and value and the language encoder output as query. Note that the fusion encoder of FLAVA is 'symmetric', not differentiating between vision and language encoder outputs. Thus, when we construct CVLM with FLAVA and X-VLM as teacher models, we connect vision and language encoders with the copy of the corresponding teacher's fusion encoder.

2.2 Phase 2: Multimodal Knowledge Transfer

The CVLM constructed in the previous step can be fine-tuned and used for downstream tasks like other multimodal pre-trained models. Since the two unimodal backbones already have an accomplished ability to understand their respective modalities, and the fusion module also has initial knowledge for merging multimodal representations, CVLM shows high performance only with the fine-tuning without additional pre-training.

However, the fine-tuning of CVLM only using downstream task data suffers from the limited quantity of labeled task data, which might be insufficient to learn the dependencies (or alignments) among the three modules of two unimodal backbones and fusion encoder. To address this issue, we apply a knowledge distillation technique to transfer knowledge about the interaction between modalities and representation alignment. Our approach is inspired by the DoKTra framework (Choi et al., 2022), which applies knowledge distillation for domain knowledge transfer from an existing domain-specific pre-trained language model to a new generic pre-trained language model. We modify the DoKTra framework for multimodal knowledge transfer from an existing multimodal pre-trained model to the CVLM with unimodal pre-trained models in order to distil the the knowledge about intermodality interactions and representation alignment, as described below.

First, we set the pre-trained multimodal model (FLAVA, X-VLM) as the teacher and CVLM as the student. To provide both models with initial knowledge about the target downstream task, we fine-tune each model on the training data of the task using task-specific loss $\mathcal{L}_{task}$, such as cross entropy, binary cross entropy with logits, or contrastive loss.

After initial fine-tuning of the teacher model and CVLM, knowledge distillation is applied to transfer the teacher's multimodal understanding knowledge to the CVLM. We set the final hidden state of the first token of each of the three modules (vision and language encoders, and fusion encoder) as the distillation target (that is, [CLS] embeddings). This is because the classification embedding serves as an encoding that condenses the entire input into a single vector and is commonly used in discriminative subtasks. Our distillation objective $\mathcal{L}_{KD}$ is formulated as follows:

$$\mathcal{L}_{KD}(\mathbf{x}^j) = \mathcal{L}_{dist}(S_V^j, T_V^j) + \mathcal{L}_{dist}(S_L^j, T_L^j) + \mathcal{L}_{dist}(S_F^j, T_F^j), \quad (3)$$

where $\mathcal{L}_{dist}$ is a distance metric and S_V^j, S_L^j, S_F^j are the classification embeddings of student model's vision, language and fusion encoders, respectively. Note that S_V^j, S_L^j, S_F^j are the [CLS] embeddings of

H_V^j, H_L^j, H_F^j from Eq. (1) and (2) and that knowledge distillation is performed using the projected outputs after matching the output dimensions of the unimodal encoders to the input dimension of the fusion encoder. Similarly, T_V^j, T_L^j, T_F^j are the classification embeddings of teacher model's encoders. In this study, we use mean squared error (MSE) as distance metric, while DoKTra employed activation boundary transfer loss (Heo et al., 2019). Please see Appendices C and D for more details.

Our knowledge distillation method utilizes classification embeddings instead of classification logits. However, using only the above distillation objective for a fine-tuned model can cause catastrophic forgetting (McCloskey and Cohen, 1989) about the output probability distribution calculation. To prevent this, we design the loss to minimize both distillation objective and task-specific loss simultaneously in the knowledge distillation procedure. The final objective for knowledge distillation between the teacher and CVLM can be summarized as follows:

$$\mathcal{L}_{KT} = \alpha \mathcal{L}_{KD} + (1-\alpha)\mathcal{L}_{task}, \tag{4}$$

where $\alpha \in [0, 1)$ is a hyperparameter to balance between the two loss functions.

2.3 TransferCVLM for Encoder-Decoder Models

To perform generative tasks such as image captioning, VLMs incorporating transformer decoder modules are also being introduced (Cho et al., 2021; Wang et al., 2022; Alayrac et al., 2022). We select GIT (Wang et al., 2022) as another alternative teacher model for TransferCVLM to verify the feasibility of extending to decoder modules and generative tasks. GIT consists of a single image encoder and a single text decoder, without a fusion module. Therefore, we construct the student model using a unimodal image encoder and a decoder-only language model without reusing teacher parameters, and applied phase 2 of TransferCVLM. In generative tasks, the representation of all tokens is utilized for task execution, unlike discriminative tasks. Therefore, we made slight modifications to phase 2 by pooling the representations of all generated and ground-truth tokens and setting them as the targets for knowledge distillation.

