VISUALLY-AUGMENTED PRETRAINED LANGUAGE MODELS FOR NLP TASKS WITHOUT IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Although pre-trained language models (PLMs) have shown impressive performance by text-only self-supervised training, they are found lack of visual semantics or commonsense, *e.g.,* sizes, shapes and colors of commonplace objects. Existing solutions often rely on explicit images for visual knowledge augmentation (requiring time-consuming retrieval or generation), and they also conduct the augmentation for the whole input text, without considering whether it is actually needed in specific inputs or tasks. To address these issues, we propose a novel **v**isually-**a**ugmented fine-tuning approach that can be generally applied to various PLMs or NLP tasks, **w**ithout using any retrieved or generated **i**mages, namely **VAWI**. Specifically, we first identify the visually-hungry words (VH-words) from input text via a token selector, where three different methods have been proposed, including syntax-, attention- and learning-based strategies. Then, we adopts a fixed CLIP text encoder to generate the visually-augmented representations of these VH-words. As it has been pre-trained by visual-language alignment task on large-scale corpus, it is capable of injecting visual semantics into the aligned text representations. Finally, the visually-augmented features will be fused and transformed into several pre-designed visual prompts based on VH-words, which can be inserted into PLMs to enrich the visual semantics in word repersentations. We conduct extensive experiments on ten NLP tasks, *i.e.,* GLUE benchmark, CommonsenseQA, CommonGen and SNLI-VE. Experimental results show that our approach can consistently improve the performance of BERT, RoBERTa, BART and T5 at different scales, and outperform several competitive baselines significantly. Besides, the generated visual prompts of our framework can also be used for parameter-efficient tuning, which boosts the performance of T5-3B. We will make our code, data, and models publicly available.

1 INTRODUCTION

Recent years have witnessed the success of pre-trained language models (PLMs), such as GPT-3 (Brown et al., 2020) and T5 (Raffel et al., 2020), in a variety of natural language process (NLP) tasks. Since these PLMs are mostly trained on text-only corpus via self-supervised pre-training, they have been shown lack of visual commonsense (Liu et al., 2022) and real-world knowledge (Zhang et al., 2022). As a result, PLMs can't well solve visually related language tasks [1], *e.g.,* answering the color and size of common things, especially those requiring complex commonsense knowledge.

To alleviate this problem, existing work mainly enhances PLMs by infusing visual information. Typically, given a text input, these studies firstly augment visual information from retrieved or generated images about the input and then leverage their visual representations to improve PLMs on NLP tasks. Such an approach leads to *visually-augmented pre-trained language models (VaLMs)*, where they adopt either visually-augmented pre-training (Tan & Bansal, 2020; Wang et al., 2022) or visually-augmented fine-tuning techniques (Lu et al., 2022). Despite the effectiveness, there are two major shortcomings in these methods. First, it is very costly to retrieve or generate high-quality images that are paired with the input. These methods often rely on pre-learned complementary retrievers or generators, and also require time-consuming inference for obtaining the images, which largely

[1] In this work, we mainly consider the text-only NLP tasks that may need visual information as complementary, but not the visual-language tasks with images.

limits the applicability of this approach. Second, as the augmented visual information comes from retrieved or generated images, it is inevitable to involve irrelevant or redundant visual information. If simply integrating the augmented visual information, the original text representations might be affected or even "polluted". Increasing evidence shows that visual information is not always useful for NLP tasks (Dai et al., 2022), and sometimes leads to a performance degradation.

Considering these issues, we aim to develop a more efficient and effective way to visually augment the PLMs and the solution is twofold:

• Firstly, we don't explicitly produce (retrieval or generate) the images but instead generate visually-aligned representations of the text on-the-fly. Recently studies (Radford et al., 2021; Jia et al., 2021) have shown that the visual-language pre-trained models (VL-PTMs) can well learn the alignment between the representations of texts and images from large-scale text-image pairs. Thus, our idea is to employ the output representations of a text from VL-PTMs' text encoders as a surrogate for the visual representations of related images. Such a way is simple and efficient: we can only keep the text encoder of a VL-PTM to produce the visually-aligned representations of texts, getting rid of the complicated image retrieval or generation process. It is widely recognized that there is large semantic gap between different modalities (Liang et al., 2022). Our method can alleviate this issue to some extent, since the visual augmentations are derived from the text representation itself, *i.e.,* visually-aligned text representations from VL-PTMs.

• Secondly, instead of directly feeding visual augmentations into the PLM, we propose to use the augmented visual information only when it actually needs. In fact, for a text input of a NLP task, PLMs are not always hungry for the visual background knowledge to effectively understand it, especially for visually-irrelevant expressions. Unlike previous work which injects visual information into a text (Wang et al., 2022) from the whole, we consider identifying *visually-hungry words* (those that require visual knowledge to derive complete semantics) from the text input, and only infuse the visual augmentations through these trigger words. We conduct visual augmentations at the word level, because it is more flexible and controllable, considering the augmented information is often irrelevant or noisy.

