Scheduled Sampling in Vision-Language Pretraining with Decoupled Encoder-Decoder Network

Yehao Li [1], Yingwei Pan [1], Ting Yao [1], Jingwen Chen [2], Tao Mei [1]

[1] JD AI Research, Beijing, China
[2] Sun Yat-sen University, Guangzhou, China
{yehaoli.sysu, panyw.ustc, tingyao.ustc, chenjingwen.sysu}@gmail.com, tmei@jd.com

Abstract

Despite having impressive vision-language (VL) pretraining with BERT-based encoder for VL understanding, the pretraining of a universal encoder-decoder for both VL understanding and generation remains challenging. The difficulty originates from the inherently different peculiarities of the two disciplines, e.g., VL understanding tasks capitalize on the unrestricted message passing across modalities, while generation tasks only employ visual-to-textual message passing. In this paper, we start with a two-stream decoupled design of encoder-decoder structure, in which two decoupled cross-modal encoder and decoder are involved to separately perform each type of proxy tasks, for simultaneous VL understanding and generation pretraining. Moreover, for VL pretraining, the dominant way is to replace some input visual/word tokens with mask tokens and enforce the multimodal encoder/decoder to reconstruct the original tokens, but no mask token is involved when fine-tuning on downstream tasks. As an alternative, we propose a primary scheduled sampling strategy that elegantly mitigates such discrepancy via pretraining encoder-decoder in a two-pass manner. Extensive experiments demonstrate the compelling generalizability of our pretrained encoder-decoder by fine-tuning on four VL understanding and generation downstream tasks. Source code is available at https://github.com/YehLi/TDEN.

Introduction

Vision and language are two fundamental capabilities of human intelligence. The interactions in between support the series of uniquely human capacity, such as visual-language (VL) understanding (e.g., visual question answering (Antol et al. 2015; Anderson et al. 2018)), and VL generation (e.g., image captioning (Vinyals et al. 2015; Pan et al. 2020b; Li et al. 2019c; Yao et al. 2017a) and video captioning (Chen et al. 2019a; Li et al. 2018; Xu et al. 2016; Pan et al. 2016, 2017)). Inspired by recent development of natural language pretraining (e.g., BERT (Devlin et al. 2019)), there has been a steady momentum of breakthroughs that push the limits of VL tasks via VL pretraining. Most of existing VL pretraining techniques (e.g., ViLBERT (Lu et al. 2019), LXMERT (Tan and Bansal 2019), VL-BERT (Su et al. 2019)) focus on learning a universal multi-modal encoder for VL understanding downstream tasks, which can

not be directly applied to VL generation task as the lack of a decoder for language modeling. Nevertheless, how to learn a pre-trainable encoder-decoder over large-scale VL benchmarks that can be fine-tuned on both VL understanding and generation downstream tasks is a seldom explored territory.

In this work, we target for pretraining a universal encoder-decoder structure, that facilitates both VL understanding and generation downstream tasks. This direction has recently been studied in Unified VLP (Zhou et al. 2020) by using a single-stream BERT-based encoder-decoder with a joint pretraining of VL understanding proxy task (MLM: masked language modeling) and generation proxy task (MSG: masked sentence generation). Nevertheless, the design of single-stream structure processes both modality inputs via the same transformer blocks, leaving the inherent different peculiarity of each modality and each VL proxy task not fully exploited. That severely limits the generalization of pre-trained encoder-decoder across different kinds of VL downstream tasks (as observed in (Zhou et al. 2020) that joint pre-training with MLM & MSG is less effective than separate pre-training with MLM/MSG on each downstream task). Instead, this paper considers a two-stream decoupled design of encoder-decoder, aiming to exploit the mutual relationship between different modalities/VL proxy tasks for enhancing the robustness on VL downstream tasks. Our design, the Two-stream Decoupled Encoder-decoder Network (**TDEN**), consists of two encoders to process each modality inputs, together with the decoupled cross-modal encoder and decoder for each kind of proxy task. In particular, the shared object/sentence encoder independently learns the representations of each modality via intra-interaction, which offers a fertile ground for multi-modal reasoning that supports both VL understanding and generation. Moreover, during pretraining, MLM makes the predictions of masked word/visual tokens in a single shot based on unrestricted message passing between two modalities, while MSG requires the autoregressive reconstruction of input sentence and solely triggers the visual-to-textual message passing. In view of such fundamental differences in between, we involve two decoupled cross-modal encoder and decoder that separately perform each kind of proxy task in a multi-task fashion.

In the meantime, we notice that several VL proxy tasks, e.g., MLM and masked object classification (MOC), have become a default choice for VL pretraining, which replace
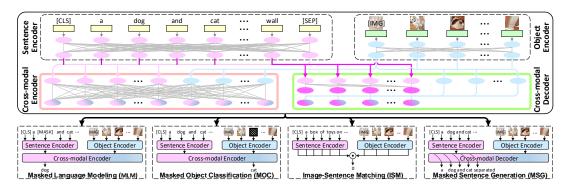
Figure 1: **Top**: An overview of our TDEN with a two-stream decoupled architecture design, consisting of four modules (i.e., the shared object and sentence encoders for contextually encoding each modality inputs, cross-modal encoder for VL understanding, and cross-modal decoder for VL generation. **Below**: Four VL proxy tasks in TDEN, i.e., MLM, MOC, and ISM for VL understanding, and MSG for VL generation.

some input word/visual tokens with mask tokens ([MASK]) and enforce the network to recover the primary inputs. Though promising results are reported on downstream tasks, these mask-based VL proxy tasks inevitably introduce a discrepancy at network fine-tuning for downstream tasks, where no artificial mask token is involved. To mitigate such discrepancy (in analogy to exposure bias (Bengio et al. 2015) in RNN based sequence-to-sequence generation), we derive a particular form of scheduled sampling for VL pretraining. Technically, a two-pass pretraining scheme is devised to enable scheduled sampling in our BERT-based encoder-decoder structure. Our launching point is to additionally formulate a more practical pretraining pass by replacing the artificial mask tokens with the generated real ones, which are directly sampled from the previous predictions of cross-modal encoder/decoder in the prior pass.

