Efficient Image Captioning for Edge Devices

Ning Wang, Jiangrong Xie, Hang Luo, Qinglin Cheng, Jihao Wu, Mingbo Jia, Linlin Li

Huawei Inc.

wn6149@mail.ustc.edu.cn, xiexjr@foxmail.com, {lhjeremy, qlincheng}@outlook.com,
{wujihao, jiamingbo, lynn.lilinlin}@huawei.com

Abstract

Recent years have witnessed the rapid progress of image captioning. However, the demands for large memory storage and heavy computational burden prevent these captioning models from being deployed on mobile devices. The main obstacles lie in the heavyweight visual feature extractors (*i.e.,* object detectors) and complicated cross-modal fusion networks. To this end, we propose LightCap, a lightweight image captioner for resource-limited devices. The core design is built on the recent CLIP model for efficient image captioning. To be specific, on the one hand, we leverage the CLIP model to extract the compact grid features without relying on the time-consuming object detectors. On the other hand, we transfer the image-text retrieval design of CLIP to image captioning scenarios by devising a novel visual concept extractor and a cross-modal modulator. We further optimize the cross-modal fusion model and parallel prediction heads via sequential and ensemble distillations. With the carefully designed architecture, our model merely contains 40M parameters, saving the model size by more than 75% and the FLOPs by more than 98% in comparison with the current state-of-the-art methods. In spite of the low capacity, our model still exhibits state-of-the-art performance on prevalent datasets, *e.g.,* 136.6 CIDEr on COCO Karpathy test split. Testing on the smartphone with only a single CPU, the proposed LightCap exhibits a fast inference speed of 188ms per image, which is ready for practical applications.

1 Introduction

Image captioning aims to automatically generate natural and readable sentences to describe the image contents, which provides a promising manner to help visually impaired people. The recent decade has witnessed a surge of captioning algorithms, benefiting from the development of large-scale pre-training (Zhou et al. 2020; Li et al. 2020b; Hu et al. 2021a; Wang et al. 2021), advanced representation learning (Zhang et al. 2021a; Huang et al. 2021), and modern cross-modal modeling (Xu et al. 2021; Li et al. 2020b; Fang et al. 2021a). In spite of the remarkable advances, current heavyweight captioning algorithms are not available to visually impaired people, who generally rely on low-resource devices such as portable phones to assist the daily life, instead of carrying on heavy computer servers with modern GPUs.

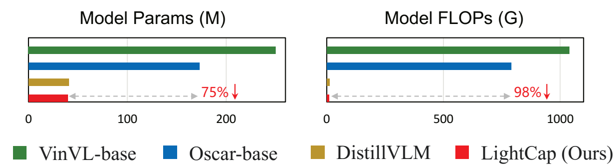

Figure 1: Compared to the state-of-the-art VinVL (Zhang et al. 2021a) and Oscar (Li et al. 2020b), our method saves more than 75% parameters and 98% FLOPs. Compared with the lightweight DistillVLM (Fang et al. 2021b), our method not only yields fewer parameters and FLOPs, but also outperforms it by a notable margin.

Designing computationally efficient and memory-friendly captioning methods is vital for practical applications but has been largely overlooked in the literature.

To achieve excellent performance, recent image captioners typically adopt deep object detectors as well as large cross-modal fusion networks. For example, the recent VinVL and LEMON algorithms (Zhang et al. 2021a; Hu et al. 2021a) utilize a strong but heavyweight ResNeXt-152 based detection model and a base or large BERT model (Devlin et al. 2018). Some methods even scale the model size from base to huge to attain superior captioning performance (Hu et al. 2021a), but how to effectively reduce the model size for edge devices is rarely touched in these works. These sophisticated image captioning models struggle to meet the real-time requirement of real-world applications, let alone the huge power consumption and memory storage. It is therefore non-trivial to investigate how to design an efficient image captioner with smaller memory storage, faster inference speed, and satisfactory performance.

In this paper, we propose LightCap, a lightweight yet high-performance image captioning method for mobile devices. Our core design is largely inspired by the recent CLIP method (Radford et al. 2021). CLIP is an impressive image-text retrieval model, which readily tells what objects exist in the image but fails to generate a description for the given image. In this work, we investigate how to transfer such a strong cross-modal retrieval model to an image captioner, and meanwhile break the obstacles that hinder image captioners from being deployed on the mobile devices. The main obstacles that hinder image captioners from being deployed on mobile devices are their cross-modal fusion and image feature extraction models. For visual representations, we leverage the efficient yet compact grid features from the CLIP without relying on time-consuming Region of Interest (ROI) features from sophisticated object detectors. To unveil the potential of a capacity-limited model, we propose the following designs. (1) *Visual concept extractor*. To take advantage of the cross-modal retrieval capability of CLIP, we train a region-based alignment model to retrieve the visual concepts from an off-the-shelf dictionary. These visual concepts serve as the description hints of the image to facilitate caption generation. (2) *Cross-modal modulator*. Before being fed to the fusion model, the feature dimension of the CLIP feature is highly compressed (*i.e.,* from 2048 to 312), which inevitably loses semantic representations. To retain the valuable semantics, we propose a cross-modal modulator that takes the textual concepts as inputs to activate the informative feature channels of the CLIP model. (3) *Ensemble head*. We jointly optimize and distill an ensemble of head networks for collaborative prediction. We disentangle the key parameters and share the rest weights of different heads for lightweight design. Last but not least, for the cross-modal fusion model, instead of the widely-used $\text{BERT}_{\text{base}}$ (Devlin et al. 2018), we chose the efficient TinyBERT (Jiao et al. 2019) to fuse cross-modal features. By virtue of our designed sequential knowledge distillations in both pre-training and fine-tuning stages and the ensemble distillations from multiple teachers, a TinyBERT almost matches the performance of the standard BERT.