Formally, in decoder-based CVLM, the representations of each module are as follows:

$$\begin{cases} H_V^j = \text{projection}_V(E_V(x_i^j)) \\ H_L^j = \text{emb}_L(x_t^j) \\ H_D^j = D_L([H_V^j; t_{[BOS]}; H_L^j; t_{[EOS]}]), \end{cases} \tag{5}$$

where E_V is a vision pre-trained encoder, emb_L is an embedding of language model, and D_L is a pre-trained language decoder model. Similar to the encoder-based TransferCVLM, the distillation objective is formulated as follows:

$$\mathcal{L}_{KD}(\mathbf{x}^j) = \mathcal{L}_{dist}(S_V^j, T_V^j) + \mathcal{L}_{dist}(S_{DV}^j, T_{DV}^j) + \mathcal{L}_{dist}(S_{DL}^j, T_{DL}^j), \tag{6}$$

where S, T are pooled representations of the student and teacher, respectively. S_{DV}, S_{DL} refer to the pooled representations of the regions corresponding to the image and text, respectively, in the CVLM's decoder output. Each region was average-pooled separately, formulated as follows.

$$S_{DV}^j = \frac{\sum_{k=1}^{l_V} H_D^j(k)}{l_V}, S_{DL}^j = \frac{\sum_{k=l_V+2}^{l_V+l_L+2} H_D^j(k)}{l_L}. \tag{7}$$

Teacher's pooled representations T_{DV}^j, T_{DL}^j can also be calculated in the same manner as described above.

2.4 Extensibility of TransferCVLM for Model Compression

In this section, we explain the adaptation of TransferCVLM for multimodal model compression. Recent studies have proposed multimodal model compression based on knowledge distillation (Fang et al., 2021; Wang et al., 2023a), but they require costly pre-training on extensive unlabeled data (see Section 4.2 for details). However, TransferCVLM does not require such large-scale pre-training and can be used for model compression if we select small backbones for CVLM construction.

For model compression, we revise Phase 2 (called Phase 2^{MC}, MC stands for Model Compression.) to perform knowledge distillation from a 'large' CVLM, fine-tuned and knowledge-distilled by Phases 1 and 2 (teacher), to a 'small' CVLM fine-tuned by Phase 1 (student). This approach is motivated by the observation that Phases 1 and 2 of TransferCVLM can produce a student model

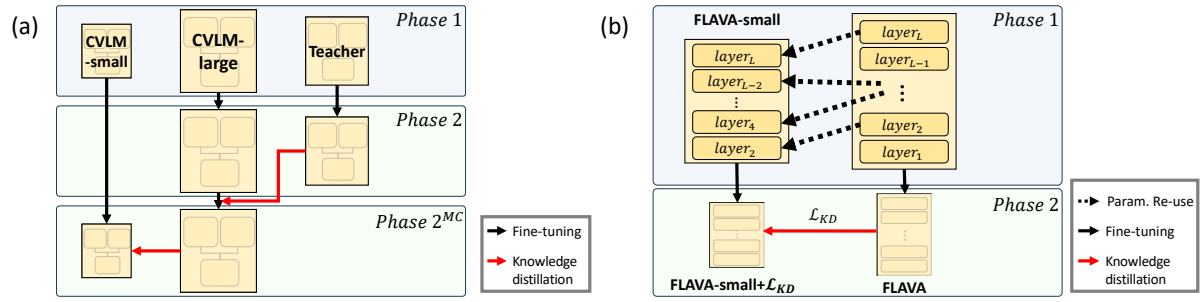

Figure 2: Visualization of (a) revised TransferCVLM for model compression and (b) construction of FLAVA-small for comparison.

that outperforms the teacher model, especially with large backbones (see Section 3.3 for details).

Since there is no prior work on model compression via knowledge distillation for multimodal models without pre-training a student model, we also devised a robust baseline method for comparison. We constructed a baseline student model by reducing the parameters of the teacher model, excluding odd-numbered layers from each module, resulting in a model nearly half the size. After fine-tuning with task-specific data (Phase 1), we applied the same knowledge distillation method as TransferCVLM (Phase 2) to gain additional insights from the fine-tuned teacher model.

Figure 2 illustrates the overview of revised TransferCVLM for model compression and construction of the baseline for comparison.

3 Experiments

3.1 Downstream Tasks

We evaluated our proposed methodology on various types of tasks. For classification tasks, we selected visual question answering using VQAv2 (Goyal et al., 2017) and visual entailment using SNLI-VE (Xie et al., 2019). For the image-text retrieval task, we used the Flickr30k (Plummer et al., 2015) dataset. Additionally, for the decoder-based models, we utilized Flickr30k for generative captioning as well. We also used COCO (Lin et al., 2014) for generative captioning task. We used accuracy as the evaluation metric for classification, Recall@1 and 5 for retrieval, and CIDEr (Vedantam et al., 2015) for captioning. As the task-specific losses ($\mathcal{L}_{task}$), we used binary cross-entropy for VQAv2, cross-entropy for SNLI-VE and captioning, and contrastive loss and matching loss, both based on cross-entropy, for image-text retrieval. Please refer to Appendix A for the detailed explanation and statistics of the downstream task datasets.