To this end, in this paper, we propose a general Visually-Augmented fine-tuning approach to improveing PLMs for NLP tasks Without Images, namely **VAWI**. Our approach consists of three ingredients, namely visually-hungry words extraction, visual knowledge augmentation, and visually-enhanced fine-tuning. Given the text input from a NLP task, we first extract the visually-hungry words (VH-words) from the input sentence. As the annotations of VH-words are generally unavailable, we propose three strategies to automatically extract the VH-words, relying on the syntax trees, attention distributions of the PLM, and an adaptive learnable module, respectively. Then, based on the extracted VH-words, we leverage the text encoder of CLIP (Radford et al., 2021) (being fixed in our approach), a VL-PTM that has been pre-trained on millions of text-image pairs, to encode the VH-tokens for obtaining their visually-aligned representations. Finally, we design visually-enhanced fine-tuning strategies to infuse the visually-aligned representations into PLMs. For small PLMs, we directly incorporate these visual representations to enrich the word embeddings and fine-tune the parameters of the PLM and our approach. For large-scale PLMs, we also propose a parameter-efficient prompt-tuning strategy that only tunes the very few parameters in our approach, with the PLM being frozen.

To summarize, our approach provides an adaptive, flexible and efficient way to leverage visual information for enhancing text-based PLMs. For verifying the effectiveness of our framework **VAWI**, we test it on four PLMs (*i.e.,* BERT, BART, RoBERTa, and T5) at different scales (*i.e.,* 110M, 340M, 3B), and conduct extensive experiments in natural language understanding, commonsense reasoning, and text generation. Experimental results show that our **VAWI** can boost the performance of these PLMs significantly, *e.g.,* 3.11%, 2.54%, and 2.16% absolute improvements on the commonsenseQA task using RoBERTa-base, RoBERTa-large, and T5-3b, respectively. Besides, **VAWI** can outperform (or be on par with) several competitive baselines that adopt complicated visually-augmented methods. Additionally, **VAWI** further improves the performance of a VL-PTM (*i.e.,* ALBEF(Li et al., 2021)) on the cross-modal reasoning task by enhancing its text-encoder.

2 RELATED WORK

Pre-trained Language Models. Recent years have witnessed the success of pre-trained language models (PLMs) (Devlin et al., 2019; Radford et al., 2019; Lewis et al., 2020; Raffel et al., 2020). After pre-trained on large-scale corpora, PLMs can be fine-tuned on multiple NLP tasks and achieve remarkable performance. Typically, PLMs incorporate the Transformer architecture (Vaswani et al., 2017) and require large-scale corpora. With the increasing scale of the model parameters and pre-training corpora, PLMs can even achieve the human performance (Devlin et al., 2019; Radford et al., 2019). However, since PLMs are just pre-trained with text-only data, they may suffer from the reporting bias problem (Gordon & Van Durme, 2013; Paik et al., 2021; Zhang et al., 2022), where the frequency distribution of visual commonsense in text may not reflect the real-world distribution of the commonsense. Existing works have also found that such a problem can not be addressed by enlarging the model or pre-training corpora (Paik et al., 2021; Zhang et al., 2022). In this work, we aim to reduce the influence of the reporting bias problem on PLMs by adding visual knowledge during fine-tuning.

Vision-Language Pre-Trained Models. To better accomplish the vision-language tasks, vision-language pre-trained models (VL-PTMs) (Su et al., 2019; Lu et al., 2019; Li et al., 2021) become a hot point in recent years, which require large-scale image-text pairs for pre-training. Existing VL-PTMs fall into two categories based on the way to modeling vision-language interaction. The first category of models (Su et al., 2019; Lu et al., 2019; Li et al., 2021) adopts an explicit vision-language interaction layer to fuse the text embeddings and image features using a single-stream or dual-stream network structure. These models are more suitable for the tasks that need to capture fine-grained semantic interactions between vision and language (*e.g.,* NLVR, VQA, and VE), and have achieved remarkable performance on them (Li et al., 2021; Alayrac et al., 2022). The second category of models (Radford et al., 2021; Jia et al., 2021) incorporates separate encoders to independently model the vision and language information, and rely on pre-training tasks (*e.g.,* cross-modal contrastive learning) to align their representations into the same space. Such a way is capable of producing enriched single-modal representations and has achieved remarkable performance on both vision and image-text retrieval tasks (Radford et al., 2021).

Visually-Augmented Language Model. Visually-augmented pre-trained language model (VaLM) (Wang et al., 2022) has become an emerging research topic that aims to introduce visual information into PLMs for alleviating the reporting bias problem. Existing VaLMs can be categorized into visually-augmented pre-training and fine-tuning methods. Visually-augmented pre-training approaches (Tan & Bansal, 2020; Wang et al., 2022) have to continually pre-train the PLMs with visual information. They mostly retrieve visual information related to input tokens or sentences from a large-scale images dataset and revise the masked language model task for better capturing the visual semantics. Visually-augmented fine-tuning methods (Lu et al., 2022) introduce the visual information into PLMs during the fine-tuning stage. These methods also leverage the image retrieval or generation models to augment the visual information and design a special fusion module to inject it into the PLMs. However, existing VaLM approaches mostly need to retrieve or generate visual information for utilization. Such a way is not only time-consuming, but also may involve unrelated or noisy information into PLMs, leading to performance degradation. In this work, we aim to first detect the visually-hungry words from the text, and then utilize a VL-PTM to generate their visual representations without the usage of external images or generation models. As a comparison, our approach is more flexible and efficient.