By highly limiting the capacity of each component in our image captioner, the overall model merely contains 40M parameters and 9.8G FLOPs, saving the model size by more than 75% and the FLOPs by more than 98% compared to the current popular image captioning models (Figure 1). Despite its low capacity, the proposed method still exhibits state-of-the-art performance on prevalent captioning datasets, *e.g.,* 136.6 CIDEr on COCO Karpathy split (Lin et al. 2014). The model storage memory of LightCap is about 112MB, which is affordable on most mobile devices. It merely costs about 188ms to process an image when testing the proposed LightCap on the mobile phone with only one CPU, which is readily ready for practical usage.

In summary, in this paper, we systematically show how to obtain a lightweight, efficient, and high-performance captioner by careful designs and training:

- **Model Design.** We propose a *visual concept extractor* and a *cross-modal modulator* to better exploit the cross-modal capability of the CLIP model for image captioning. We further design a partially parameter-sharing *ensemble head* for collaborative prediction.
- **Model Training.** We present the *sequential knowledge distillations* from pre-training to fine-tuning to distill the tiny model. We leverage the *ensemble distillation* to better optimize the TinyBERT model and ensemble heads.

2 Related Work

Image Captioning. Image captioning methods generally contain a visual encoder to extract the image representations and a cross-modal fusion model to generate the caption. Previous methods (Huang et al. 2019; Pan et al. 2020; Anderson et al. 2018; Ji et al. 2021; Song et al. 2021; Fei 2022; Yang, Liu, and Wang 2022) typically utilize the object detection methods such as Faster-RCNN (Ren et al. 2016) to extract ROI features. The recent VinVL method (Zhang et al. 2021a) shows that a strong visual feature extractor consistently improves the performance on image captioning.

To reduce the computational burden, MiniVLM (Wang et al. 2020a) designs a lightweight object detector using EfficientNet backbone (Tan and Le 2019). DistillVLM (Fang et al. 2021b) leverages knowledge distillation to acquire a thinner transformer architecture for vision-language tasks. In contrast to the ROI features from object detectors, some cross-modal algorithms turn to the grid features for high efficiency, which are known as the detector-free approaches in the literature (Fang et al. 2021a; Xu et al. 2021; Wang et al. 2021; Wang, Xu, and Sun 2022). Nevertheless, these models (Fang et al. 2021a; Wang et al. 2021; Wang, Xu, and Sun 2022) still struggle to be deployed on edge devices. Compared with them, our method leverages a light yet powerful CLIP model to extract the grid features. We further propose a concept extractor and a cross-modal modulator to unveil the cross-modal representation power of the CLIP. Our approach outperforms previous efficient captioners such as MiniVLM (Wang et al. 2020a) and DistillVLM (Fang et al. 2021b) with lower model capacity and faster inference speed, and is even comparable to the recent heavyweight captioners.

Recent works (Shen et al. 2021; Cornia et al. 2021) also take advantage of CLIP model for image captioning. Nevertheless, they simply utilize the standard CLIP model to extract features or image tags. In contrast, to reduce the model size, we train a lightweight region-level concept extractor as well as a feature modulator to better exploit the cross-modal characteristic of CLIP.

VL Pre-training. Vision-language (VL) pre-training aims to learn robust cross-modal representations to bridge the domain gap between vision and language signals (Dou et al. 2021). CLIP (Radford et al. 2021) and ALIGN (Jia et al. 2021) align the VL representations via a light fusion manner (*i.e,* dot-product) using the contrastive learning technique. Nevertheless, their light fusion manner fails to conduct the cross-modal generation task such as image captioning. In contrast, recent VL pre-training approaches (Zhou et al. 2020; Chen et al. 2020; Li et al. 2020b,a; Zhang et al. 2021a) adopt a relatively heavy transformer architecture (Vaswani et al. 2017) to fuse the VL representations, which are qualified to perform more VL downstream tasks. Inspired by previous arts, our approach also involves VL pre-training to facilitate the downstream captioning task. Differently, we do not employ the widely-adopted bidirectional masked language modeling, and shed light on the unidirectional language modeling to fully focus on the text generation task, *e.g.*, image captioning. Furthermore, similar to previous arts (Jiao et al. 2019; Mukherjee and Awadallah 2020), we adopt the sequential knowledge distillation (KD) to preserve the model representational capability within a tiny network. Based on the general KD, we also investigate how to better leverage KD in the captioning task by intro-

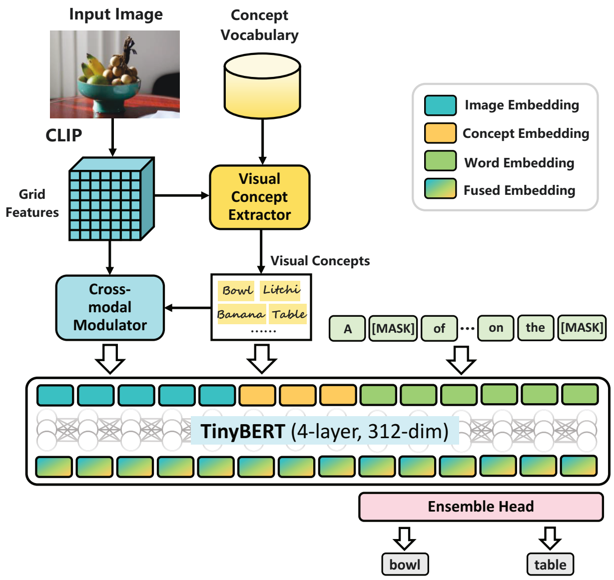

Figure 2: The overall framework of our LightCap. The input image is encoded to the grid visual feature via a ResNet-50 model. Then, we leverage a concept extractor to extract the visual concepts and a cross-modal modulator to reinforce the visual features. Finally, a TinyBERT fuses multi-modal embeddings and an ensemble head performs image captioning.

ducing concept distillation to facilitate the modality alignment and ensemble distillation for multi-head optimization.