The main contribution of this work is the proposal of a universal encoder-decoder structure that facilitates both VL understanding and generation tasks. This also leads to the elegant view of how a pre-trainable encoder-decoder structure should be designed for fully exploiting the mutual but also fuzzy relations between different modalities and VL proxy tasks, and how to bridge the discrepancy between pre-training and finetuning tailored to VL pretraining. Through an extensive set of experiments on four VL understanding and generation downstream tasks, we demonstrate that our pre-trained TDEN achieves new state-of-the-art performances for each task.

Related Work

Language Pretraining. Language pretraining has attracted increasing attention in NLP field, and obtained impressive performances in various natural language understanding tasks. GPT (Radford et al. 2018) is one of the early successes for language pretraining by exploiting the unidirectional word context to learn general language representations. ELMo (Peters et al. 2018) also provides the context sensitive features to downstream tasks. By integrating proxy tasks (masked language modeling and next sentence prediction) for pretraining, BERT (Devlin et al. 2019) enables the learning of bidirectional representations. Recently, XLNet

(Yang et al. 2019) upgrades BERT with a generalized autoregressive pretraining mechanism. Unlike BERT-type pretraining models which only support language understanding via one encoder, several recent works go beyond the traditional language pretraining and aim to pretrain an encoder-decoder network for language generation through generative proxy tasks (e.g., masked sequence to sequence learning in MASS (Song et al. 2019), sequence-to-sequence modeling in UNILM (Dong et al. 2019), and denoising sequence-to-sequence modeling in BART (Lewis et al. 2019)). Our work pursues their vision-language counterpart by pretraining a universal encoder-decoder structure and finetuning it on both VL understanding and generation tasks.

Vision-Language Pretraining. Inspired by language pretraining, the research community starts to pay more attention to VL pretraining in multi-modal scenario. In particular, VisualBERT (Li et al. 2019b) directly migrates BERT to VL pretraining by involving visually-grounded proxy tasks (e.g., MLM coupled with image and image-sentence matching (ISM)). After that, a new VL understanding proxy task, named masked object classification (MOC), is widely adopted in VL pretraining techniques (UNITER (Chen et al. 2019b), Unicoder-VL (Li et al. 2019a), and VL-BERT (Su et al. 2019)), which further enhances the region-level vision-language alignment. Different from BERT-type techniques that capitalize on a single-stream encoder, ViLBERT (Lu et al. 2019) and LXMERT (Tan and Bansal 2019) utilize a more detailed two-stream encoder structure for VL pretraining, which consists of two separate encoders and one cross-modal encoder. Most recently, unlike the aforementioned methods that only support VL understanding tasks via a multi-modal encoder, Unified VLP (Zhou et al. 2020) learns a single-stream encoder-decoder that can be generalized to both VL understanding and generation tasks. In this paper, we also focus on the latter challenging task in VL pretraining. Instead of learning a single-stream encoder-decoder structure in Unified VLP, we utilize a two-stream decoupled design of encoder-decoder that reflects the mutual relationship between different modalities/VL proxy tasks. In addition, a novel scheduled sampling strategy is proposed to mitigate the pretraining-finetuning discrepancy when pretraining the universal encoder-decoder.
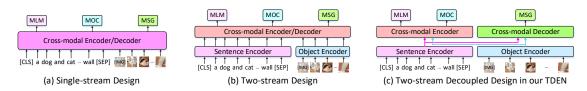
Figure 2: Three architectures for VL pretraining: (a) single-stream design (e.g., VL-BERT (Su et al. 2019), Unified VLP (Zhou et al. 2020)); (b) two-stream design (e.g., ViLBERT (Lu et al. 2019), LXMERT (Tan and Bansal 2019)); (c) two-steam decoupled design (our TDEN).

Approach

We present a pre-trainable Two-stream Decoupled Encoder-decoder Network (TDEN) that facilitates both vision-language understanding and generation tasks, as shown in Figure 1. In this section, we firstly elaborate the rationale behind our two-stream decoupled design of encoder-decoder, followed by a brief introduction of four VL understanding and generation proxy tasks adopted in our TDEN. Finally, a primary scheduled sampling strategy tailored to VL pretraining is described in detail.

Architecture Design

The core architectures in existing VL pretraining techniques for VL understanding tasks can be briefly grouped into two directions: single-stream design (e.g., Unicoder-VL (Li et al. 2019a) and VL-BERT (Su et al. 2019)) and two-stream design (e.g., ViLBERT (Lu et al. 2019) and LXMERT (Tan and Bansal 2019)). As depicted in Figure 2 (a), the single-stream design directly fuses the primary multi-modal inputs from the start, and leaves the inherent peculiarity of each modality unexploited. Instead, the two-stream design (Figure 2 (b)) independently processes each modality via separate encoders, followed with a shared cross-modal encoder for cross-modal interaction. Such design enables a flexible and specialized encoding process for each modality, and fully modulates features of each modality into high-level semantic features for multi-modal reasoning. Nevertheless, when directly migrating existing single-stream/two-stream design into the pretraining of encoder-decoder for both VL understanding and generation tasks (e.g., Unified VLP (Zhou et al. 2020)), two different kinds of VL proxy tasks (MLM and MSG) are enforced to be learnt through a shared cross-modal encoder-decoder. This way apparently ignores the inherently different peculiarity of each VL proxy task (e.g., MLM capitalizes on the unrestricted message passing between two modalities, while MSG only exploits the visual-to-textual message passing), and thus severely limits the generalization of pre-trained encoder-decoder across different types of VL downstream tasks.