3.2 Implementation Details

For the construction of encoder-only CVLM, we select encoder part of T5 (Raffel et al., 2020) as the language backbone and Swin Transformer (Liu et al., 2021) as the vision backbone. Swin provides pre-trained models of various patch sizes, window sizes, and image sizes, and we use the `-patch4-window12-384` models. The maximum sequence length of the language encoder is set to 24. We use `-base` or `-large` as the size of the unimodal backbones. For the encoder-decoder CVLM, we combined the decoder-only language model GPT-2-`small` or `-medium` (Radford et al., 2019) with Swin.

PyTorch and Transformers were used for the implementation of all experimental stages. Our experiments were performed with NVIDIA RTX 3090 24GB GPU for base size models and NVIDIA A100 40GB GPU for large size models. Please refer to Appendix E for details about hyperparameter search.

3.3 Experimental Results

Table 1 presents the main experimental results of TransferCVLM on the three multimodal downstream tasks. Since recent developments in vision-language models and their knowledge distillation have primarily focused on pre-training, there are no methodologies available for fair comparison to ours. Therefore, to enable performance comparison, we designed baseline models under the same conditions as our methodology, utilizing only one vision-language pretrained encoder and two unimodal pretrained encoders. We established three baselines, incrementally increasing the model size. The first baseline involves concatenating two unimodal outputs directly for classification (UE concat). The second baseline employs the same architecture as our CVLM but initializes the fusion

Model	# Params	VQAv2	SNLI-VE	Flickr30k-TR		Flickr30k-IR	
				R@1	R@5	R@1	R@5
FLAVA (Teacher)	241M	68.34±.24	78.38±.25	60.90±.14	83.40±.08	46.92±.18	75.32±.11
UE concat	200M	60.37±.17	72.21±.14	-	-	-	-
CVLM w/ random FE	241M	68.57±.15	75.63±.25	-	-	-	-
UE+ME concat	442M	62.44±.31	78.92±.05	-	-	-	-
TransferCVLM-base-phase 1	241M	69.51±.66	75.80±.19	62.80±.35	87.23±.32	51.05±.40	80.54±.34
TransferCVLM-base-phase 2		71.66±.12	78.52±.05	66.90±.14	88.50±.11	52.88±.18	81.34±.11
TransferCVLM-large-phase 1	575M	70.43±.99	76.38±.40	72.30±.46	92.20±.41	59.84±.47	86.14±.39
TransferCVLM-large-phase 2		**72.77±.02**	**79.61±.17**	**73.30±.10**	**93.20±.09**	**59.96±.14**	**86.30±.11**

Table 1: Main experimental results of TransferCVLM on three vision-language downstream tasks. The results that outperform the teacher (FLAVA) are underlined. (TR: Image-Text Retrieval, IR: Text-Image Retrieval)

encoder randomly (CVLM w/ random FE). The last approach involves concatenating the outputs of three encoders (two unimodal and one multimodal) and feeding them to the classifier (UE+ME concat). Please refer to the appendix B for the visualized architecture of each baseline model. Note that our baseline designs are only feasible for classification tasks (VQAv2 and SNLI-VE), and therefore we do not have experimental results for retrieval task. We denote the construction phase of CVLM as Phase 1 and the knowledge distillation phase as Phase 2. Below are our findings from the experiment results.

First, the fine-tuned CVLM (Phase 1) shows performance compatible with the teacher model FLAVA. The fine-tuned CVLM with large-sized backbones even outperforms FLAVA. This outcome suggests that enhanced unimodal comprehension directly benefits multimodal downstream tasks without further pre-training for unsupervised multimodal alignment. Please note that the performance on VQAv2 reported in the FLAVA paper slightly differs from our fine-tuning results. This discrepancy is primarily due to a difference in image resolution between the model used in the paper and the model that was actually released.

Second, the performance of both base-sized and large-sized CVLM could be enhanced by an average of 1.73% through the application of knowledge distillation (Phase 2). As a result, the enhanced base-sized CVLM by Phase 2 outperforms the teacher model on every task, where the base-sized CVLM has the same model size as the teacher. This result is noteworthy because unlike general knowledge distillation, where the final performance of the student model is reportedly bound to the teacher's performance, TransferCVLM can produce a student that surpasses the teacher as keeping the original strength of the student (Phase 1) and synergistically distilling additional knowledge of cross-modal dependencies from the teacher to the student. Furthermore, the consistent improvement of both sized CVLMs by Phase 2 suggests that TransferCVLM might be effective for a CVLM of even a larger size, as larger-size unimodal pre-trained models continue to emerge.