3 Methodology

In this section, we introduce the technical details of the proposed method. First, in Section 3.1, we elaborate on the model design of each block. Then, in Section 3.2, we show the training details. Finally, we exhibit the model distillation in both pre-training and fine-tuning stages in Section 3.3.

3.1 Model Architecture

The overall framework is shown in Figure 2. Our LightCap contains an image encoder to extract the visual representations, a concept extractor to retrieve the visual concepts from an off-the-shelf vocabulary, and a cross-modal modulator to enhance the visual representations with the textual (concept) information. Finally, we use a lightweight TinyBERT to fuse multi-modal representations and an ensemble head module to generate the image caption.

Image Encoder. Instead of extracting expensive ROI features from object detectors, we leverage the ResNet backbone (He et al. 2016) to acquire grid representations. Specifically, we choose the recent CLIP model (ResNet-50 version) (Radford et al. 2021) due to (1) its impressive generalization capability, especially in the cross-modal domain; (2) its promising potential in extracting visual concepts from images, which is beneficial to the image captioning task. CLIP model contains a visual encoder and a text encoder. In the visual encoder, after obtaining the image feature map, CLIP additionally learns a transformer block (*i.e.*, attention pooler) to obtain the global image embedding. In our framework, to save the model capacity, we only utilize the ResNet-50 backbone in CLIP visual encoder *without* the attention pooler to extract the visual features $\boldsymbol{v} \in \mathbb{R}^{7\times7\times2048}$, which only involves 4.1G FLOPs.

Visual Concept Extractor. Intuitively, knowing the semantic concepts of the image is highly beneficial to image captioning. Although CLIP model is ready for cross-modal retrieval, there still exist two issues. First, CLIP relies on a heavy attention pooler to obtain the global image representation, which contains 14.8M parameters and is in conflict with our lightweight model design. Second, CLIP model is pre-trained using global image features and thus is not effective enough in recognizing image regions. To this end, we design and train an efficient region-based visual concept extractor on top of the CLIP feature.

The overall architecture of the proposed visual concept extractor is shown in Figure 3 (left). First, we collect the common object categories from the Visual Genome dataset (Krishna et al. 2017), and form these category words using the description form `a photo of [object]`. We take advantage of the CLIP text encoder to extract the textual embeddings of these descriptions to form an off-the-shelf vocabulary. Note that this vocabulary contains textual embeddings instead of the raw words to avoid unnecessary computations in the captioning stage. Then, we train an efficient foreground-background object detector without knowing object classes. This detector is designed to roughly predict the foreground bounding boxes, whose architecture is tiny YOLOv5n with only 1.9M parameters. After obtaining the object proposals, we employ ROI-Align (He et al. 2017) to pool the region embeddings. These ROI embeddings are further processed by two linear blocks to align with the concept embeddings in the aforementioned vocabulary. To train this concept extractor, we freeze the CLIP ResNet-50 parameters and only train two linear layers using the standard contrastive loss in CLIP.

In summary, compared to the original CLIP, we transfer it from global image-text retrieval to region-level content retrieval. In the image captioning stage, for each foreground proposal, the object category with the highest similarity score is assigned as its label. All the retrieved labels are assembled to form the visual concept of the image.

Cross-modal Modulator. ResNet-50 backbone yields the feature map with a high channel dimension of 2048, which requires to be highly compressed before multi-modal fusion. It has been well recognized that different feature channels contain certain semantics. After extracting the visual concepts that reside in the image, we propose to utilize these textual hints to promote the visual representations. Specifically, we train a modulator that receives the concept tokens to activate the informative channels of the CLIP feature. As shown in Figure 3 (middle), this cross-modal modulator contains an embedding layer to embed the concept words, two fully-connected layers with a non-linear ReLU function to project the word embeddings, and a Sigmoid function to restrict the output weight. Finally, we average the output weights of all the concepts to obtain the final channel activation weight $\mathbf{w} \in \mathbb{R}^{1\times1\times2048}$, which is applied to the raw CLIP feature $\boldsymbol{v}$

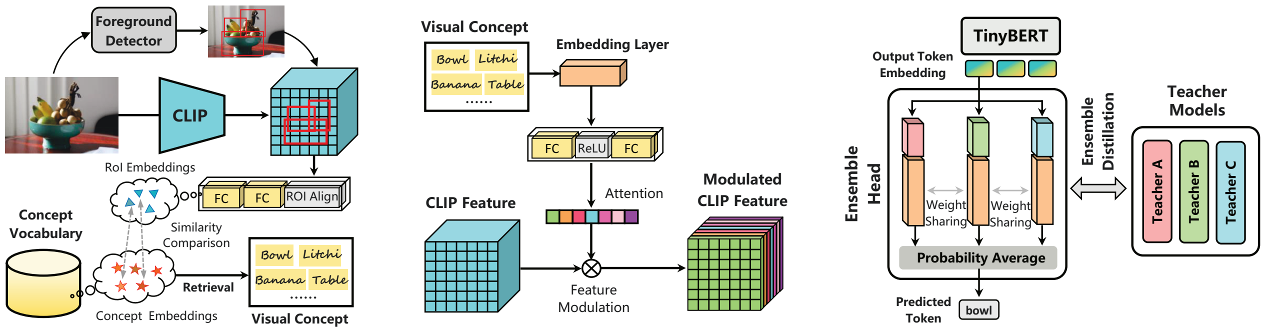

Figure 3: Left: visual concept extractor block. Middle: cross-modal modulator block. Right: ensemble head block.

to reweigh the channel importance via $\boldsymbol{v}^{\diamond} = \mathbf{w} \otimes \boldsymbol{v}$, where $\otimes$ denotes the channel-wise multiplication, and $\boldsymbol{v}^{\diamond}$ is the modulated CLIP feature.