Accordingly, we design a specific two-stream decoupled encoder-decoder structure (as shown in Figure 2 (c)) with two principles: (*i*) we want the encoders of each modality to be independent to each other for capturing intra-interactions, but shared across different proxy tasks to offer a fertile ground for multi-modal reasoning; (*ii*) we want to decouple the two kinds of proxy tasks during pretraining by involving two decoupled cross-modal encoder and decoder, and each one is in charge of one kind of proxy task. The detailed architecture of our TDEN is illustrated as following:

Notations. The inputs of our TDEN are image-sentence pairs $\{\mathcal{I}, \mathcal{S}\}$ derived from a large-scale image-sentence benchmark (Conceptual Captions (Sharma et al. 2018)). Each input image $\mathcal{I}$ is initially represented as a group of detected image regions (i.e., visual tokens) $\mathcal{R}_I = \{\boldsymbol{r}_i\}_{i=1}^{N_I}$, which are generated via object detector (Faster R-CNN (Ren et al. 2015)). For each corresponding sentence $\mathcal{S}$, we represent it as a sequence of word tokens $\mathcal{W}_S = \{\boldsymbol{w}_j\}_{j=1}^{N_S}$. Note that both of the input visual and word tokens are position-aware features by additionally exploiting the 2D/1D position information of each region/word as in (Lu et al. 2019).

Object Encoder and Sentence Encoder. In our TDEN, both of object encoder and sentence encoder are implemented as a series of transformer blocks, that independently encode the inputs of each modality with intra-modal contextual information. Formally, for object encoder, we firstly integrate the primary visual tokens with a special visual token [IMG] (the mean-pooling object representation over all detected regions) as the input visual token sequence ($\tilde{\mathcal{R}}_I = \{\boldsymbol{r}_i\}_{i=0}^{N_I}$). Here [IMG] indicates the beginning of visual token sequence. After that, K_I stacked transformer blocks are utilized to perform self-attention over input visual token sequence, aiming to enhance each visual token representation with intra-modal context information mined from the pervious transformer block. Similarly, in sentence encoder, we leverage K_S stacked transformer blocks to capture the intra-modal context information within input word token sequence. Two special tokens [CLS] and [SEP] are included to indicate the beginning and ending of the input word token sentence ($\tilde{\mathcal{W}}_S = \{\boldsymbol{w}_j\}_{j=0}^{N_S+1}$).

Cross-modal Encoder. The cross-modal encoder is devised to fully capture the inter-modal interaction across different modalities for VL understanding. Technically, given the enhanced visual and word tokens from each encoder ($\mathcal{H}_S^0 = \{\boldsymbol{h}_j\}_{j=0}^{N_S+1}$, $\mathcal{H}_I^0 = \{\boldsymbol{h}_i\}_{i=0}^{N_I}$), we first concatenate them as multi-modal input: $\mathcal{H}_{SI}^0 = [\mathcal{H}_S^0, \mathcal{H}_I^0]$. We further feed $\mathcal{H}_{SI}^0$ into a stack of K_E transformer blocks, leading to the contextual multi-modal features of each visual/word token ($\mathcal{H}_{SI}^E = \{\boldsymbol{h}_i^E\}_{i=0}^{N_I+N_S+2}$).

Cross-modal Decoder. We additionally involve a cross-modal decoder that learns to auto-regressively reconstruct the input sentence word-by-word conditioned on the input image, which mimics the sequence generation process and thus supports VL generation downstream task. Here we implement the cross-modal decoder by stacking K_D transformer-based decoder layers (Vaswani et al. 2017). In particular, each transformer-based decoder layer first col-
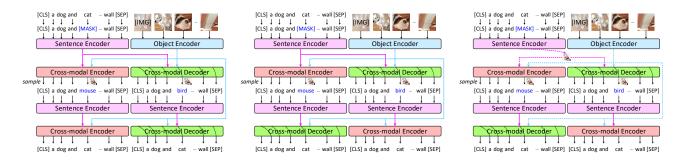
Figure 3: Three designs of two-pass pretraining schemes for VL pretraining with scheduled sampling. **Left (Two-Pass-A)**: In the second pass, we replace the input masked word tokens of cross-modal encoder/decoder with the generated ones sampled from the corresponding output word distributions of cross-modal encoder/decoder in the first pass. **Middle (Two-Pass-B)**: In the second pass, we exchange the input word token sequences of cross-modal encoder and decoder in Two-Pass-A. **Right (Two-Pass-C)**: Two-Pass-C is a cost-efficient version of Two-Pass-B by involving a switch between sentence encoder and cross-modal encoder/decoder in the first pass, aiming to randomly choose only the cross-modal encoder or decoder (50% probability) for pretraining. Therefore, in the second pass, only the cross-modal decoder/encoder is activated for a more practical pretraining.

lects contextual information from all the "past" word tokens via self-attention, and then performs cross-attention over all visual tokens for next word prediction.