3 METHOD

In this section, we firstly introduce the task setting, and then describe our proposed visual augmentation approach for infusing visual knowledge into PLMs during fine-tuning.

3.1 TASK SETTING AND SOLUTION OVERVIEW

This work aims to improve the fine-tuning performance of pre-trained language models (PLMs) on NLP tasks by leveraging related visual information without images. For a NLP task, a set of n

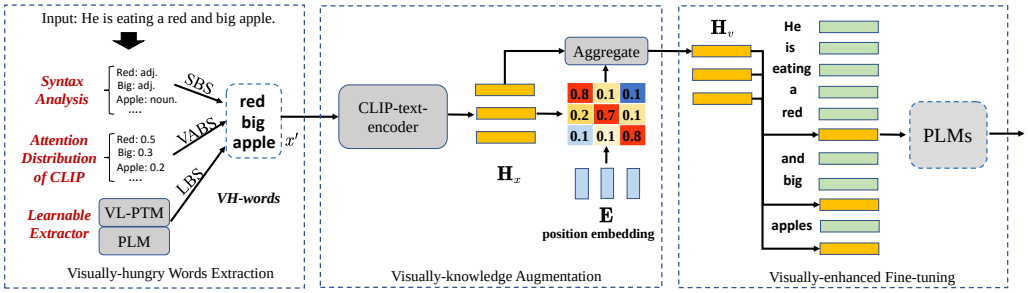

Figure 1: The illustration of our VAMI approach, consisting of visually-hungry words extraction, visual knowledge augmentation and visually-enhanced fine-tuning.

labeled texts $\{\langle x_i, y_i \rangle\}$ are available, where x_i is the i-th text data consisting of a sequence of words, denoted as $x_i = \{w_1, w_2, ..., w_m\}$, and y_i is the ground-truth output. Note that we don't specify the form of y_i, which can be a discrete label (classification), a continuous value (regression) or a text sequence (generation), since our approach is generally applicable to various settings (either natural language understanding or generation tasks).

To solve the target task, we assume that a text-based PLM is given (either for understanding or generation). Let f denote a PLM parameterized by θ_{PLM} that has already been pre-trained on general-purpose large-scale text data. We mainly focus on the encoder part of the PLM, which learns the input embedding via the "`[CLS]`" token. Given the labeled training data, we can train the PLM (*e.g.,* cross-entropy loss) and further solves the target task. However, existing works (Zhang et al., 2022) have revealed that PLMs may be unaware of visual knowledge that is not explicitly mentioned in the pre-trained text-only data (*e.g.,* the shape of coins and the color of the sky), leading to the lack of world commonsense and generating wrong statements.

In this work, we focus on devising an efficient and effective way to infuse such visual knowledge into PLMs during fine-tuning. Our approach is based on *visually-hungry words* (short as *VH-words*), which require visual information to derive complete semantic representations. The overall illustration of our approach is shown in Figure 1, which consists of three important ingredients, namely VH-words extraction, visual knowledge augmentation, and visually-enhanced fine-tuning. Given the input text x_i and its label y_i, we first detect and extract a set of VH-words. Then, we adopt a visual knowledge augmentation module to enhance the visual background knowledge of their tokens, and generate their visually-aligned representations. Finally, we infuse the visually-aligned text representations into the PLM to improve its fine-tuning performance, where we consider both the general fine-tuning of small PLMs and the parameter-efficient fine-tuning of large-scale PLMs.

3.2 Visually-Hungry Words Extraction

In our approach, visually-hungry words (VH-words) are the trigger units for visual augmentation, which require visual knowledge for deriving complete semantic representations. For example, some visually-related words (*e.g.,* color, shape and object) may not be well learned by the text-only PLM. Therefore, we propose to first detect the VH-words from the input text, and then inject proper visual knowledge that they are hungry for into the PLM. However, the annotations about VH-words are generally not available in NLP datasets. To address this problem, we devise three different strategies to extract the VH-words from the input text, including two feature-based strategies based on syntax tree and attention distribution of PLMs, and a learnable model-based strategy.

Syntax-based Strategy. In natural language, entity words and descriptive words usually convey more visual semantics than others. For example, for the sentence "*He is eating a green apple*", where underlined words are more related to visual semantics. Such words are mostly the nouns or adjectives in the input text, which can be detected by syntactic analysis. Therefore, we design a rule-based strategy that leverages the syntactic information of the words in s for VH-words extraction. Concretely, we firstly delete all stopwords in a text, and then adopt an open-resource toolkit SPACY to convert the input text into a syntax dependency tree. Based on the syntax tree, we extract the words

that have a particular part of speech (POS), *e.g.,* nouns or adjectives as the VH-words $\mathcal{W}^{(VH)}$. In this way, we can efficiently extract the VH-words from input text by using fast parser toolkit. However, as such a strategy mainly relies on the syntax information, it may select some nouns or adjectives that are explicitly related to visual semantics (*e.g.,* sweet and bitter). In the following, we will introduce two improved strategies to alleviate this problem.

