Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks

Xizhou Zhu[1*], Jinguo Zhu[2*†], Hao Li[4*†], Xiaoshi Wu[4*†]
Hongsheng Li[4], Xiaohua Wang[2], Jifeng Dai[3✉]
[1]SenseTime Research [2]Xi'an Jiaotong University [3]Tsinghua University
[4]CUHK-SenseTime Joint Laboratory, The Chinese University of Hong Kong
zhuwalter@sensetime.com, lechatelia@stu.xjtu.edu.cn, {haoli, wuxiaoshi}@link.cuhk.edu.hk
hsli@ee.cuhk.edu.hk, xhw@mail.xjtu.edu.cn, daijifeng001@gmail.com

Abstract

Biological intelligence systems of animals perceive the world by integrating information in different modalities and processing simultaneously for various tasks. In contrast, current machine learning research follows a task-specific paradigm, leading to inefficient collaboration between tasks and high marginal costs of developing perception models for new tasks. In this paper, we present a generic perception architecture named Uni-Perceiver, which processes a variety of modalities and tasks with unified modeling and shared parameters. Specifically, Uni-Perceiver encodes different task inputs and targets from arbitrary modalities into a unified representation space with a modality-agnostic Transformer encoder and lightweight modality-specific tokenizers. Different perception tasks are modeled as the same formulation, that is, finding the maximum likelihood target for each input through the similarity of their representations. The model is pre-trained on several uni-modal and multi-modal tasks, and evaluated on a variety of downstream tasks, including novel tasks that did not appear in the pre-training stage. Results show that our pre-trained model without any tuning can achieve reasonable performance even on novel tasks. The performance can be improved to a level close to state-of-the-art methods by conducting prompt tuning on 1% of downstream task data. Full-data fine-tuning further delivers results on par with or better than state-of-the-art results. Code and pre-trained weights shall be released.

1. Introduction

Biological intelligence systems of animals perceive the world by receiving information in different modalities, integrating with the complex central nervous system, and processing simultaneously for different tasks. However, designing a generic artificial perception model that handles multiple modalities and numerous tasks has always been considered too difficult. To simplify this problem, previous machine learning research has focused on developing specialized models for inputs from certain restricted modality, *e.g.*, Convolutional Neural Networks [45] for visual recognition and Transformers [80] for natural language processing. Recently, Transformers have been proved to have competitive performance in more scenarios such as image [10, 20, 51, 76, 78, 82, 84, 90] and video [4, 6, 87] recognition, which triggers a new paradigm of designing unified architectures for different modalities. Following this paradigm, recent works [1, 27, 33, 64] adopt Transformers as the backbone for multi-modal applications such as

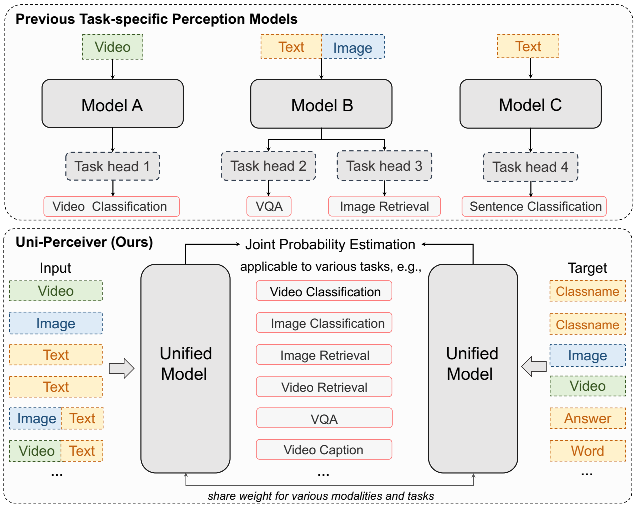

Figure 1. Comparing previous task-specific perception models with our proposed Uni-Perceiver, which processes various modalities and tasks with a single siamese model and shared parameters.

[*] Equal contribution. [†] This work is done when Jinguo Zhu, Hao Li, and Xiaoshi Wu are interns at SenseTime Research. [✉] Corresponding author.

visual-linguistic recognition. They convert the inputs from different modalities into unified input token sequences with modality-specific tokenizers. Models are pre-trained with large-scale multi-modal datasets, and then adapted to downstream tasks with fine-tuning.

Despite the ability of processing multi-modal information with unified architectures, current methods still require specific design and training for different tasks. This limitation is caused by two reasons. First, the input of a particular model is the combination of specific modalities required by its target task. Second, previous works require prediction heads specifically designed and trained for the target tasks.