VL Understanding and Generation Proxy Tasks

When pretraining our TDEN, we consider four specific objectives with regard to VL understanding and generation proxy tasks: masked language modeling (MLM), masked object classification (MOC), and image-sentence matching (ISM) that pursue VL alignment at both region-level and image-level for VL understanding, and masked sentence generation (MSG) that simulates the auto-regressive process of language modeling for VL generation.

Masked Language Modeling (MLM). Motivated by single-modal MLM proxy task in BERT, we leverage a multi-modal version of MLM for pretraining. The objective of MLM is to recover the masked word tokens depending on the unmasked word tokens and visual tokens. In particular, we randomly mask the input word tokens (15% probability) by replacing each masked word token with a mask token [MASK]. The output contextual multi-modal features of masked word tokens in cross-modal encoder are thus utilized to re-produce the original word tokens via a classifier over the whole vocabulary, driven by softmax cross-entropy loss.

Masked Object Classification (MOC). In analogy to MLM, MOC is additionally involved to enable the reconstruction of objects reflected in masked visual tokens conditioned on the unmasked visual tokens and word tokens, which further enhances the region-level VL grounding. Similarly, the input visual tokens are randomly replaced (15% probability) with mask token. We further feed the output contextual multi-modal features of masked visual tokens into a classifier for object classification. The objective of MOC is expressed as KL divergency loss that measures the mismatch between the estimated object distribution and the ground-truth object distribution (directly obtained from the off-the-shelf object detector) for each masked visual token.

Image-Sentence Matching (ISM). Different from MLM and MOC that only exploit the local multi-modal contextual information, we involve ISM to strengthen image-sentence alignment for VL understanding. Nevertheless, the typical implementation of ISM (e.g., (Lu et al. 2019)), i.e., capitalizing on the outputs of cross-modal encoder to measure image-sentence similarity, would hurt performances on VL understanding tasks. As pointed in (Su et al. 2019), one probable reason for the performance drop is that the introduced mismatched image-sentence pairs through shared cross-modal encoder would hamper the pretraining of other proxy tasks. Instead, we directly measure the image-sentence similarity for ISM based on the outputs of object and sentence encoders, which triggers an earlier image-sentence alignment and thus avoids the negative impacts of mismatched image-sentence pairs on cross-modal encoder. In particular, we feed the outputs of object and sentence encoders into an attention-based two-layer MLP (Yu et al. 2019) to calculate the image-sentence similarity. The objective of ISM is measured as triplet ranking loss, where the negative pair is generated via Multibatch strategy (Tadmor et al. 2016).

Masked Sentence Generation (MSG). Inspired by (Pan et al. 2020a; Zhou et al. 2020), MSG is introduced to teach cross-modal decoder how to auto-regressively decode the input sentence word-by-word conditioned on input image, endowing TDEN with the capability of sentence generation. The objective of MSG is measured as the joint negative log probability for reconstructing the sequential word tokens depending on all the "past" word tokens and input image.

Scheduled Sampling in VL Pretraining

The most typical way to pretrain TDEN is to optimize the whole encoder-decoder structure with all the four objectives of VL understanding and generation proxy tasks simultaneously. Nevertheless, when taking a close look on the objectives, we find that most of them correspond to mask-driven proxy tasks (e.g., MLM and MOC), which corrupt the multi-modal inputs by replacing the original word/visual tokens with artificial mask tokens ([MASK]). This inevitably results in a discrepancy at the subsequent finetuning process for downstream tasks, since no artificial mask token

is involved during finetuning. Inspired by scheduled sampling in RNN based sequence modeling (Bengio et al. 2015), we derive a particular form of scheduled sampling to elegantly mitigate such pretraining-finetuning discrepancy via pretraining our TDEN in a two-pass manner. Specifically, a two-pass pretraining scheme is designed to trigger scheduled sampling during pretraining. In the first pass, a standard VL pretraining is performed with the masked multi-modal inputs, leading to the estimated word distribution of each masked word token in cross-modal encoder/decoder. Next, the second pass conducts a more practical VL pretraining by replacing the masked word tokens with plausible alternatives sampled from the first pass.

The detailed implementation of our two-pass pretraining scheme can be varied by adopting different sampling sources (i.e., each alternative input word token of encoder/decoder in the second pass can be sampled from the output distribution of cross-modal encoder or decoder in the first pass). Hence here we derive three different two-pass pretraining schemes as depicted in Figure 3, respectively, named as Two-Pass-A to Two-Pass-C. Detailed comparisons in between are provided as follows:

Two-Pass-A. In the first scheme, we perform a standard pretraining on TDEN in the first pass, and obtain the output word distribution for each position of masked word token via cross-modal encoder and decoder. After that, we corrupt the input masked word token sequence by replacing the artificial mask tokens with the generated ones sampled from the output distributions from cross-modal encoder or decoder in the first pass, leading to two unmasked word token sequences $(\mathcal{S}_E, \mathcal{S}_D)$. Next, in the second pass, after encoding $\mathcal{S}_E$ via sentence encoder, we feed the outputs (coupled with the encoded visual tokens in the first pass) into cross-modal encoder to perform MLM and MOC proxy tasks. Meanwhile, the encoded word tokens of $\mathcal{S}_D$ are fed into cross-modal decoder to perform MSG proxy task conditioned on the input image. We jointly optimize our TDEN through the two passes and the overall objective is measured as:

$$\mathcal{L}_{TDEN}^A = \mathcal{L}_{TDEN} + \mathcal{L}_{\text{MLM}}(\mathcal{S}_E, \mathcal{I}) + \mathcal{L}_{\text{MOC}}(\mathcal{S}_E, \mathcal{I}) \\ + \mathcal{L}_{\text{MSG}}(\mathcal{S}_D, \mathcal{I}), \quad (1)$$

where $\mathcal{L}_{TDEN}$ denotes the objective of a standard VL pretraining with all the four proxy tasks in the first pass, $\mathcal{L}_{\text{MLM}}$, $\mathcal{L}_{\text{MOC}}$, and $\mathcal{L}_{\text{MSG}}$ represents the objective in MLM, MOC, and MSG, respectively.