Visually-enhanced Attention based Strategy. The attention-based strategy utilizes the attention distribution of a VL-PTM to detect the VH-words. Since VL-PTMs (Radford et al., 2021) are pre-trained on large-scale image-text pairs, their text encoders are able to focus more on the words corresponding to some specific visual concept in an image, which are likely to be VH-words. Inspired by this idea, we can use the attention scores calculated by the text-encoder of VL-PLMs to select the VH-words. Specifically, we adopt the text encoder of CLIP (Radford et al., 2021), a VL-PTM that has been pre-trained on millions of image-text pairs, to help extract the VH-words. As CLIP adopts an autoregressive GPT-2 model as the text encoder, we calculate the average attention scores between each token and the "`[EOS]`" token on the last multi-head self-attention layer, denoted as s_{w_i}. Then, we select the top-K ranked words according to $\{s_{w_i}\}$ as the VH-words denoted by $\mathcal{W}^{(VH)}$. In this way, we can leverage the intrinsic knowledge encoded by the VL-PTM to extract VH-words.

Learning-based Strategy. Considering that diverse PLMs and NLP tasks may be hungry for different complementary visual information, we devise a learning-based strategy that can adaptively extract VH-words according to task requirements. Concretely, we add a parameterized VH-words extractor layer for the PLM, which can be updated by optimization algorithms to fit the need for some specific task. Given the input text x_i, we first leverage the PLM and a text encoder of a VL-PTM (*i.e.,* CLIP (Radford et al., 2021)) to produce the contextualized representations of the contained words in x_i. Then, we concatenate the representations of each word from the two models, and utilize a MLP layer to project it into the score s_{w_i}:

$$s_{w_i} = \text{MLP}([\mathbf{h}_{w_i}^{(P)}; \mathbf{h}_{w_i}^{(V)}]) \tag{1}$$

where $\mathbf{h}_{w_i}^{(P)}$ and $\mathbf{h}_{w_i}^{(V)}$ are the output word representations from the PLM and VL-PTM, respectively, and scores s_{w_i} are calculated by the learned model based on supervision information from downstream tasks. Based on the scores of all words, we incorporate the gumbel-softmax function (Jang et al., 2016) to extract the top-k words as the VH-words in a differentiable way. In this way, the gradients of the fine-tuned tasks can be back-propagated to the extractor layer, which learns to adaptively select the more suitable VH-words.

3.3 Visual Knowledge Augmentation

Existing works (Lu et al., 2022) mainly utilize an image retrieval or generation module to augment related visual knowledge. Such a way is time-consuming and may also involve noisy images that will influence the task performance. Inspired by recent works that show the effective visual-language alignment in VL-PTMs (Radford et al., 2021; Li et al., 2021), we utilize the visually-aligned text encoders as the visual augmentation representations of VH-words. As the text encoders have been aligned to the image encoders during pre-training on large-scale image-text pairs, their output textual representations can be used as surrogate of visual augmentation based on real images related to the input text. As will be shown in experiments (Section 4), this approach is not only efficient but very effective for NLG tasks.

Based on the extracted VH-words, we first add a prefix text in image caption style before the VH-words, *e.g.,* "*a photo of:* ", to compose the input text x'. Then, we utilize the text encoder of CLIP (Radford et al., 2021) to encode x' and obtain the contextualized word representations as the visually-aligned representations $\mathbf{H}_x \in \mathbb{R}^{k \times d}$, where k is the number of VH-words and d is the embedding size. Next, we incorporate a reformulation layer to aggregate the visually-aligned representation $\mathbf{H}_x$ into visually-augmented representations of the VH-tokens. To effectively reserve the visual semantics, position information is particularly important to consider. For example, given a sentence "*a green apple*", it indicates that word "*green*" is a prefix for word "*apple*", which is useful to accurately convey visual semantics. Therefore, we design a position-aware attention mechanism to inject position information into visual representations. Specifically, we leverage a position

embedding matrix $\mathbf{E} \in \mathbb{R}^{l \times d}$ to reserve the position information of VH-words, and perform the cross-attention between it and the pseudo visual representations as:

$$\mathbf{Q} = \mathbf{E}, \quad \mathbf{K} = \mathbf{H}_s \mathbf{W}^K + \boldsymbol{b}^K, \quad \mathbf{V} = \mathbf{H}_s \mathbf{W}^V + \boldsymbol{b}^V, \tag{2}$$

$$\mathbf{H}_v = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}, \tag{3}$$

where $\mathbf{H}_v \in \mathbb{R}^{k \times d}$ is the obtained visually-augmented representations of VH-words, which can be leveraged for augmenting the visual knowledge of the PLM during fine-tuning. Note that in Eq. 2, we only use position information to set the *query* matrix $\mathbf{Q}$, which is more efficient than standard self-attention (Vaswani et al., 2017), and the visual semantics are mainly captured and injected through the *key* and *value* matrices.