Multi-modal Fusion Module. The proposed method adopts TinyBERT_4 (Jiao et al. 2019) as the cross-modal fusion module, which is extremely shallow consisting of only 4 transformer blocks and a hidden size of 312.

Following previous arts (Li et al. 2020b; Zhang et al. 2021a), we apply the `seq2seq` attention mask to generate the caption token in an auto-regressive way. Our TinyBERT takes as input the concatenation of the modulated image features $\boldsymbol{v}^{\diamond}$ and visual concept embeddings $\boldsymbol{c}$, and starts the caption generation by appending a mask token `[MASK]` to the inputs. Then, the previous `[MASK]` is replaced by the predicted token, and a new `[MASK]` is appended to generate the next word. The words are predicted one by one until the TinyBERT outputs the `[STOP]` token.

Ensemble Head Module. Multi-model ensemble is an intuitive way to improve the performance, but will greatly increase the model size. In this work, we propose a parameter-efficient ensemble head to predict the token. The ensemble head contains three branches to parallelly tackle the word embeddings, as shown in Figure 3 (right). We recognize that the parameter burden of head network mainly resides in the word embedding layer, whose shape is 312×30522 (dictionary size). To reduce the storage room, word embedding layers in different branches share the model weights, while the lightweight project layers (shape: 312×312) before the word embedding layer are individually optimized for diversity. These parallel head networks are individually distilled by different teacher networks to further enhance the prediction diversity, which will be discussed in the next section.

3.2 Model Training

Pre-training Stage. Most VL pre-training methods (Tan and Bansal 2019; Chen et al. 2020; Li et al. 2020b; Zhang et al. 2021a) utilize the popular masked language modeling (MLM) loss to pre-train the cross-modal fusion model. Since our work focuses on the image captioning scenario, we do not apply the bi-directional modeling manner and choose the sequence-to-sequence MLM to facilitate the text generation. To simulate the uni-directional generation process, the self-attention mask is constrained such that the caption token can only attend to the previous tokens. To be specific, we randomly mask 15% of the caption tokens following BERT and replace them with the special token `[MASK]`. The fusion model takes the Image-Concept-Caption triple $(\boldsymbol{v}^{\diamond}, \boldsymbol{c}, \boldsymbol{x})$ from the dataset $\mathcal{D}$ as input, where $\boldsymbol{x} = \{x_1, \cdots, x_T\}$ are the masked input tokens. The training objective is to reconstruct the masked token x_t based on the previous tokens $(\boldsymbol{x}_{<t})$, concepts $(\boldsymbol{c})$, and image features $(\boldsymbol{v}^{\diamond})$ by minimizing the following negative log-likelihood:

$$\mathcal{L}_{\text{caption}} = -\mathbb{E}_{(\boldsymbol{v}^{\diamond}, \boldsymbol{c}, \boldsymbol{x}) \in \mathcal{D}} \Big[\sum_t \log P(x_t | \boldsymbol{v}^{\diamond}, \boldsymbol{c}, \boldsymbol{x}_{<t}) \Big]. \quad (1)$$

Recent works (Li et al. 2020b; Zhang et al. 2021a) observe that image detection tags are qualified to serve as the anchor points to facilitate the multi-modal representation alignment. Inspired by these methods (Li et al. 2020b; Zhang et al. 2021a), we treat our retrieved visual concepts as the anchors to form the modality contrastive loss. To be specific, we "pollute" the image concept by replacing it with probability 50% with a different concept from the dataset $\mathcal{D}$. The potentially polluted image concept is denoted by $\boldsymbol{c}^{\star}$. We use a binary classifier $f(\cdot)$ on the top of the TinyBERT `[CLS]` embedding to judge whether the triple $(\boldsymbol{v}^{\diamond}, \boldsymbol{c}^{\star}, \boldsymbol{x})$ is polluted $(y = 0)$ or not $(y = 1)$. This concept contrastive loss $\mathcal{L}_{\text{concept}}$ is defined as follows:

$$\mathcal{L}_{\text{concept}} = -\mathbb{E}_{(\boldsymbol{v}^{\diamond}, \boldsymbol{c}^{\star}, \boldsymbol{x}) \in \mathcal{D}} \big[\log P(y | f(\boldsymbol{v}^{\diamond}, \boldsymbol{c}^{\star}, \boldsymbol{x})) \big]. \quad (2)$$

The aforementioned two losses are equally combined to form the final training objective in the pre-training stage: $\mathcal{L}_{\text{pre-train}} = \mathcal{L}_{\text{caption}} + \mathcal{L}_{\text{concept}}$.

Fine-tuning Stage. After model pre-training on the noisy pre-training data, our LightCap model is further fine-tuned on the well-annotated captioning dataset such as COCO. In the fine-tuning stage, we do not adopt the contrastive loss and only utilize Eq. (1) as the training objective to fully concentrate on the image captioning scenario.

3.3 Knowledge Distillation

We further adopt knowledge distillation (KD) to remedy the performance drop caused by the limited model capacity. We train teacher networks with the architecture of $\text{BERT}_{\text{base}}$, and then sequentially distill the student model.