We argue that this task-specific paradigm conflicts with the objective of designing generic perceptual models. Specifically, during pre-training, the specialised designs for different tasks hinder the collaboration between tasks, which may hurt the representational capacity. Meanwhile, when a pre-trained model is applied to a new task, the input format and the prediction head need to be re-designed and fine-tuned on sufficient downstream data. Considerable effort in collecting and annotating data is required. Also, all parameters need to be copied and maintained for each downstream task, which becomes inefficient and inconvenient as the number of tasks and the model size grow. On the other hand, when fine-tuning is performed with insufficient training data, it may forget the pre-trained knowledge that is beneficial to the downstream task, thereby hurting generalization performance [14]. All of these issues increase the marginal cost of developing perception models for new tasks and limit the capability to meet the rapidly growing demands of diverse scenarios, indicating task-specific paradigm is not suitable for generic perceptual modeling.

Our core idea is to replace task-specific designs by encoding different task inputs and targets from arbitrary modalities into a unified representation space, and model the joint probability of inputs and targets through the similarity of their representations. This design eliminates the gap between the formulations of different perception tasks, and therefore encourages the collaboration between different modalities and tasks in representation learning. Moreover, by aligning the formulations of pre-training and downstream tasks, the knowledge can be better transferred when applying the pre-trained model to the target tasks. The model can even conduct zero-shot inference on novel tasks that do not appear in the pre-training stage.

In this paper, we propose a unified architecture named Uni-Perceiver, which processes various modalities and tasks with a single siamese model and shared parameters. Specifically, the task inputs and targets from arbitrary combinations of modalities are first converted into unified token sequences with lightweight modality-specific tokenizers. The sequences are then encoded by a modality-agnostic Transformer encoder into a unified representation space. Different perception tasks are modeled as the same formulation, finding the maximum likelihood target for each input through the similarity of their representations, so as to facilitate the generic perceptual modeling.

Uni-Perceiver is pre-trained on various uni-modal tasks such as image / video classification and language modeling, and multi-modal tasks such as image-text retrieval and language modeling with image clues. When applied to downstream tasks, thanks to the generic modeling of perception tasks, the pre-trained model shows the ability of zero-shot inference on novel tasks that did not appear in the pre-training stage. Moreover, the performance can be further boosted with additional task-specific data. For the few-shot scenario, we adapt the model to downstream tasks with prompt tuning [47], where only a small amount of additional parameters are optimized for specific tasks. The performance of our model can be further improved with full-model fine-tuning on sufficient downstream training data.

We pre-train our model on several uni-modal and multi-modal tasks, and evaluate its performance on a variety of downstream tasks, including novel tasks that did not appear in the pre-training stage. Results show that our pre-trained model without any tuning can achieve reasonable performance even on novel tasks. Its performance can be boosted to a level close to state-of-the-art methods by conducting prompt tuning with 1% of the downstream task data. When fine-tuning the pre-trained model with 100% of the target data, our model achieves result on par with or better than state-of-the-art methods on almost all the tasks, which demonstrates the strong representation ability.

2. Related Works

Architecture. For visual recognition, Convolutional Neural Networks (CNN) [45] used to be the main architecture paradigm. Motivated by the success of Transformers in natural language processing [8, 18, 37, 40, 50, 80], attempts have been made to apply Transformers to image and video modalities. For image recognition, vision Transformers [10, 20, 51, 76, 78, 82, 84, 90] replace CNNs by an image patch tokenizer and a transformer encoder, which have been proved to obtain competitive performance as CNNs. [4, 6, 87] make attempts to apply Transformers on video recognition in a convolution-free fashion. For visual-linguistic recognition, recent works [15, 42, 43, 54, 61, 72, 75, 92] also adopt Transformers as the backbone, while they usually take regional features as inputs, which are typically extracted by off-the-shelf object detectors (*e.g.*, Faster R-CNN [65] pre-trained on Visual Genome [36]). [29] attempts to eliminate the need for object detectors by directly extracting features from the raw pixels with CNNs. [1, 27, 33, 64] take a further step by applying Transformers to raw image patches and word tokens. Transformers have enabled a unified architecture paradigm for different modalities, which only need the modality-specific tokenizers to convert inputs from different modalities into unified input token sequences.

Nevertheless, previous architecture requires prediction heads specifically designed and trained for different perception tasks. Instead, we replace the task-specific design by encoding different task inputs and targets into a unified representation space, and model their relationship by representational similarity. This modification enables our model to conduct zero-shot inference even on novel downstream tasks that did not appear in the pre-training stage.