3.4 Visually-Enhanced Fine-tuning

After obtaining the visually-augmented representations of VH-words (*i.e.,* $\mathbf{H}_v$ in Eq. 3), we propose a visually-enhanced fine-tuning strategy to inject the captured visual knowledge. Here, we consider two cases: (1) full-parameter fine-tuning for small PLMs, and (2) parameter-efficient prompt-tuning for large-scale PLMs. Before introducing the learning method, we simply review the parameters to consider, consisting of the parameters in the underlying PLM (denoted by Θ_{plm}) and the parameters from our approach (denoted by Θ_{ref}). Note that we will always fix the text encoder from CLIP.

Fine-tuning for Small PLMs. For small PLMs, we can perform full-parameter fine-tuning, which update both Θ_{plm} and Θ_{ref}. Specifically, given the visually-augmented representations $\mathbf{H}_v$ of VH-words, we directly incorporate them into the embedding layer of the PLM. For a certain VH-word, we directly integrate the original word embedding and the visually-augmented representation, which leverages the visual semantics to enrich the word representations.

Prompt-tuning for Large-Scale PLMs. For large-scale PLMs, we fix the parameters in it, *i.e.,* Θ_{plm} and employ a parameter-efficient prompt-tuning way to optimize it on downstream NLP tasks. Concretely, given the visually-augmented representations $\mathbf{H}_v$ of VH-words, we directly prefix them before the input embeddings as *soft prompts*. Then, following the typical prompt-tuning paradigm (Li & Liang, 2021), we only tune the parameters of the reformulation layer (*i.e.,* Θ_{ref}) to control the soft prompts.

Our approach can be generally applied to various PLMs (*e.g.,* BERT (Devlin et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2020)) and NLP tasks (natural language understanding and text generation). Unlike other complicated visually-augmented methods (Tan & Bansal, 2020; Wang et al., 2022), it is more efficient, without the explicit need of images; and meanwhile, it only introduces introduces a relatively small number of parameters (Eq. 2), which are easier to learn.

4 Experiments

4.1 Experimental Setup

Dateset. We conduct experiments on four types of NLP tasks: (1) Nature Language Understanding (NLU) task: we extract 6 datasets from the popular GLUE benchmark (Wang et al., 2018); (2) Commonsense reasoning task: we select CommonsenseQA (Talmor et al., 2019), a 5-way multiple choice QA dataset that requires commonsense knowledge; (3) Text generation task: we select CommonGen (Lin et al., 2019b), a constrained text generation task about generative commonsense reasoning. (4) Cross-modal reasoning task: we select SNLI-VE (Xie et al., 2019), to evaluate the capacity of predicting whether the image semantically entails the text.

Baseline Models. We compare our approach with the following baselines, including pre-trained language models (PLMs), visual-language pre-trained models (VL-PTMs), and visually-augmented pre-trained language modes (VaLMs): (1) **PLMs**: We choose BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020), T5 (Raffel et al., 2020) as the PLM backbones, and directly fine-tune them as baselines. (2) **VL-PTMs**: We select ALBEF (Li et al., 2021),

BLIP (Li et al., 2022), and CLIP (Radford et al., 2021) as baselines, which have been pre-trained on large-scale image-text pairs via cross-modal contractive learning. (3) **VaLMs**: we select VOKEN (Tan & Bansal, 2020) and iACE (Lu et al., 2022), which introduce visual information into PLMs by pre-training on retrieved images and fine-tuning on generated images, respectively.

Implementation Details. We implement all methods based on Huggingface Transformers (Wolf et al., 2020). For all the baselines, we set the hyper-parameters according to the original papers. In our approach, we leverage the text encoder of CLIP (ViT-B/32) to implement the learnable model-based VH-words extractor and generate the visual representations of VH-words in the visual knowledge augmentation module. The hidden size of visual representations is set to 512. For different NLP tasks, we tune the number of visually hungry words in $\{2, 3, 4, 5\}$. During fine-tuning, we perform parameter-efficient tuning on T5-3B and BART-Large, and full-parameter tuning on other PLMs. For all tasks and all backbones, we utilize Adam as the optimizer, set the learning rate to 2e-5, weight decay to 0.01, and a linear warmup for the first 6% steps. For NLU, GommonGen, and SNLI-VE datasets, we fine-tune our model for 3 epochs with 32 batch size. For CommonsenseQA, we tune our model for 10 epochs with 32 batch size. We use cross-entropy loss for classification and mean squared error loss for regression. All experiments are conducted on a V100 GPU.

4.2 Main Experimental Results

In this part, we conduct a series of experiments on NLU, commonsense reasoning, text generation and cross-modal commonsense reasoning tasks.

Evaluation on NLU tasks. We present the experimental results of different methods on 6 NLU tasks in Table 1. First, we observe that VL-PTMs perform worse than PLMs, a possible reason is that they have been continually pre-trained on large-scale image-text pairs, which may cause catastrophic forgetting problem. Second, VaLMs (*i.e.,* VOKEN, iACE and VAWI) achieve better performance over PLMs. As VaLMs infuse external visual knowledge into the PLMs, they can help the PLMs better understand the background knowledge of some words (*e.g.,* color, shape, and size of objects). Between the two VaLM baselines, iACE is slightly better. This is because iACE is enhanced based on VOKEN and incorporates an image generation model, so it produces more visual information to utilize. However, the generated images are inevitable to contain noise and redundant information, which limits the performance gain of iACE.