KD in Pre-training Stage. In the pre-training stage, we first encourage the student model to mimic the transformer attentions and hidden state representations of its teacher:

$$\mathcal{L}_{\text{KD-1}} = \mathcal{L}^{\text{KD}}_{\text{atten}} + \mathcal{L}^{\text{KD}}_{\text{hidden}} = \frac{1}{h}\sum_{i=1}^{h} \texttt{MSE}\left(\mathbf{A}^{\text{S}}_i, \mathbf{A}^{\text{T}}_i\right) + \frac{1}{l}\sum_{j=1}^{l} \texttt{MSE}\left(\mathbf{H}^{\text{S}}_j \mathbf{W}, \mathbf{H}^{\text{T}}_{3\times j}\right), \tag{3}$$

where $\texttt{MSE}(\cdot,\cdot)$ denotes the mean-squared loss; $\mathbf{A}^{\text{S}}_i$ and $\mathbf{A}^{\text{T}}_i$ are the attentions from the i-th head of the student model and teacher model, respectively; $\mathbf{H}^{\text{S}}_j$ and $\mathbf{H}^{\text{T}}_{3\times j}$ denote the j-th and $(3 \times j)$-th layer's hidden state representations from the student and teacher models, respectively (we empirically adopt this setting since the teacher model is 3 times deeper than the student model); $\mathbf{W}$ is an 1×1 linear block to facilitate the student model to match its teacher's feature dimension for hidden state distillation.

After the attention and hidden representation distillations, we further perform the second-stage KD, *i.e.,* prediction-level distillation $\mathcal{L}_{\text{KD-2}}$ as follows:

$$\mathcal{L}_{\text{KD-2}} = \mathcal{L}^{\text{KD}}_{\text{caption}} + \mathcal{L}^{\text{KD}}_{\text{concept}} = \texttt{CE}\left(\mathbf{z}^{\text{S}}/\tau, \mathbf{z}^{\text{T}}/\tau\right) + \texttt{CE}\left(\mathbf{y}^{\text{S}}/\tau, \mathbf{y}^{\text{T}}/\tau\right), \tag{4}$$

where $\texttt{CE}(\cdot,\cdot)$ denotes the cross-entropy loss, $\mathbf{z}^{\text{S}}$ and $\mathbf{z}^{\text{T}}$ denote the soft predictions of the tokens of the student and teacher; $\mathbf{y}^{\text{S}}$ and $\mathbf{y}^{\text{T}}$ are the "pollution" probability of the visual concepts of the student and teacher; τ refers to the temperature in KD. In this distillation stage, the student model not only mimics the captioning capability (*i.e.,* token prediction probability) of the teacher via $\mathcal{L}^{\text{KD}}_{\text{caption}}$, but also preserves the cross-modal alignment capability (*i.e.,* concept "pollution" probability) via $\mathcal{L}^{\text{KD}}_{\text{concept}}$.

KD in Fine-tuning Stage. In the fine-tuning stage, we also first conduct knowledge distillation on attention weights and hidden states as in Eq. (3), and then conduct knowledge distillation on the output probability. However, the model fine-tuning stage only involves a simple captioning constraint without the concept contrastive learning. Consequently, we merely force the student to mimic the token prediction of its teacher via $\mathcal{L}^{\text{KD}}_{\text{caption}} = \texttt{CE}\left(\mathbf{z}^{\text{S}}/\tau, \mathbf{z}^{\text{T}}/\tau\right)$.

Ensemble KD. Actually, instead of adopting a single head, we construct the ensemble head with three parallel branches. We train three teacher models with different model initializations. These teachers jointly distill different branches of the ensemble head model, as shown in Figure 3 (right).

4 Experiments

4.1 Datasets and Metrics

Pre-training Datasets. In the experiments, we collect the image-text pairs from Google Conceptual Captions (CC3M) (Sharma et al. 2018), SBU Captions (Ordonez, Kulkarni, and Berg 2011), OpenImages (Shao et al. 2019), and MS-COCO (Lin et al. 2014) to form the pre-training data. In total, our pre-training corpus consists of about 5.8M image-text pairs.

Evaluation Datasets and Metrics. We evaluate the proposed method on the COCO caption of Karpathy split (Lin et al. 2014) and nocaps validation dataset (Agrawal et al. 2019). To evaluate the quality of the generated captions, we use standard metrics in the image captioning task, including BLEU@4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), and SPICE (Anderson et al. 2016). In the captioning stage, beam search (beam size = 5) is adopted in all experiments and the maximum generation length is restricted to 20 words.

4.2 Implementation Details

Visual Encoder. We take the ResNet-50 backbone from the CLIP model (Radford et al. 2021) as the visual feature extractor, whose parameters are frozen in both pre-training and fine-tuning stages. The input image resolution is 224×224.

Visual Concept Extractor. We follow the tiny YOLOv5n and its default settings to train a binary (foreground-background) object detector. This tiny detector is trained using Visual Genome dataset (Krishna et al. 2017), where all the object bounding boxes are treated as the foreground annotations. After obtaining the foreground object detector, we train the alignment module using the region-level CLIP features and textual embeddings from the Visual Genome dataset. This alignment module only contains two linear blocks (2048×1024 and 1024×1024) and is trained for 60 epochs with a learning rate of 1×10^{-5}.

Cross-modal Modulator. The cross-modal modulator contains two sequential linear blocks with sizes of 312×39 and 39×2048. The token embedding layer in this modulator shares weights with the embedding layer in TinyBERT.

Cross-modal Fusion Model. For the TinyBERT, we initialize it with the pre-trained weights (Jiao et al. 2019). The visual concepts, as well as the caption words, are tokenized and projected via an embedding layer before being fed to the TinyBERT. The modulated visual embeddings are compressed via the 1×1 linear block to match the TinyBERT's embedding dimension. In the pre-training stage, the fusion model is trained 1.0M steps with a learning rate of 5×10^{-5} and batch size of 512. In the fine-tuning stage, the fusion model is trained 120 epochs with a learning rate of 3×10^{-5}. Except for the TinyBERT, we also train large fusion models $\text{BERT}_{\text{base}}$ (Devlin et al. 2018) following the above steps.