Model	VQAv2	SNLI-VE	Avg.
X-VLM	74.40	80.02	77.21
T-CVLM$_{\text{X-VLM}}$-base-phase 1	70.40	76.71	73.56
T-CVLM$_{\text{X-VLM}}$-base-phase 2	73.36	79.14	76.25
T-CVLM$_{\text{X-VLM}}$-large-phase 1	71.20	77.33	74.27
T-CVLM$_{\text{X-VLM}}$-large-phase 2	**74.56**	**79.92**	**77.24**

Table 2: Experimental results of X-VLM based TransferCVLM. T-CVLM refers to TransferCVLM.

Model	Flickr30k	COCO
GIT-base	78.58	123.35
T-CVLM$_{\text{GIT}}$-small-phase 1	72.67	120.24
T-CVLM$_{\text{GIT}}$-small-phase 2	77.36	122.64
T-CVLM$_{\text{GIT}}$-medium-phase 1	75.74	121.15
T-CVLM$_{\text{GIT}}$-medium-phase 2	**79.91**	**123.88**

Table 3: Experimental results of GIT based TransferCVLM on two captioning datasets.

Additionally, TransferCVLM consistently outperformed all three types of baseline models. This suggests that TransferCVLM is the optimal choice when employing one multimodal model and two unimodal models. Note that the UE+ME concat model is a type of ensemble model and comes with roughly double the parameters, although it exhibited slightly better results than TransferCVLM in SNLI-VE.

We also conducted an ablation study on the initialization of the fusion encoder in CVLM, the selection of the distillation objective function and distillation target. The experimental results indi-

Model	# Params	VQAv2	SNLI-VE	Avg.	Retain(%)
FLAVA (Teacher)	241M	68.34±.24	78.38±.25	73.36	
FLAVA-small		56.78±.45	71.07±.13	63.93	
FLAVA-small + MSE	135M	66.27±.01	74.83±.24	70.55	96.17
FLAVA-small + ABT		65.84±.02	75.93±.17	70.89	96.63
T-CVLM-small-phase 1		64.79±.17	72.31±.19	68.55	
T-CVLM-small-phase 2	128M	67.83±.20	**76.18±.11**	72.01	98.16
T-CVLM-small-phase 2^{MC}		**68.26±.04**	75.98±.06	**72.12**	**98.31**

Table 4: Comparisons of task-specific multimodal model compression methods. The column 'Retain' indicates the ratio of the average performance of the corresponding model over the teacher model performance.

cate that our choice of reusing the parameters of the teacher model and selecting MSE yield the best results. For detailed information and results of the ablation study, please refer to Appendix C.

3.4 Flexibility in Model Architecture Selection

Table 2 presents evaluation results of X-VLM-based TransferCVLM. Firstly, the CVLM with cross-attention (Phase 1) exhibited better performance than the FLAVA-based CVLM in Table 1. This suggests that using not only advanced unimodal encoders but also advanced fusion encoders can contribute to the performance of fine-tuned CVLM. Secondly, we found that multimodal knowledge distillation (Phase 2) for the X-VLM-based CVLM could lead to a significant performance improvement of 2.84% in average. Thirdly, large-size model of X-VLM-based TransferCVLM slightly outperformed the teacher model on the VQAv2 task, though the other models did not outperform the teacher like FLAVA-based TransferCVLM. We leave it as future work to study the relationship between teacher model architecture and the effect of TransferCVLM on different downstream tasks.

Table 3 illustrates the results of the GIT-based TransferCVLM on the two captioning tasks. TransferCVLM exhibited substantial performance improvements of decoder-based models for the generative task. Remarkably, it consistently enhanced performance regardless of the student model's size, and medium-sized models with sufficiently robust initial performance even outperformed the teacher when TransferCVLM was applied. Considering that the GPT-2 model we utilized as the backbone is relatively outdated, this suggests the potential for achieving significantly higher performance by employing more advanced text decoders.

3.5 Extensibility of TransferCVLM for Model Compression

In this section, we demonstrate the effectiveness of TransferCVLM as a model compression method (Section 2.4). We applied TransferCVLM with small-sized unimodal backbones, namely `T5-small` and `Swin-small-patch4-window7-224`, and compared it with the baseline model described in Section 2.4. As for the baseline, we also compared two distance metrics, Mean Squared Error (MSE) and activation boundary transfer (ABT) (Heo et al., 2019) (see Appendices C and D for details).

Table 4 presents the evaluation results of the two compression methods on the two vision-language tasks. Even though both methodologies undergo a similar number of training stages, including teacher training, student training, and knowledge distillation, TransferCVLM with small-sized backbones outperformed the baseline. Moreover, the results of phase 2^{MC} suggest the broad applicability of TransferCVLM-phase 2 models, which are not only well-performing models but also better teachers for TransferCVLM-small models.