Finally, by comparing our approach with all baselines, it is obvious that VAWI performs consistently better than them on the six datasets. In our approach, we adopt an efficient and effective way that augments the visual representations using the text encoder of CLIP to encode the VH-words from the input text. Benefiting from pre-training on large-scale image-text pairs, the text encoder of CLIP has been well aligned with the semantic space of images, so that it can generate high-quality visual representations of the VH-words to enrich them. Such a way not only saves the costs of time and computation but also reduces the influence of inevitable noise from retrieved or generated images. Additionally, among three VH-words extraction strategies, LBS slightly outperforms others in most NLU tasks. The reason is that LBS incorporates a learnable model-based strategy to select the VH-words. Such a way can adaptively extract proper VH-words with the consideration of the intrinsic knowledge of the PLM. However, LBS will increase the computation cost due to its involved VH-words extractor layer. Therefore, in the following experiments, we mostly utilize the SBS strategy in our approach for comparison.

Evaluation on Commonsense Reasoning Tasks. Following existing works (Lin et al., 2019a), we also rely on a rule-based strategy to extract the examples containing visible objects, to construct a new dataset called CommonsenseQA-3K. It consists of 2,903 and 342 examples in the training set and dev set, respectively. In this way, we can ensure that most of the instances in this dataset require visual information to enrich the commonsense knowledge. Based on the CommonsenseQA and CommonsenseQA-3k, we also report the results with different amounts of training data, to further evaluate the performance of different methods in the few-shot setting.

We show the experimental results of different methods on the commonsense reasoning tasks in Table 2. We can also see that with the help of visual information from either retrieved images or our VAWI-SBS, the performance of PLMs can be improved significantly. It indicates that visual in-

Table 1: Performance comparison of different methods on NLU tasks, the **BEST** results are highlighted in bold. +*None* denotes that we directly fine-tune the backbone without adding visual information. SBS, VABS, and LBS represent using the syntax-based strategy, visually-enhanced attention based strategy and learning based strategy in our approach, respectively.

Base Model	Method	SST-2	QNLI	QQP	MNLI	MRPC	STS-B	Avg.
CLIP	+None	73.3	74.5	72.8	68.4	74.3	73.8	72.85
BLIP	+None	76.3	77.4	78.8	72.5	77.8	76.4	76.53
ALBEF$_{14M}$	+None	78.9	78.2	79.4	73.4	76.5	77.5	77.31
BERT$_{base}$	+None	89.3	87.9	87.2	79.4	81.7	84.4	84.98
	+VOKEN	92.2	88.6	88.6	82.6	83.5	86.0	86.83
	+iACE	91.7	88.6	89.1	82.8	85.8	86.6	87.43
	+VAWI-SBS	**92.9**	88.4	89.6	82.2	85.5	86.9	87.58
	+VAWI-VABS	92.7	88.9	89.5	82.7	**85.8**	**87.2**	**87.80**
	+VAWI-LBS	92.4	**89.1**	**89.7**	**83.0**	85.6	86.9	87.78
RoBERTa$_{base}$	+None	89.2	87.5	86.2	79.0	81.4	85.4	84.78
	+VOKEN	90.5	89.2	87.8	81.0	87.0	86.9	87.06
	+iACE	91.6	89.1	**87.9**	82.6	87.7	86.9	87.63
	+VAWI-SBS	91.4	89.4	87.7	82.2	88.2	87.7	87.76
	+VAWI-VABS	**91.7**	89.1	**87.9**	**82.6**	88.3	88.1	87.95
	+VAWI-LBS	91.6	**90.6**	**87.9**	82.4	**88.5**	**88.3**	**88.21**

formation is indeed helpful to improve PLMs for understanding commonsense knowledge. Besides, our approach outperforms the method using retrieved images from search engines. The reason is that our approach omits the image retrieval process due to its inevitably involved noise, and relies on the text encoder of CLIP to augment the visual representations. Such a way can guarantee the relevance between the augmented visual knowledge and the text input, reducing the influence of retrieved noisy images and redundant information. Furthermore, we perform parameter-efficient tuning on T5-3B-encoder with our approach and also boost its performance. It shows that our approach is able to be applied to large-scale PLMs to meet their thirst for visual information.

Table 2: Performance comparison on CommonsenseQA-3k and CommonsenseQA with different amounts of training data. We report the average performance on the dev set over three runs, and the **BEST** results are highlighted in bold. +*retrieved images* denotes that we add retrieved images about the VH-words using web search engines, and encode them via CLIP-ViT.