4.3 Ablation Study

Model Pre-training. It has been well recognized that model pre-training on large-scale image-text corpus benefits the image captioning. As shown in Table 1, for the student model with limited capacity, model pre-training significantly improves the performance by 8.0 CIDEr score.

Visual Concept Extractor. The proposed visual concept extractor provides valuable clues for image captioning via an efficient image-text retrieval manner. As shown in Table 1, for the student model, the visual concept extractor improves the captioning performance by 3.4 CIDEr score on the COCO dataset. This mechanism also improves the strong teacher model by 3.7 CIDEr score.

Cross-modal Modulator. The cross-modal modulator takes advantage of the retrieved visual concepts to modulate the raw CLIP features. As shown in Table 1, based on the student model with a visual concept extractor, the proposed cross-modal modulator further improves the captioning performance by 1.8 CIDEr score. This tiny block promotes the strong teacher model by 2.1 CIDEr score.

Student	Pre-train	Concept	Modulator	B@4	M	C	S
✓				32.1	26.9	103.6	19.9
✓	✓			33.6	27.7	111.6	20.8
✓	✓	✓		34.3	28.3	115.0	21.3
✓	✓	✓	✓	34.9	28.9	116.8	21.9
Teacher	**Pre-train**	**Concept**	**Modulator**	**B@4**	**M**	**C**	**S**
✓				34.2	28.3	113.8	21.2
✓	✓			36.2	29.0	120.5	22.1
✓	✓	✓		37.0	29.6	124.2	23.5
✓	✓	✓	✓	37.5	29.9	126.3	24.3

Table 1: Ablative study of the proposed LightCap. To better investigate the performance of each component, the student model does not employ any knowledge distillation and uses a single head model. The evaluation metrics are BLEU@4 (B@4), METEOR (M), CIDEr (C), and SPICE (S) scores on the COCO-caption Karpathy test split (Lin et al. 2014).

Pre-training			Fine-tuning		Ensemble	COCO test	
A&R	Caption	Concept	A&R	Caption	Distill	B@4	C
						34.9	116.8
✓						35.2	117.6
✓	✓					35.6	119.6
✓	✓	✓				36.2	120.8
✓	✓	✓	✓			36.4	121.9
✓	✓	✓		✓		36.8	123.4
✓	✓	✓	✓	✓		37.1	124.1
✓	✓	✓	✓	✓	✓	37.4	125.8

Table 2: Ablative study of the proposed LightCap method using different distillation techniques. "A&R", "Caption", and "Concept" denote the knowledge distillations on attention weight and hidden representation, token probability, and concept probability, respectively. Finally, we adopt the ensemble head block and leverage the ensemble distillation to optimize the overall model.

Sequential Model Distillation. In Table 2, we ablate the model knowledge distillation (KD) techniques in our approach. First, we investigate KD in the pre-training stage in Table 2 (top). In these experiments, we only adopt the standard cross-entropy optimization without any KD in the fine-tuning stage. In the pre-training stage, the "attention & representation distillation" improves 0.8 CIDEr score, and the distillation of output token probability improves 2.0 CIDEr score. Considering the characteristic of cross-modal training, we further propose to distill the soft prediction of the anchor words (*i.e.,* visual concepts), which brings an additional 1.2 CIDEr gain. This indicates the concept distillation facilitates the cross-modal alignment.

Next, we investigate KD in the model fine-tuning stage. As shown in Table 2, based on the distilled fusion model from the pre-training stage, in the fine-tuning stage, "attention & representation distillation" and "output token distillation" further improve 1.1 CIDEr and 2.6 CIDEr, respectively. Combining the above KD techniques achieves the best result of 3.3 CIDEr gain. Finally, by virtue of the model distillation in both pre-training and fine-tuning, our lightweight student model achieves a promising captioning performance of 37.1 BLEU@4 and 124.1 CIDEr, and even matches the strong teacher model (*i.e.,* 37.5 BLUE@4 and 126.3 CIDEr in Table 1).

Ensemble Model Distillation. The above experiments are based on the single head setting. Actually, our model adopts the ensemble head for superior performance. To encourage the prediction diversity, we prepare three teachers to individually distill these heads. As shown in Table 2, ensemble head module and ensemble KD improve 1.7 CIDEr.

	Img. Encoder	Concept	Modulator	Fusion
Model	ResNet50	YOLOv5n	2 FC	TinyBERT_4
Params (M)	23.5M	1.9M	9.4×10^{-2}M	14.5M
Size (MB)	56.5MB	7.6MB	0.4MB	58.0MB
FLOPs (G)	4.1G	4.5G	1.9×10^{-4}G	1.2G

Table 3: Illustration of model details including number of parameters (in M), model size (in MB), and computational complexity (FLOPs, in G) of the proposed LightCap.

4.4 Inference on the Mobile Device

Table 3 exhibits the model FLOPs and parameters of each block in the LightCap. Note that the ResNet50 backbone in CLIP adopts the half-precision model training and thus the model storage size of the visual encoder is 56.5MB. Overall, our LightCap consumes a total storage space of 112.5MB, which is affordable for most mobile devices.

Then, we test the inference latency of LightCap model on Huawei P40 smartphone with a Kirin 990 chip. To purely investigate the model inference speed, we set the beam search size to 1. It merely takes about 188ms for our light model to process a single image on the CPU from mobile devices, which meets the real-world efficiency requirements.