Two-Pass-B. Two-Pass-B is a variant of Two-Pass-A by exchanging the input unmasked word token sequences of cross-modal encoder and decoder in the second pass. That is, we feed encoded word tokens of $\mathcal{S}_E/\mathcal{S}_D$ into cross-modal decoder/encoder. The overall objective is thus calculated as:

$$\mathcal{L}_{TDEN}^B = \mathcal{L}_{TDEN} + \mathcal{L}_{\text{MSG}}(\mathcal{S}_E, \mathcal{I}) + \mathcal{L}_{\text{MLM}}(\mathcal{S}_D, \mathcal{I}) \\ + \mathcal{L}_{\text{MOC}}(\mathcal{S}_D, \mathcal{I}). \quad (2)$$

Two-Pass-C. Due to an additional pass without mask tokens is involved in our two-pass pretraining scheme, the pretraining cost of Two-Pass-A and Two-Pass-B is doubled against a single standard VL pretraining. Therefore, we design a cost-efficient version of our two-pass pretraining scheme (Two-Pass-C), which slightly modifies Two-Pass-B by involving a switch between sentence encoder and cross-modal encoder/decoder in the first pass. The switch is utilized to randomly choose only one path (cross-modal encoder/decoder) for pretraining in the first pass. Next, when one path (e.g., cross-modal encoder) is selected, only the corresponding path (e.g., cross-modal decoder) is activated with the input unmasked word token sequence in the second pass. As such, we effectively reduce the pretraining cost by approximately a factor of 2. Accordingly, the overall objective can be expressed as:

$$\mathcal{L}_{TDEN}^C = \alpha[\mathcal{L}_{\text{MLM}}(\mathcal{S}, \mathcal{I}) + \mathcal{L}_{\text{MOC}}(\mathcal{S}, \mathcal{I}) + \mathcal{L}_{\text{MSG}}(\mathcal{S}_E, \mathcal{I})] \\ + (1 - \alpha)[\mathcal{L}_{\text{MSG}}(\mathcal{S}, \mathcal{I}) + \mathcal{L}_{\text{MLM}}(\mathcal{S}_D, \mathcal{I}) + \mathcal{L}_{\text{MOC}}(\mathcal{S}_D, \mathcal{I})] \quad (3) \\ + \mathcal{L}_{\text{ISM}}(\mathcal{S}, \mathcal{I}),$$

where $\alpha \in \{0, 1\}$ indicates the switching between cross-modal encoder and decoder in the first pass.

Experiments

Experimental Settings

Pretraining Data and Details. We conduct the experiments for pretraining over the large-scale image captioning benchmark—Conceptual Captions (Sharma et al. 2018). Conceptual Captions contains 3.3 million image-sentence pairs, which are automatically collected from billions of webpages. During pretraining, the adopted off-the-shelf Faster-RCNN is pre-trained on ImageNet (Deng et al. 2009) and Visual Genome (Krishna et al. 2017) as in (Anderson et al. 2018). At most 100 image regions with detection confidences higher than 0.2 are selected as inputs. Each input image region representation is a 2,048-dimensional vector. The number of stacked transformer-based layers in object encoder, sentence encoder, cross-modal encoder, and cross-modal decoder is set as in ViLBERT (Lu et al. 2019). We implement the whole architecture with PyTorch (Paszke et al. 2019), optimized with Adam (Kingma and Ba 2015) on 16 Tesla P40 GPUs. The mini-batch size is 1,024 and learning rate is set as 0.0001. The maximum iteration is 10 epoches.

Finetuning Data and Details on Downstream Tasks. In **Visual Question Answering (VQA)**, the model predicts an answer to the given question with regard to an image. VQA 2.0 (Antol et al. 2015) is adopted for finetuning our TDEN, which consists of 1.1 million questions about images in COCO (Chen et al. 2015). During finetuning, we follow the official split (Anderson et al. 2018) and formulate this task as a multi-label classification problem. In particular, by feeding the input image-question pair into TDEN, we learn the holistic image-question feature based on the multi-modal outputs of cross-modal encoder & decoder via attention mechanism (Yu et al. 2019). The holistic image-question feature is further embedded into 3,129 possible answers via a fully-connected layer with sigmoid function. As in (Anderson et al. 2018), we optimize the output answer predictions with regard to the soft answer labels via cross-entropy loss (mini-batch size: 96, learning rate: 0.00005, maximum iteration: 20 epoches).

Table 1: Performance comparison with state-of-the-art (**SOTA**) task-specific models and VL pretraining techniques (**Pre-E**: pre-trainable encoder; **Pre-ED**: pre-trainable encoder-decoder) on four VL downstream tasks. TDEN$^-$ is one variant of our TDEN which is directly trained with task-specific training data, without pretraining.* indicates using a stronger backbone (ResNeXt).