Base Model	Method	CommonsenseQA-3k				CommonsenseQA			
		5%	10%	20%	100%	5%	10%	20%	100%
RoBERTa$_{base}$	+None	41.88	46.04	50.58	61.88	44.88	50.04	57.08	67.90
	+retrieved images	42.37	48.09	52.81	64.22	45.72	51.17	58.96	69.64
	+VAWI-SBS	**42.94**	**49.27**	**53.97**	**65.10**	**46.51**	**52.44**	**59.87**	**71.01**
RoBERTa$_{large}$	+None	48.39	56.30	59.06	74.19	51.24	59.95	65.52	76.65
	+retrieved images	49.55	57.78	61.29	75.61	52.18	60.93	66.08	78.39
	+VAWI-SBS	**50.27**	**58.17**	**62.22**	**76.54**	**52.98**	**61.97**	**67.40**	**79.19**
T5-3B	Baseline	70.16	73.02	75.04	81.81	71.99	75.27	77.72	82.40
	+retrieved images	70.96	73.60	75.91	82.40	72.87	76.17	78.71	83.64
	VAWI-SBS+PET	**71.52**	**74.19**	**76.49**	**83.61**	**73.58**	**73.58**	**79.66**	**84.56**

Evaluation on the Text Generation Task. As shown in previous experiments, it is useful to improve the performance of VAWI on the commonsense reasoning and nature language understanding. Here, we would like to study the effectiveness of our approach on the text generation task (*i.e.,* CommonGen) using large PLMs. As shown in Table 3, our model VAWI also consistently boosts the performance of BART-large and T5-3B among all metrics. It further shows that our approach can also improve PLMs on the text generation task. As a comparison, we can see that the retrieved images are not very helpful and even cause performance degradation. The reason may be that the text generation task is more sensitive to the inevitable noise from the retrieved images. Finally, the

parameter-efficient tuning strategy of our approach also achieves comparable performance with the full-parameter tuning. It indicates that our parameter-efficient strategy is able to efficiently optimize the parameters of large-scale PLMs, and shows a promising future to apply our approach to much larger PLMs, *e.g.,* GPT-3.

Table 3: Performance comparison on CommonGen. We also show the performance of parameter-efficient tuning of our approach, denoted as *+PET*. The best results are highlighted.

Method	Base Model	BLUE-3	BLUE-4	METOR	Rouge-L	CIDER	SPICE
BART-large	+None	42.80	32.42	31.36	57.57	16.56	32.94
	+retrieved images	42.67	32.67	32.12	57.46	16.78	32.81
	+VAWI-SBS	**44.56**	**34.17**	**32.47**	**58.46**	**17.23**	**33.67**
	+VAWI-SBS+PET	43.12	33.76	32.20	58.12	16.91	33.17
T5-3b	+None	45.92	35.92	33.02	58.57	17.71	33.51
	+retrieved images	45.69	35.50	33.55	58.94	17.51	32.91
	+VAWI-SBS	**47.67**	**37.54**	33.41	**59.94**	**18.34**	**34.67**
	+VAWI-SBS+PET	47.40	37.36	**33.71**	59.78	18.18	34.17

Evaluation on the cross-modal commonsense reasoning. To verify the generality of our method, we further evaluate the effectiveness of our VAWI on improving a VL-PTM (*i.e.,* ALBEF), and conduct experiments on a cross-modal reasoning dataset, SNLI-VE. Concretely, we implement our approach on ALBEF by inserting the visually-augmented representation into the embedding layer of the text encoder before the fusion layer, and keep others unchanged. As shown in Table 4, our VAWI can also improve the performance of ALBEF using different amounts of training data. It further shows the generality of our approach in VL-PTMs, as it can also provide rich information to enhance the encoder of the text input, helping it better perform cross-modal reasoning.

Table 4: Experimental results on the test set of SNLI-VE task.

Method	SNLI-VE			
	10%	20%	50%	100%
ALBEF	65.46	67.52	75.47	80.91
ALBEF+VAWI +SBS	**65.94**	**68.23**	**76.14**	**81.64**

5 Conclusion

In this paper, we proposed a general visually-augmented fine-tuning approach that can be applied to a variety of PLMs or NLP tasks, without using any retrieved or generated images, namely **VAWI**. Specifically, we first identified and extracted the visually-hungry words (VH-words) from input text via a token selector, where three different methods have been proposed, including syntax-, attention- and learning-based strategies. Then, we adopted a fixed CLIP text encoder to generate the visually-augmented representations of these VH-words. As it has been pre-trained by visual-language alignment task on large-scale corpora, it is capable of injecting visual semantics into the aligned text representations. Finally, we transformed the visually-augmented features into several pre-designed visual prompts based on VH-words, and inserted them into PLMs to enrich the visual semantics of word representations in PLMs. Extensive experiments conducted on ten NLP tasks, *i.e.,* GLUE benchmark, CommonsenseQA, CommonGen, and SNLI-VE show that our approach can consistently improve the performance of BERT, RoBERTa, BART, and T5 at different scales, and outperform several competitive baselines significantly. Besides, the visual prompts of our framework can also be used for parameter-efficient tuning, which boosts the performance of T5-3B.