4.5 State-of-the-art Comparison

Comparison on Model Size and Efficiency. In Table 4, we compare our LightCap with the state-of-the-art captioning methods in terms of model size and inference efficiency in FLOPs. Most existing pre-training methods such as VLP (Zhou et al. 2020), Oscar (Li et al. 2020b), and UNIMO (Li et al. 2020a) use the Faster R-CNN as the feature extractor and a $\text{BERT}_{\text{base}}$ as the fusion model, yielding about 173M parameters and about 800G FLOPs. It is worth noting that the current performance leaders such as VinVL (Zhang et al. 2021a) and LEMON (Hu et al. 2021a) contain a huge FLOPs of more than 1000G. As illustrated in Section 4.4, the overall FLOPs of our LightCap is only 9.8G. Consequently, compared with the recent popular image captioners, our LightCap saves more than 98% of the FLOPs.

To the best of our knowledge, DistillVLM (Fang et al. 2021b) and MiniVLM (Wang et al. 2020a) are the representative lightweight image captioners in the literature. These methods design a tiny object detector called Eff-DET based on the EfficientNet (Tan and Le 2019). Nevertheless, their fusion model (*i.e.,* MiniLM (Wang et al. 2020b)) is still much larger than our TinyBERT_4. As discussed in MiniVLM, changing the fusion model from MiniLM to a

Method	Image Encoder		Fusion Model	
	Params	FLOPs	Params	FLOPs
VinVL$_B$, LEMON$_B$	141.2M	1017.2G	109M	22.5G
Oscar$_B$, VLP$_B$	63.8M	767.0G	109M	22.5G
ViTCAP, BLIP$_B$	86.4M	55.5G	109M	22.5G
DistillVLM, MiniVLM	7.5M	4.4G	33M	8.3G
LightCap (Ours)	23.5M	4.1G	14.5M	1.2G

Table 4: Comparison of different captioning methods in terms of the model structure, inference speed in FLOPs (in G), number of parameters (in M).

Method	Cross-Entropy				CIDEr Optimization			
	B@4	M	C	S	B@4	M	C	S
w/o Pre-training								
BUTD	36.2	27.0	113.5	20.3	36.3	27.7	120.1	21.4
LBPF	37.4	28.1	116.4	21.2	38.3	28.5	127.6	22.0
AoANet	37.2	28.4	119.8	21.3	38.9	29.2	129.8	22.4
X-LAN	38.2	28.8	122.0	21.9	39.5	29.5	132.0	23.4
RSTNet	-	-	-	-	39.3	29.4	133.3	23.0
DLCT	-	-	-	-	39.8	29.5	133.8	23.0
Normal model								
VLP$_B$	36.5	28.4	116.9	21.2	39.5	29.3	129.3	23.2
Oscar$_B$	36.5	30.3	123.7	23.1	40.5	29.7	137.6	22.8
UNIMO$_B$	38.8	-	124.4	-	-	-	-	-
ViTCAP	36.3	29.3	125.2	22.6	41.2	30.1	138.1	24.1
VinVL$_B$	38.2	30.3	129.3	23.6	40.9	30.9	140.4	25.1
LEMON$_B$	40.3	30.2	133.3	23.3	41.6	31.0	142.7	25.1
BLIP$_B$	39.7	-	133.3	23.3	-	-	-	-
Light model								
E2E-VLP	36.2	-	117.3	-	-	-	-	-
MiniVLM	35.6	28.6	119.8	21.6	39.2	29.7	131.7	23.5
DistillVLM	35.6	28.7	120.8	22.1	-	-	-	-
LightCap (Ours)	37.4	29.9	125.8	22.6	40.1	29.9	136.6	24.2

Table 5: Performance comparisons on the COCO Karpathy test split (Lin et al. 2014).

TinyBERT$_4$, the captioning performance will drop sharply (about 10 CIDEr). Thanks to our designed concept extractor, cross-modal modulator, and ensemble head, a lightweight TinyBERT$_4$ also works well in our framework.

Evaluation on COCO. In Table 5, we present the performance of state-of-the-art captioning methods on the COCO Karpathy test split (Karpathy and Fei-Fei 2015). These approaches are generally trained with the cross-entropy loss and further optimized with CIDEr as a reinforcement learning reward. Previous captioners without model pre-training such as BUTD, AoANet, and X-LAN mostly use the Faster R-CNN as the visual feature extractor. The proposed LightCap outperforms all previous pretraining-free algorithms.

Recent "pre-training then fine-tuning" methods typically choose the BERT model as the cross-modal fusion model. These methods struggle to achieve a fast inference speed with the large visual backbone and the heavyweight BERT model. Using similar pre-training data and the same cross-entropy optimization, our LightCap (125.8 CIDEr) is superior to the heavyweight Oscar$_B$ (123.7 CIDEr) and UNIMO$_B$ (124.4 CIDEr). Compared with other lightweight captioning methods such as MiniVLM and DistillVLM, our LightCap retains fewer parameters and FLOPs, but surpasses them by a notable margin of about 5 CIDEr score. Note that BLIP and LEMON algorithms collect large-scale high-quality pre-training datasets containing 129 and 200 million image-text pairs (more than 20× larger than ours) for pre-training, respectively. We believe that the proposed LightCap can be further improved by involving more pre-training data, which leaves as our future work.