3.6 Efficiency Evaluation

To quantitatively evaluate the efficiency of TransferCVLM, we estimated and compared the relative training time of conventional vision-language pre-training and TransferCVLM. We set the goal of the two methods as obtaining a base-sized vision-language model with T5 and Swin for the VQAv2 task. We computed the relative training time under the following assumptions: 1) All iterations take the same time. 2) All subtasks of the two methods take the same time per iteration if performed without gradient accumulation. 3) Training with a large batch size in our computing environment requires gradient accumulation, and the time increases linearly with the number of steps in gradient accumu-

Method	Stage	# of training data	Required batch size	Relative training time per iteration	# of iterations	Relative training time per subtask	Total relative training time
VLP	Pre-training	9M	4096	256	100k	25.6M	26.22M
	Fine-tuning	658k	160	10	62k	0.62M	
T-CVLM	CVLM fine-tuning	658k	160	10	62k	0.62M	1.644M
	Teacher fine-tuning		128	8	51k	0.408M	
	Knowledge distillation		64	4	154k	0.616M	

Table 5: Efficiency evaluation results for comparison between vision-language pre-training and TransferCVLM in terms of relative training time. (VLP: Vision-Language Pre-training)

lation per iteration. For instance, if the time taken for one iteration without gradient accumulation is 1, it is 256 with 256 steps of gradient accumulation.

We performed efficiency experiments using an RTX 3090 24GB GPU, with a maximum batch size of 16. We estimated the time per iteration without gradient accumulation by running the fine-tuning subtask for a few iterations. For vision-language pre-training, we used the hyperparameters set by METER (Dou et al., 2022). The hyperparameters for the other stages were based on previous experiments.

Table 5 displays the approximated training time for the two methods. As indicated, TransferCVLM can obtain a vision-language model for VQAv2 approximately 16 times faster than pre-training. The primary reason for the inefficiency of pre-training is its need for a sufficiently large batch size for stable training. Several recent vision-language models require a batch size of at least 1024 for their pre-training (Dou et al., 2022). Although this can be addressed with advanced computational devices, it remains impractical for general environments with only a few GPUs. Furthermore, handling and preprocessing over 9 million image-text pair data is labor-intensive. In contrast, TransferCVLM is suitable for general devices since it only requires a reasonable batch size and task-specific data.

4 Related Works

4.1 Vision-Language Pre-training

Recently, many pre-trained multimodal models capable of processing both modalities' inputs simultaneously have been developed. CLIP (Radford et al., 2021) is a prominent example that improves visual understanding through language supervision rather than performing multimodal tasks. CLIP employs contrastive learning, aligning images and text within the same embedding space. This trains the model to group similar images and text closely and separate dissimilar ones. Contrastive learning, often combined with image-text matching and masked language modeling objectives, is standard in vision-language pre-training, typically utilizing large-scale datasets such as COCO (Lin et al., 2014), Visual Genome (Krishna et al., 2017), and SBU captions (Ordonez et al., 2011).

There have been various attempts on model architectures for modality fusion, which can be grouped into categories: early fusion (Kim et al., 2021; Bao et al., 2022; Wang et al., 2023b), late fusion (Dou et al., 2022; Singh et al., 2022), intermediate fusion (Xu et al., 2023), early-late fusion (Driess et al., 2023; Li et al., 2023; Wang et al., 2022), and intermediate-late fusion (Gao et al., 2023). The proposed method is developed for late fusion and early-late fusion models as student models. For late fusion model (FLAVA, X-VLM), each modal data is processed with a large unimodal model, combined with the other modal processed data, and passed to a cross-modal encoder. For early-late fusion (GIT), an image is processed with a unimodal image encoder, and passed to a text decoder that simultaneously performs two roles: language encoding and multimodal fusion. However, it can be adapted for other fusion methods when the teacher and student models have similar structures. Knowledge distillation can be applied if the teacher and student models have corresponding vision, language, and fusion representations, even with different dimension sizes. We leave adaptation of TransferCVLM for the other fusion models as future work.

4.2 Knowledge Distillation for Multimodal Models

Knowledge distillation (Hinton et al., 2015) is a technique for transferring knowledge typically from a large model (called teacher) to a small model (called student), and is mainly used for model compression. In general, knowledge is transferred through the student learning to imitate the teacher's logit or intermediate representation. Re-

cently, various approaches to apply knowledge distillation to multimodal models have been proposed. Fang et al. (2021) proposed DistillVLM, which compresses vision-language models using knowledge distillation during pre-training, enhancing smaller models with additional knowledge from the teacher. Wang et al. (2023b) also proposed knowledge distillation-based model compression method for vision-language models. They incorporated several knowledge distillation objectives to enhance the pre-training of a small model, called EfficientVLM. Our method TransferCVLM also transfers knowledge between multimodal models, but is different from the two methods in that TransferCVLM does not involve any further pre-training.

Approaches to improving unimodal understanding through the knowledge distillation from multimodal models were also proposed. Kim et al. (2022) transferred the representation of a pretrained multimodal model to a small vision model. A technique called cross-modal similarity matching was added to efficiently utilize the characteristic of image-text alignment. Similarly, Tang et al. (2021) proposed a technique for distilling knowledge from video-language models to improve the understanding of language models. Please note that the proposed student model consists of pre-trained unimodal models and a cross-model encoder, but the knowledge transfer is mainly for learning inter-modal dependencies between the unimodal models.