	Model	Image Captioning (test 5k/test server c40)				VQA (test-dev/test-std)				CIR			VCR (dev/test)		
		B@4	M	C	S	Overall	Yes/No	Number	Other	R1	R5	R10	Q→A	QA→R	Q→AR
SOTA	BUTD (Anderson et al. 2018)	36.3	27.7	120.1	21.4	65.3/65.7	81.8/82.2	44.2/43.9	56.1/56.3	-	-	-	-	-	-
	AoANet (Huang et al. 2019)	38.9	29.2	129.8	22.4	-	-	-	-	-	-	-	-	-	-
	BAN (Kim, Jun, and Zhang 2018)	-	-	-	-	70.0/70.4	85.4/85.8	54.0/53.7	60.5/60.7	-	-	-	-	-	-
	DFAF (Gao et al. 2019)	-	-	-	-	70.2/70.3	86.1/-	53.3/-	60.5/-	-	-	-	-	-	-
	MCAN (Yu et al. 2019)	-	-	-	-	70.6/70.9	86.8/-	53.3/-	60.7/-	-	-	-	-	-	-
	SCAN (Lee et al. 2018)	-	-	-	-	-	-	-	-	48.6	77.7	85.2	-	-	-
	R2C (Zellers et al. 2019)	-	-	-	-	-	-	-	-	-	-	-	63.8/65.1	67.2/67.3	43.1/44.0
	TDEN$^-$	38.8	28.8	128.2	22.8	69.8/-	86.3/-	49.5/-	60.4/-	55.6	81.8	88.5	73.0/-	74.4/-	54.5/-
Pre-E	LXMERT (Tan and Bansal 2019)	-	-	-	-	72.4/72.5	-/88.2	-/**54.2**	-/63.1	-	-	-	-	-	-
	VisualBERT (Li et al. 2019b)	-	-	-	-	70.8/71.0	-	-	-	-	-	-	70.8/71.6	73.2/73.2	52.2/52.4
	ViLBERT (Lu et al. 2019)	-	-	-	-	70.6/70.9	-	-	-	58.2	84.9	91.5	72.4/73.3	74.5/74.6	54.0/54.8
	VL-BERT (Su et al. 2019)	-	-	-	-	71.2/72.2	-	-	-	-	-	-	73.8/-	74.4/-	55.2/-
Pre-ED	Unified VLP (Zhou et al. 2020) *	39.5	29.3	129.3	23.2	70.5/70.7	87.2/87.4	52.1/52.1	60.3/60.5	-	-	-	-	-	-
	TDEN (Single-stream)	38.6	28.8	128.6	22.6	70.7/-	86.9/-	53.7/-	60.6/-	56.2	84.3	90.3	73.3/-	74.9/-	55.2/-
	TDEN (Two-stream)	39.1	29.1	129.6	22.9	70.9/-	87.2/-	52.8/-	61.0/-	59.2	85.4	90.9	73.9/-	75.3/-	55.8/-
	TDEN (Two-stream decoupled)	**40.2**/70.7	**29.7**/38.7	**133.4**/130.4	**23.5**	**72.5**/72.8	**88.5**/88.8	**54.7**/53.8	**63.0**/63.2	**63.6**	**88.2**	**92.9**	**75.2**/75.7	**76.7**/76.4	**58.1**/58.0

Caption-based image retrieval (CIR) aims to search an image from a pool given its caption that depicts the image content. We utilize Flickr30k (Plummer et al. 2015) in this task and each image is equipped with five human-annotated sentences. We follow the commonly adopted split in (Lee et al. 2018) and formulate this task as a ranking problem that sorts images according to the image-sentence similarities, which are measured as in ISM. The whole model is optimized with triplet ranking loss (mini-batch size: 512, learning rate: 0.00002, maximum iteration: 30 epoches).

Visual commonsense reasoning (VCR) tackles two problems: visual question answering (Q→A) and answer justification (QA→R), that requires the model to predict an answer or judge the correctness of the chosen rationale respectively. Each problem is framed as multiple choice task. In addition, VCR includes a holistic setting (Q→AR) that the model should choose the right answer (from four answer choices) and then select the correct rationale for that answer (from four rationale choices). The Visual Commonsense Reasoning (VCR) benchmark (Zellers et al. 2019) is utilized for evaluation. During finetuning, we concatenate the question and each possible response (answer/rationale) as the input sentence, which is fed into TDEN along with the image. As in VQA, we obtain the holistic image-sentence feature and then utilize a linear layer to predict the score for each possible response. The final prediction (i.e., all scores of the four response choices) is thus trained with cross-entropy loss (mini-batch size: 64, learning rate: 0.00002, maximum iteration: 20 epoches).

Image captioning (IC) aims to generate the natural sentence that depicts input image. COCO (Chen et al. 2015) is utilized for fine-tuning and evaluating TDEN. We utilize the widely adopted Karpathy split (Karpathy and Fei-Fei 2015; Yao et al. 2017b, 2018, 2019) for evaluation. For finetuning, we firstly optimize the whole architecture with cross-entropy loss. The mini-batch size is 16 and the learning rate is 0.00003. We set the maximum iteration as 10 epoches. The fine-tuned TDEN is further trained with self-critical training strategy (Rennie et al. 2017), which enables the sequence-level optimization with CIDEr reward. The learning rate is 0.000005 and the maximum iteration is 30 epoches.

Performance Comparison

Table 1 summarizes the quantitative results of our TDEN on four VL understanding (VQA, CIR, and VCR) and generation (IC) downstream tasks. We compare TDEN with state-of-the-art task-specific models and VL pre-training approaches (including both pre-trainable encoder and pre-trainable encoder-decoder models) in each downstream task. Note that we additionally include three variants of our TDEN (with two-stream decoupled structure): (i) **TDEN$^-$** is directly trained with task-specific training data for each task, without pretraining on Conceptual Captions; (ii) **TDEN (Single-stream)** is implemented as a shared cross-modal encoder-decoder structure across different modalities and proxy tasks; (iii) **TDEN (Two-stream)** consists of two object and sentence encoders, followed by a shared cross-modal encoder-decoder across different proxy tasks.