Method	Out-of-domain		Overall	
	C	S	C	S
BUTD	31.3	8.3	55.3	10.1
BUTD + CBS	66.4	9.7	73.1	11.1
Oscar$_B$	45.3	9.7	63.8	11.2
Oscar$_B$ + CBS	77.6	10.6	81.1	11.7
VIVO$_B$	71.1	10.6	81.5	12.2
VIVO$_B$ + CBS	87.5	11.5	88.3	12.4
VinVL$_B$ + CBS	87.4	11.6	90.9	12.8
ViTCAP	78.1	11.9	89.2	12.7
ViTCAP + CBS	95.4	12.7	93.8	13.0
SimVLM$_B$ (w/ pre-train)	-	-	94.8	13.1
LEMON$_B$	62.6	10.6	79.0	12.3
LEMON$_B$ (w/ pre-train)	107.9	13.1	106.8	14.1
BLIP$_B$ (w/ pre-train)	111.5	14.2	109.6	14.7
Human Performance	95.7	14.0	87.1	14.2
LightCap (Ours)	76.5	11.2	85.1	12.3
LightCap (Ours) + CBS	90.5	11.5	90.8	12.8

Table 6: Performance comparisons on the nocaps validation split (Agrawal et al. 2019). We report the results of both without and with constrained beam search (CBS) decoding.

Evaluation on Nocaps. Nocaps benchmark (Agrawal et al. 2019) contains 15,100 images collected from OpenImages (Shao et al. 2019). We evaluate the proposed method on the nocaps dataset to assess the model generalizability. Due to the limited space, we only present the out-of-domain and overall performance in Table 6. Following the protocol of this benchmark, we merely train the LightCap model on the COCO-caption *without* additional pre-training. Our captioning model is much smaller than all the comparison methods such as VIVO and ViTCap. It is also worth mentioning that our method surpasses the human CIDEr score and even slightly outperforms the strong VinVL method in the out-of-domain, which can be largely contributed to the representational power of the CLIP feature and our designed concept extractor to retrieve novel concepts.

5 Conclusion

In this paper, we propose a lightweight image captioning approach for resource-limited devices. To unveil the potential of a capacity-limited tiny model, we design a visual concept extractor, a cross-modal modulator, and an ensemble head to improve the captioning quality. By virtue of the sequential knowledge distillation and ensemble distillation, our LightCap exhibits competitive performance under a limited model capacity. Extensive experiments verify the super-balanced performance and efficiency of the proposed LightCap.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019. Nocaps: Novel object captioning at scale. In *ICCV*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop*.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Cornia, M.; Baraldi, L.; Fiameni, G.; and Cucchiara, R. 2021. Universal captioner: Long-tail vision-and-language model training through content-style separation. *arXiv preprint arXiv:2111.12727*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Liu, Z.; Zeng, M.; et al. 2021. An Empirical Study of Training End-to-End Vision-and-Language Transformers. *arXiv preprint arXiv:2111.02387*.

Fang, Z.; Wang, J.; Hu, X.; Liang, L.; Gan, Z.; Wang, L.; Yang, Y.; and Liu, Z. 2021a. Injecting semantic concepts into end-to-end image captioning. *arXiv preprint arXiv:2112.05230*.

Fang, Z.; Wang, J.; Hu, X.; Wang, L.; Yang, Y.; and Liu, Z. 2021b. Compressing visual-linguistic model via knowledge distillation. *arXiv preprint arXiv:2104.02096*.

Fei, Z. 2022. Attention-Aligned Transformer for Image Captioning. In *AAAI*.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Hu, X.; Gan, Z.; Wang, J.; Yang, Z.; Liu, Z.; Lu, Y.; and Wang, L. 2021a. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*.

Hu, X.; Yin, X.; Lin, K.; Wang, L.; Zhang, L.; Gao, J.; and Liu, Z. 2021b. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. In *AAAI*.

Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019. Attention on attention for image captioning. In *ICCV*.

Huang, Z.; Zeng, Z.; Huang, Y.; Liu, B.; Fu, D.; and Fu, J. 2021. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *AAAI*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*.

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

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

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

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086*.

Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; and Wang, H. 2020a. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Luo, Y.; Ji, J.; Sun, X.; Cao, L.; Wu, Y.; Huang, F.; Lin, C.-W.; and Ji, R. 2021. Dual-level collaborative transformer for image captioning. In *AAAI*.

Mukherjee, S.; and Awadallah, A. 2020. XtremeDistil: Multi-stage distillation for massive multilingual models. *arXiv preprint arXiv:2004.05686*.

Ordonez, V.; Kulkarni, G.; and Berg, T. 2011. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020. X-linear attention networks for image captioning. In *CVPR*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Qin, Y.; Du, J.; Zhang, Y.; and Lu, H. 2019. Look back and predict forward in image captioning. In *CVPR*.

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

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6): 1137–1149.

Shao, S.; Li, Z.; Zhang, T.; Peng, C.; Yu, G.; Zhang, X.; Li, J.; and Sun, J. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *ICCV*.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Shen, S.; Li, L. H.; Tan, H.; Bansal, M.; Rohrbach, A.; Chang, K.-W.; Yao, Z.; and Keutzer, K. 2021. How much can CLIP benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*.

Song, Z.; Zhou, X.; Mao, Z.; and Tan, J. 2021. Image captioning with context-aware auxiliary guidance. In *AAAI*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

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

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020b. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *arXiv preprint arXiv:2002.10957*.

Wang, Y.; Xu, J.; and Sun, Y. 2022. End-to-End Transformer Based Model for Image Captioning. In *AAAI*.

Wang, Z.; Yu, J.; Yu, A. W.; Dai, Z.; Tsvetkov, Y.; and Cao, Y. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

Xu, H.; Yan, M.; Li, C.; Bi, B.; Huang, S.; Xiao, W.; and Huang, F. 2021. E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning. *arXiv preprint arXiv:2106.01804*.

Yang, X.; Liu, Y.; and Wang, X. 2022. Reformer: The relational transformer for image captioning. In *ACM MM*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021a. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*.

Zhang, X.; Sun, X.; Luo, Y.; Ji, J.; Zhou, Y.; Wu, Y.; Huang, F.; and Ji, R. 2021b. RSTNet: Captioning with Adaptive Attention on Visual and Non-Visual Words. In *CVPR*.

Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; and Gao, J. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.