5 Conclusion

We present a novel knowledge transfer method TransferCVLM for quickly building enhanced multimodal models of CVLM by using latest unimodal pre-trained models and by distilling cross-modal knowledge from a teacher of multimodal pre-trained model, without any pre-training with large amount of multimodal data. The proposed method consistently improves performance of CVLM across downstream tasks and across model architectures, even outperforming the teacher model in many cases when using large-size unimodal pre-trained models. When using small-size unimodal pre-trained models, TransferCVLM can be used as model compression method without large-scale pre-training.

Limitations

Here, we summarize the limitations of TransferCVLM to inspire future works.

Firstly, we explored only a limited number of models and tasks. There is a need for validation on more multimodal discriminative tasks, such as NLVR2 (Suhr et al., 2018) or HatefulMemes (Kiela et al., 2020), which are commonly used to evaluate the performance of vision-language models. Additionally, experiments with emerging billion-scale and advanced unimodal backbones should be conducted. Furthermore, future work could consider validation for various fusion encoders, including multiway transformers (Bao et al., 2022), bridge layers (Xu et al., 2023), and other diverse fusion encoders.

Secondly, our methodology has limitations in multimodal fusion methods. Since our methodology assumes the use of unimodal backbones, it is only applicable to models employing the late fusion or early-late fusion method introduced in Section 4.1. Future work could explore variations that are applicable across various fusion methodologies, such as early fusion and intermediate-late fusion.

Acknowledgements

This research was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (2021R1A2C2006268) and Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) [No.2019-0-01842, Artificial Intelligence Graduate School Program (GIST)]. We appreciate the high-performance GPU computing support of HPC-AI Open Infrastructure via GIST SCENT.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. Vlmo:

Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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 *European conference on computer vision*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Dongha Choi, HongSeok Choi, and Hyunju Lee. 2022. Domain knowledge transferring for pre-trained language model via calibrated activation boundary distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1658–1669.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. In *arXiv preprint arXiv:2303.03378*.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1428–1438.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

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. *Transactions of the Association for Computational Linguistics*, 9:570–585.

Byeongho Heo, Minsik Lee, Sangdoo Yun, and Jin Young Choi. 2019. Knowledge transfer via distillation of activation boundaries formed by hidden neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3779–3787.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Byoungjip Kim, Sungik Choi, Dasol Hwang, Moontae Lee, and Honglak Lee. 2022. Transferring pre-trained multimodal representations with cross-modal similarity matching. *Advances in Neural Information Processing Systems*, 35:30826–30839.

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

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.

Tsung-Yi Lin, Michael Maire, Serge 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 13*, pages 740–755. Springer.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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 *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Zineng Tang, Jaemin Cho, Hao Tan, and Mohit Bansal. 2021. Vidlankd: Improving language understanding via video-distilled knowledge transfer. *Advances in Neural Information Processing Systems*, 34:24468–24481.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Tiannan Wang, Wangchunshu Zhou, Yan Zeng, and Xinsong Zhang. 2023a. EfficientVLM: Fast and accurate vision-language models via knowledge distillation and modal-adaptive pruning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13899–13913, Toronto, Canada. Association for Computational Linguistics.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023b. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186.

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

Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, pages 10637–10647.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR.

Appendices

A Details and Statistics of the Downstream Task Datasets

VQAv2 (Goyal et al., 2017) is a visual question-answering task frequently used in the evaluation

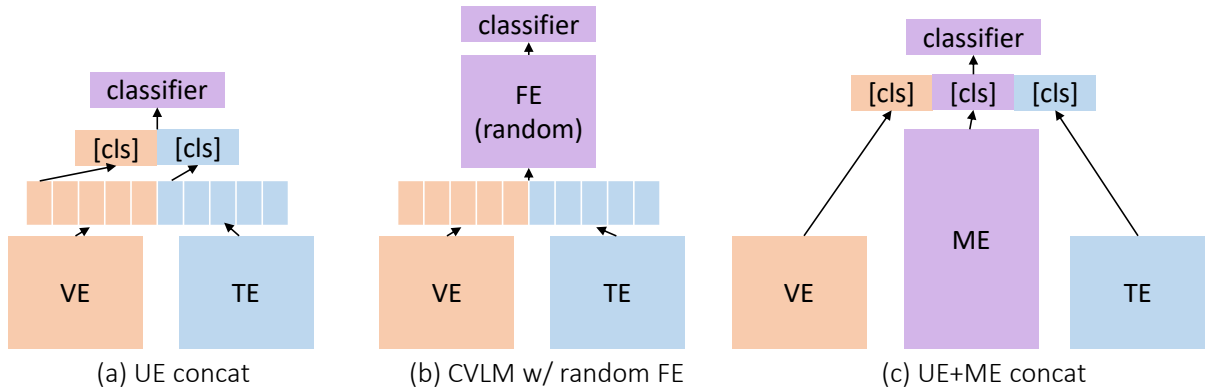

Figure B1: Model architecture of the baseline models. (VE: Vision Encoder, TE: Text encoder, FE: Fusion Encoder, ME: Multimodal Encoder)