Comparison with SOTA Task-specific Models. In general, under the same task-specific training setting without VL pre-training, our TDEN$^-$ achieves comparable results with other SOTA baselines on all downstream tasks. The results basically demonstrate the effectiveness of the adopted two-stream decoupled Transformer based encoder-decoder structure. Furthermore, when pretraining TDEN on Conceptual Captions and then finetuning it on task-specific data, our TDEN consistently exhibits better performances than other SOTA task-specific baselines across the most evaluation metrics on four tasks. The performance improvements generally demonstrate the key advantage of exploiting VL pretraining via TDEN, that facilitates both VL understanding and generation downstream tasks.

Table 2: Ablation study on the use of different VL proxy tasks and scheduled sampling for VL pretraining. * indicates a different implementation of ISM based on the outputs of cross-modal encoder.

#	MLM	MOC	MSG	ISM	Scheduled Sampling	Image Captioning				VQA (test-dev)				CIR			VCR		
						B@4	M	C	S	Overall	Yes/No	Number	Other	R1	R5	R10	Q→A	QA→R	Q→AR
1	✓	✓		✓		39.2	28.9	129.1	22.8	71.9	88.1	53.7	62.3	61.8	87.1	92.2	74.6	75.9	56.9
2			✓	✓		39.3	29.3	131.0	23.1	71.7	87.9	53.4	62.0	61.7	87.0	92.2	74.5	75.9	56.8
3	✓	✓	✓			39.7	29.3	131.8	23.1	72.0	88.1	54.3	62.2	60.6	86.1	91.4	74.8	76.0	57.3
4	✓	✓	✓	*		39.5	29.2	130.3	23.0	71.9	88.0	54.2	62.2	60.2	85.7	91.2	74.6	76.4	57.2
5	✓	✓	✓	✓		39.7	29.4	132.4	23.2	72.1	88.1	54.2	62.6	62.5	86.7	92.3	74.8	76.1	57.3
6	✓	✓	✓	✓	Two-Pass-A	40.2	29.7	133.4	23.5	72.5	88.5	54.7	63.0	63.6	88.2	92.9	75.2	76.7	58.1
7	✓	✓	✓	✓	Two-Pass-B	40.1	29.6	133.0	23.4	72.3	88.4	54.2	62.7	63.6	87.8	92.9	75.3	76.9	58.2
8	✓	✓	✓	✓	Two-Pass-C	40.0	29.6	133.0	23.4	72.2	88.1	54.6	62.7	63.4	87.4	92.7	75.1	77.0	58.3

Comparison with VL Pretraining Approaches. Overall, the results across all VL understanding and generation tasks consistently indicate that our TDEN exhibits better performances that other state-of-the-art pre-trainable encoder modules (e.g., ViLBERT and LXMERT) and encoder-decoder structure (Unified VLP). In the IC and VQA tasks, the CIDEr and Overall of TDEN can achieve 133.4% and 72.5%, making 4.1% and 1.3% absolute improvements over the best competitors Unified VLP and VL-BERT, respectively. In particular, all the pre-trainable encoder modules lead to a large performance boost over SOTA task-specific approaches for VL understanding tasks. However, the pre-trainable encoder modules can not be directly adapted to VL generation task (IC), which needs an additional language decoder for sentence generation. By enabling the simultaneous pretraining of encoder and decoder, Unified VLP outperforms SOTA task-specific approaches on VL generation task. However, the results of Unified VLP on VL understanding tasks are still lower than those of pre-trainable encoders (e.g., VL-BERT). The same observation is also noticed when comparing our TDEN (Single-stream) and TDEN (Two-stream) against VL-BERT. We speculate that the performance degradations on VL understanding tasks may be caused by the joint pre-training with two different kinds of VL proxy tasks (MLM and MSG) in a shared encoder-decoder structure. Furthermore, by decoupling the two kinds of proxy tasks with two decoupled cross-modal encoder and decoder during pretraining, our TDEN boosts the performances in all the VL understanding and generation downstream tasks. This confirms the effectiveness of two-stream decoupled design in our TDEN.

Ablation Analysis

To better understand our TDEN, we perform an ablation analysis in Table 2.

Pretraining Cross-modal Encoder and Decoder Simultaneously v.s. Separately. We tried pretraining the cross-modal encoder and decoder separately. In particular, the cross-modal encoder is pretrained with ISM, MLM, and MOC, which is only applied for VL understanding downstream tasks. Meanwhile, we pretrain the cross-modal decoder with ISM and MSG, which is specifically utilized for VL generation downstream task. When separating pretraining process, the final results on four downstream tasks (Row 1&2 in Table 2) are lower than simultaneous pretraining (Row 5 in Table 2). This verifies the merit of simultaneous pretraining strategy, which jointly exploits the underlying common vision-language associations across understanding and generation proxy tasks.

Performing ISM with Object/Sentence Encoders v.s. Cross-modal Encoder. We also experimented by using the multi-modal outputs of cross-modal encoder for ISM as in (Lu et al. 2019). The performances on all tasks (Row 4 in Table 2) are even lower than our design without ISM (Row 3 in Table 2). The results align with the observation in (Su et al. 2019) and show that the typical design of ISM hampers the pretraining of other proxy tasks by introducing mismatched image-sentence pairs in the shared cross-modal encoder. Instead, by capitalizing on the outputs of object and sentence encoders for ISM, our design (Row 5 in Table 2) boosts up the performances. This basically validates the effectiveness of an earlier image-sentence alignment via our adopted ISM.