In the future, we will investigate to devise more effective approaches to extract visual-hungry words from the input text, and incorporate more effective VL-PTMs to augment the visual representations. Besides, we will also devise more effective visually-augmented pre-training strategies without images, to further improve the performance of PLMs.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

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

Wenliang Dai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. Enabling multimodal generation on clip via vision-language knowledge distillation. *arXiv preprint arXiv:2203.06386*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pp. 25–30, 2013.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pp. 4904–4916. PMLR, 2021.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, 2020.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053*, 2022.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. Kagnet: Knowledge-aware graph networks for commonsense reasoning. *arXiv preprint arXiv:1909.02151*, 2019a.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. Commongen: A constrained text generation challenge for generative commonsense reasoning. *arXiv preprint arXiv:1911.03705*, 2019b.

Xiao Liu, Da Yin, Yansong Feng, and Dongyan Zhao. Things not written in text: Exploring spatial commonsense from visual signals. *arXiv preprint arXiv:2203.08075*, 2022.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Yujie Lu, Wanrong Zhu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. Imagination-augmented natural language understanding. *arXiv preprint arXiv:2204.08535*, 2022.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. The world of an octopus: How reporting bias influences a language model's perception of color. *arXiv preprint arXiv:2110.08182*, 2021.

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

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

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683, 2020.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. 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)*, pp. 4149–4158, 2019.

Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*, 2020.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Weizhi Wang, Li Dong, Hao Cheng, Haoyu Song, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. Visually-augmented language modeling. *arXiv preprint arXiv:2205.10178*, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020.

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

Chenyu Zhang, Benjamin Van Durme, Zhuowan Li, and Elias Stengel-Eskin. Visual commonsense in pretrained unimodal and multimodal models. *arXiv preprint arXiv:2205.01850*, 2022.

APPENDIX

A ABLATION STUDY

The effect of the source of visual representations. In this part, we verify whether the improvement is due to the involvement of visual information about the VH-words. We propose three variants that incorporate powerful PLMs, *i.e.,* RoBERTa-base, T5-large, and T5-3B respectively, to replace the text encoder of CLIP in our framework. We also replace the generated visual representations from the text encoder of CLIP with random noise, to investigate the effectiveness of the visual representations. As shown in Table 5, we can see that our approach is better than all the variants, even T5-3B with billion-scale parameters. It indicates that the improvement of our approach is indeed due to the learned visual knowledge within CLIP, and it is not effective to simply increase the model scale to improve the understanding of visual knowledge. Besides, our approach also outperforms the variant using random noise as the visual representation, showing the worse performance among all the variants. It also shows the importance of the visual representations, as they indeed contain the visual knowledge that the PLM is hungry for.

Table 5: Performance comparison of different variants of our approach.

Source of visual representation (Params)	CSQA-3k	SST-2	QQP
Random Noise (0M)	61.59	89.13	87.14
RoBERTa$_{large}$ (355M)	61.18	89.43	87.53
T5$_{large}$-encoder (375M)	62.21	89.71	87.94
T5$_{3B}$-encoder (1500M)	63.10	90.24	88.21
CLIP-text-encoder (52M)	**65.10**	**92.93**	**89.74**

The Number of VH-words. Our approach has an important hyper-parameter required to tune, the number of VH-words. Here, we would like to study whether more VH-words are better to improve performance. We conduct experiments on the QQP and CSQA-3K datasets using RoBERTa-base as the backbone, and present the results in Figure 2. We can see that with the increase of the number of VH-words, the performance gain of our approach first increases and then decreases. A possible reason is that too many VH-words may also introduce noisy or redundant information (*e.g.,* not very relevant words), which would also influence the fine-tuning performance. Instead, it is also more efficient to select a few VH-words (*e.g.,* two words for CSQA-3k) for deploying our approach in large-scale PLMs.

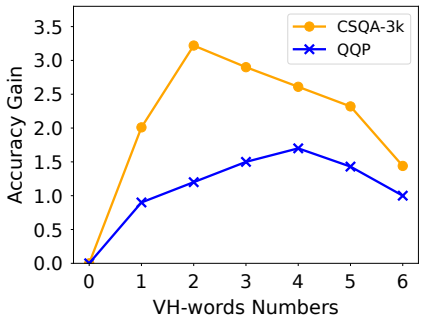

Figure 2: Performance comparison w.r.t. different numbers of VH-words.

Table 6: Performance comparison w.r.t. different insertion positions of visual representations. The base model is RoBERTa-base.

Insert Positions	CSQA-3k			
	5%	10%	20%	100%
Not Insert	41.88	46.04	50.58	61.88
Before input text	-	39.77	44.86	57.47
After input text	-	40.23	45.67	58.08
Nearby VH-tokens	**42.94**	**49.27**	**53.97**	**65.10**

Different Insertion Positions of Visual Representation.
In our visually-enhanced fine-tuning framework, we insert the visual representation into the nearby positions of the VH-words. To verify its effectiveness, we propose three variants of it that do not insert, inserts before and after the input text, respectively.

As shown in Table 6, we can observe that all these variants would lead to a performance decrease. It demonstrates that a proper position to insert the visual representation is important for the utilization of augmented visual representations. By inserting them nearby the embeddings of VH-words, PLMs can effectively aggregate the visual representations to enrich the word representations, leading to better performance on NLP tasks.