Pre-training. Large-scale pre-training has achieved great success in the field of deep learning, which can alleviate the data-hungry challenge and improve the performance of downstream tasks [83]. For image recognition, pre-training is usually performed on image classification datasets, *e.g*., ImageNet [17]. Video recognition networks are either pre-trained on image classification or video classification datasets, *e.g*., Moments in Time [57] and Kinetics [32]. In natural language processing, self-supervised language modeling [8, 18, 37, 40, 50] is adopted for pre-training on large-scale unlabeled corpora [62]. Specifically, GPT [8] performs the auto-regressive pre-training, which optimizes the probability of the next word conditioned on previous words. BERT [18] uses masked language modeling (MLM) and next sentence prediction (NSP) for pre-training. These pre-trained models can serve as robust feature extractors for downstream tasks with small architecture modifications.

Recent years have witnessed interest in large-scale cross-modal pre-training [83]. Compared with uni-modal pre-training, cross-modal pre-training needs to align information from different modalities. Such pre-training is usually performed on image-text pairs collected from Internet [9, 31, 58, 68] and manual annotated visual-linguistic datasets [36, 46, 58]. Moreover, various pre-training objectives are proposed to utilize these datasets effectively. The most widely used objectives are image-text retrieval [2, 5, 43, 55, 63, 73, 74, 75], masked language modeling with image clues [2, 23, 43, 55, 72, 73, 74, 75], and masked region modeling [15, 55, 72, 73, 75]. Among them, masked region modeling requires regional features extracted by off-the-shelf object detectors. More recently, CLIP [63] has verified the effectiveness of only performing image-text retrieval pre-training on huge webly collected data.

Previous multi-task pre-training requires task-specific heads, which hinders the collaboration among different tasks. Instead, we encode different task inputs and targets into a unified representation space, and model their relationship by a unified representational similarity, which enables the collaboration between different modalities and tasks. Our pre-training tasks include image and video classification, language modeling with and without image clues, and image-text retrieval. We do not use regional features and the corresponding pre-training tasks.

Prompt Tuning. As an alternative solution to fine-tuning, prompt tuning has recently been proposed in the NLP community, which originated from prompting methods [47]. In prompting, specially designed natural language tokens, or namely prompts, are inserted into the input sequence as hints for the target tasks. These prompt inputs are used to query a large language model (*e.g*., GPT-3 [8]). Methods [30, 69] have been proposed to automate the prompt engineering process. The prompting process does not tune any of the parameters, which is empirically sub-optimal compared to fine-tuning [49].

Prompt tuning [39] is proposed to replace hard language prompts with learnable prompt tokens that can be updated through gradient back-propagation, while other parameters are still kept fixed. Other than adding learnable input tokens, Prefix-Tuning [44] adds learnable prompts to each layer of the Transformer to boost the model capacity. For few-shot scenario, [25] proves that prompt tuning can be much better than traditional fine-tuning. When the training data is sufficient, prompt tuning performs slightly worse than fine-tuning [49]. However, the performance gap from full-model fine-tuning closes up as the pre-trained model gets larger [39, 48]. Inspired by the success of prompt tuning in NLP, [91] applies prompt tuning to visual-linguistic pre-trained models (*e.g*., CLIP [63]) to perform few-shot image classification. [59, 66] further apply a residual feature adapter to improve the few-shot performance.

In this paper, we focus on the zero-shot and few-shot scenarios, where the downstream tasks may not even appear in the pre-training stage. For few-shot learning, we adapt the model with prompt tuning proposed by [47]. The performance of our model can be further improved by fine-tuning the whole model with sufficient downstream training data.

3. Method

3.1. Unified Architecture for Generic Perception

In this section, we will describe our unified architecture for various modalities and tasks. Fig. 2 illustrates the architecture. Specifically, the model first converts different task inputs and targets from arbitrary combinations of modalities into token sequences with modality-specific tokenizers. A modality-agnostic Transformer encoder, which shares parameters for different input modalities and target tasks, is then employed to encode different token sequences into a shared representation space. Any perception task can be modeled in a single unified formulation, which finds the maximum likelihood target for each input through the similarity of their representations.

Tokenization. Given the raw inputs from text, image, and video modalities, modality-specific tokenizers are applied