Split	VQAv2	SNLI-VE	Flickr30k	COCO
Train	443,757	529,527	145,000	566,747
Dev	214,354	17,858	1,014	5,000
Test	447,793	17,901	1,000	5,000

Table A1: The statistics of the downstream task datasets

of vision-language models. VQAv2 consists of images and corresponding questions, and the model must answer the questions based on the given images. Following previous studies (Anderson et al., 2018; Kim et al., 2021; Chen et al., 2020), we converted VQAv2 task into a multi-label classification for training and used the trained model to select the most likely answer from 3129 most frequent answer candidates for inference. In addition, after searching for hyperparameters using the development set, we train the model with the combination of the training set and the development set, and report the evaluation results for the test-dev set.

SNLI-VE (Xie et al., 2019) is an inference task to predict multimodal entailment. SNLI-VE requires the model to predict whether a premise image semantically entails a given hypothetical text. Similar to the existing natural language inference tasks, the model classifies the input hypothetical text as one of entailment, neutral, or contradiction.

Flickr30k (Plummer et al., 2015) consists of 31,000 images, each with 5 human-annotated captions, and we used the commonly utilized Karpathy split (Karpathy and Fei-Fei, 2015). When applied to image-text and text-image retrieval, we calculated similarity scores for all images and texts, and then computed matching probabilities for the top-128 candidates to make inferences. For captioning, we generated captions using greedy decoding by inputting the image and [BOS] token.

The COCO Captioning dataset (Lin et al., 2014) consists of over 330,000 images, each annotated with 5 descriptive captions. We also used the standard Karpathy split and same decoding scheme with Flickr30k captioning.

Table A1 shows the statistics of the downstream task datasets.

B Baseline Model Architecture

Figure B1 illustrates the structural design of each baseline model.

C Ablation Study

We conducted an ablation study to investigate how each component contributed to the proposed method. In phase 1 of TransferCVLM, we reuse the parameters of teacher's fusion encoder to provide initial knowledge to CVLM before fine-tuning. To ablate this, we constructed a CVLM with randomly initialized fusion encoders of the same architecture (L = 6, H = 768, A = 12). For phase 2, we compared the distillation objective function (MSE) with activation boundary transfer (ABT) loss (Heo et al. 2019) which DoKTra utilized for their distillation phase (See Appendix D for detailed formulation of the objective functions). We also ablated the $\mathcal{L}_{KD}$ itself of Phase 2 to demonstrate that the performance improvement in Phase 2 is not solely due to the additional fine-tuning stage. We used same setting and configuration with TransferCVLM-base in Table 1 except for the ablated component.

Table C1 presents the results of the ablation study. In phase 1, the reuse of the parameters from the teacher's fusion encoder outperformed the random initialization of the fusion encoder. This

Phase	Ablation	VQAv2
Phase 1	Teacher fusion encoder	69.51±.66
	Random init.	68.57±.15
Phase 2	MSE	71.66±.12
	ABT	71.06±.05
	w/o KD loss ($\mathcal{L}_{task}$ only)	70.22±.22

Table C1: Ablation study of TransferCVLM on VQAv2 dataset.

KD Target	VQAv2
All encoders	71.26±.12
Fusion encoder only	70.26±.18
Unimodal encoders only	68.81±.14

Table C2: Ablation study on the distillation target of TransferCVLM on the VQAv2 dataset.

outcome suggests that CVLM could acquire initial knowledge for fusing representations from multiple modalities. Such a characteristic is well-suited for our method, given that TransferCVLM does not undergo the process of learning cross-modal dependencies through pre-training. For phase 2, training the model with only $\mathcal{L}_{task}$ once more resulted in a very slight performance improvement. However, the performance could be significantly improved only by applying the knowledge distillation loss through the transfer of the teacher model's knowledge. Moreover, distillation with MSE outperformed ABT, which is the distillation objective of DoKTra. The experiments of DoKTra were conducted on downstream tasks that are characterized by relatively limited data within specific domains. As a result, activation boundary distillation, known for its effectiveness in scenarios with scarce data, demonstrated strong performance. On the other hand, the two multimodal tasks used in our experiments are characterized by relatively abundant data, suggesting that the utilization of MSE was more appropriate.

We also performed an ablation study on the distillation target with VQAv2 for previous versions of CVLM (the only difference being the implementation of the task-specific layer). The results are presented in Table C2. As shown in the results, the best performance was achieved when distillation was applied to all encoders, so we applied the same method to the latest version of CVLM reported in section 3.