Comparison of Different Two-pass Schemes for Scheduled Sampling. We finally examine how performances on downstream tasks are affected when capitalizing on different two-pass schemes for scheduled sampling in pretraining. Overall, in our TDEN, all the three two-pass schemes (Row 6-8 in Table 2) exhibit better performance than the one without any scheduled sampling strategy (Row 5 in Table 2), which clearly indicates the advantage of exploring scheduled sampling for VL pretraining. Moreover, compared to Two-Pass-A and Two-Pass-B, Two-Pass-C effectively reduces the pretraining cost with only a slight performance drop across all tasks, which is a very practical choice.

Conclusions

In this paper, we developed TDEN, a new pre-trainable encoder-decoder structure that simultaneously supports both vision-language understanding and generation downstream tasks. Particularly, instead of using the common single-stream or two-stream architecture for pretraining, TDEN capitalizes on a two-stream decoupled design of encoder-decoder to exploit the mutual but also fuzzy relations across different modalities and proxy tasks. Furthermore, a primary scheduled sampling strategy is novelly proposed to elegantly mitigate the pretraining-finetuning discrepancy for our TDEN pretraining. Extensive experiments demonstrate the compelling generalizability of TDEN by fine-tuning it to four vision-language understanding and generation tasks. More remarkably, TDEN pretraining leads to a new state-of-the-art performance on each downstream task.

Ethical Statement

Vision-language technologies (e.g., visual question answering and image captioning) have a great potential impact for instance on robotic vision or helping visually impaired people. Nevertheless, the achievements of these technologies rely heavily on the requirement of large quantities of task-specific annotations (e.g., image-question-answer triplets/image-sentence pairs) for neural model learning. This severely hinders the scalability and generalization of vision-language techniques when only limited annotations are available. To alleviate this problem, we focus on pretraining a multi-modal encoder-decoder model over an automatically collected large-scale image captioning dataset without any human annotations, which could facilitate a wide range of vision-language applications (e.g., visual question answering, caption-based image retrieval, visual commonsense reasoning, and image captioning). However, one potential risk lies in that if the use of vision-language pretraining means vision-language understanding/generation systems may now be easily developed by those with lower levels of domain or ML expertise, this could increase the risk of the vision-language model or its outputs being used incorrectly.

References

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*.

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*.

Bengio, S.; Vinyals, O.; Jaitly, N.; and Shazeer, N. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *NeurIPS*.

Chen, J.; Pan, Y.; Li, Y.; Yao, T.; Chao, H.; and Mei, T. 2019a. Temporal deformable convolutional encoder-decoder networks for video captioning. In *AAAI*.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* .

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2019b. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740* .

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei., L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Dong, L.; Yang, N.; Wang, W.; Wei, F.; Liu, X.; Wang, Y.; Gao, J.; Zhou, M.; and Hon, H.-W. 2019. Unified language model pre-training for natural language understanding and generation. In *NeurIPS*.

Gao, P.; Jiang, Z.; You, H.; Lu, P.; Hoi, S. C.; Wang, X.; and Li, H. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *CVPR*.

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

Karpathy, A.; and Fei-Fei, L. 2015. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *CVPR*.

Kim, J.-H.; Jun, J.; and Zhang, B.-T. 2018. Bilinear attention networks. In *NeurIPS*.

Kingma, D.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *ICLR*.

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* .

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *ECCV*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* .

Li, G.; Duan, N.; Fang, Y.; Jiang, D.; and Zhou, M. 2019a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *arXiv preprint arXiv:1908.06066* .

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* .

Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2018. Jointly localizing and describing events for dense video captioning. In *CVPR*.

Li, Y.; Yao, T.; Pan, Y.; Chao, H.; and Mei, T. 2019c. Pointing novel objects in image captioning. In *CVPR*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Pan, Y.; Li, Y.; Luo, J.; Xu, J.; Yao, T.; and Mei, T. 2020a. Auto-captions on GIF: A Large-scale Video-sentence Dataset for Vision-language Pre-training. *arXiv preprint arXiv:2007.02375* .

Pan, Y.; Mei, T.; Yao, T.; Li, H.; and Rui, Y. 2016. Jointly modeling embedding and translation to bridge video and language. In *CVPR*.

Pan, Y.; Yao, T.; Li, H.; and Mei, T. 2017. Video captioning with transferred semantic attributes. In *CVPR*.

Pan, Y.; Yao, T.; Li, Y.; and Mei, T. 2020b. X-Linear Attention Networks for Image Captioning. In *CVPR*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *NAACL*.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *Technical report, OpenAI* .

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-Critical Sequence Training for Image Captioning. In *CVPR*.

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*.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *ICML*.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530* .

Tadmor, O.; Rosenwein, T.; Shalev-Shwartz, S.; Wexler, Y.; and Shashua, A. 2016. Learning a Metric Embedding for Face Recognition using the Multibatch Method. In *NeurIPS*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*.

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*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *NeurIPS*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2017a. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring visual relationship for image captioning. In *ECCV*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2019. Hierarchy parsing for image captioning. In *ICCV*.

Yao, T.; Pan, Y.; Li, Y.; Qiu, Z.; and Mei, T. 2017b. Boosting image captioning with attributes. In *ICCV*.

Yu, Z.; Yu, J.; Cui, Y.; Tao, D.; and Tian, Q. 2019. Deep modular co-attention networks for visual question answering. In *CVPR*.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

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