We conducted a similar ablation study on the encoder-decoder model and the captioning task (Flickr30k). Since we did not reuse teacher parameters in the encoder-decoder model, we excluded ablations related to Phase 1. Additionally, we conducted experiments by dividing the distillation targets into the entire model, encoder only, and decoder only. Tables C3 and C4 present the results, further demonstrating that our choices were optimal.

Phase	Ablation	Flickr30k
Phase 1	None	75.74
Phase 2	MSE	79.91
	ABT	34.17
	w/o KD loss ($\mathcal{L}_{task}$ only)	70.70

Table C3: Ablation study of TransferCVLM on Flickr30k captioning dataset.

KD Target	Flickr30k
All modules	79.91
Encoder only	74.88
Decoder only	78.55

Table C4: Ablation study on the distillation target of TransferCVLM on the Flickr30k captioning dataset.

D Formulation of the Distillation Objectives

We used mean square error as the distance metric $\mathcal{L}_{dist}$ in the knowledge distillation stage, which is formulated as:

$$\mathcal{L}_{dist}(S,T) = \mathcal{L}_{MSE}(S,T) = \sum_{i=1}^{d_F}(s_i - t_i)^2, \tag{8}$$

where $S = \{s_1, ..., s_{d_F}\}, T = \{t_1, ..., t_{d_F}\}$.

The activation boundary transfer loss we compared in our ablation study is formulated as follows:

$$\rho(x) = \begin{cases} 1, & \text{if } x > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

$$\begin{aligned} \mathcal{L}_{ABT}(S,T) = \|&\rho(T) \odot \sigma(\mu\mathbf{1} - S) \\ &+(\mathbf{1} - \rho(T)) \odot \sigma(\mu\mathbf{1} + S)\|_2^2, \end{aligned} \tag{10}$$

where $\odot$ is the element-wise product and $\mathbf{1}$ is a d_F-dimensional vector, with all values equal to 1. μ is the margin, which is a hyperparameter for training stability and set to 1.0. σ is an activation function such as ReLU.

	VQAv2	SNLI-VE	F30k-R	F30k-C	COCO
lr	1e-4	1e-5	1e-5	2e-6	2e-6
lr scheduler	cosine	cosine	linear	linear	linear
epochs	10	3	10	10	10
batch size	128	128	16	64	512
warmup ratio	0.1	0.1	0.1	0.0	0.0

Table E1: The hyperparameters for fine-tuning of teacher (FLAVA and GIT). F30K-R and -C refers to Flickr30k retrieval and captioning tasks, respectively.

	VQA			SNLI-VE			F30k-R		F30k-C		COCO	
	small	base	large	small	base	large	base	large	small	medium	small	medium
lr	1e-4	1e-4	1e-4	2e-5	1e-5	2e-5	2e-5	5e-5	1e-5	2e-5	1e-5	1e-5
lr scheduler	linear	cosine	linear	cosine	cosine	cosine	linear	linear	linear	linear	linear	linear
epochs	15	15	15	6	5	5	20	10	10	10	10	10
batch size	128	160	400	32	32	40	16	40	40	40	16	16
warmup ratio	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0

Table E2: The hyperparameters for fine-tuning of CVLM (Phase 1).

	VQA			SNLI-VE			F30k-R		F30k-C		COCO	
	small	base	large	small	base	large	base	large	small	medium	small	medium
lr	1e-4	1e-4	1e-4	1e-4	5e-5	5e-5	2e-5	2e-5	2e-5	1e-5	1e-5	1e-5
lr scheduler	cosine	cosine	cosine	cosine	cosine	cosine	linear	linear	linear	linear	linear	linear
epochs	15	15	15	9	6	8	10	10	10	10	10	10
batch size	64	64	80	64	32	40	16	32	16	64	16	64
warmup ratio	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0
α	0.5	0.5	0.5	0.9995	0.999	0.9995	0.7	0.5	0.8	0.9	0.9	0.9

Table E3: The hyperparameters for multimodal knowledge transfer (Phase 2).

E Hyperparameter Searching

In this section, we report the searching scheme and actual values of the hyperparameters we used in our experiments. All our experiments were triplicated with random seeds of [42, 123, 528].

In the knowledge distillation phase, our method uses the hyperparameter α to balance the task-specific loss with the knowledge distillation loss. The search range of the hyperparameter α was determined empirically based on the ratio of $\mathcal{L}_{KD}$ and $\mathcal{L}_{task}$. α was chosen among {0.3, 0.5, 0.7} for VQAv2 and Flickr30k retrieval, {0.99, 0.995, 0.999, 0.9995} for SNLI-VE, and {0.7, 0.8, 0.9} for captioning tasks.

Tables E1 to E3 show the hyperparameters used in each stage. We conduct a grid search on the development set to find the best hyperparameters. Since we experimented with small, base, and large models (small and medium in case of GPT-2) on different devices, we had to set different search units for the batch size.