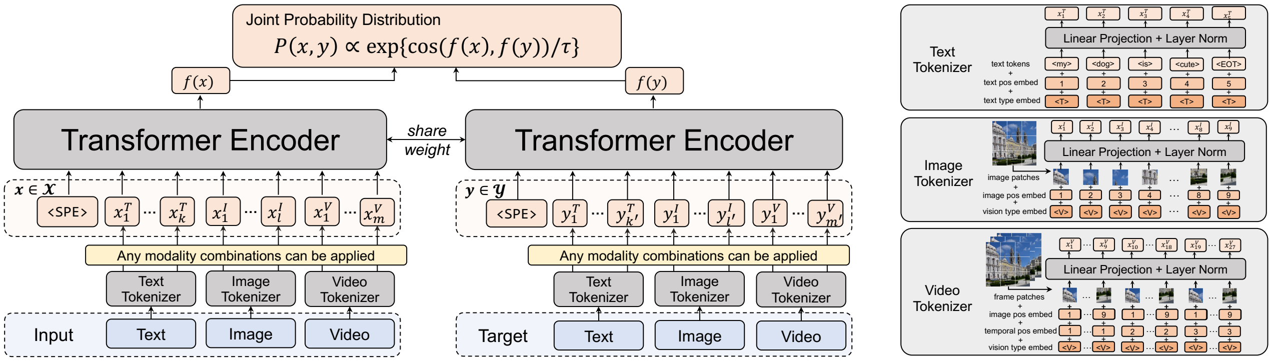
Joint Probability Distribution
$P(x,y) \propto \exp\{\cos(f(x),f(y))/\tau\}$
$f(x)$
$f(y)$
Transformer Encoder
share weight
Transformer Encoder
$x \in \mathcal{X}$
<SPE>
x_1^T ⋯ x_k^T x_1^I ⋯ x_l^I x_1^V ⋯ x_m^V
$y \in \mathcal{Y}$
y_1^T ⋯ $y_{k'}^T$ y_1^I ⋯ $y_{l'}^I$ y_1^V ⋯ $y_{m'}^V$
Any modality combinations can be applied
Text Tokenizer
Image Tokenizer
Video Tokenizer
Input
Target
Text
Image
Video
Text Tokenizer
text tokens
text pos embed
text type embed
<my> <dog> <is> <cute> <EOT>
<T>
Linear Projection + Layer Norm
x_1^T x_2^T x_3^T x_4^T x_5^T
Image Tokenizer
image patches
image pos embed
vision type embed
<V>
x_1^I x_2^I x_3^I x_4^I ⋯ x_8^I x_9^I
Video Tokenizer
frame patches
temporal pos embed
x_1^V ⋯ x_9^V x_{10}^V ⋯ x_{18}^V x_{19}^V ⋯ x_{27}^V

Figure 2. Overview of our unified architecture for generic perception. Different task inputs and targets from arbitrary modalities are converted into unified token sequences with modality-specific tokenizers. A modality-agnostic weight-sharing Transformer encoder is then applied to encode these token sequences into the shared representation space. Any perception task can be modeled as finding the maximum likelihood target for each input through the similarity of their representations.

to generate the input token sequences for the Transformer encoder. Here, we use the BPE tokenizer [67] for text modality, the image patch tokenizer [20] for image modality, and the temporal frame patch tokenizer [7] for video modality. These outputted tokens are attached with additional modality type embeddings to identify which modality the raw input belongs to. Details of the modality-specific tokenizers are described in the Appendix.

As illustrated in Fig. 2, depending on the task requirements, the input sequence x of the Transformer encoder can be composed of different combinations of text token sequence x^T, image token sequence x^I, and video token sequence x^V. At the beginning of the sequence x, a special token <SPE> is always inserted. For example, $x = [\texttt{<SPE>}, x^I, x^T]$ for image-text pair inputs, and $x = [\texttt{<SPE>}, x^V]$ for video-only inputs, where [ ] denotes the sequence concatenation. The feature of <SPE> at the encoder output serves as the representation of the input.

Generic Modeling of Perception Tasks. We model different perception tasks with a unified architecture, whose parameters are shared for all target tasks. Each task is defined with a set of inputs $\mathcal{X}$ and a set of candidate targets $\mathcal{Y}$. Given an input $x \in \mathcal{X}$, the task is formulated as finding the maximum likelihood target $y \in \mathcal{Y}$ as

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(x,y), \tag{1}$$

where $P(x,y)$ is the joint probability distribution. The joint probability is estimated through calculating the cosine similarity between the representation of x and y as

$$P(x,y) \propto \exp\left(\cos\left(f(x), f(y)\right)/\tau\right), \tag{2}$$

where $f(\cdot)$ is the Transformer encoder, and $\tau > 0$ is a learnable temperature parameter.

To obtain generic modeling capability, our unified architecture is pre-trained on a variety of multi-modal tasks simultaneously. Suppose a series of pre-training tasks is denoted as $\{\mathcal{X}_1, \mathcal{Y}_1\}, \{\mathcal{X}_2, \mathcal{Y}_2\}, ..., \{\mathcal{X}_n, \mathcal{Y}_n\}$, where $\mathcal{X}_i$ and $\mathcal{Y}_i$ is the input set and target set of the i-th task, respectively. Then the pre-training loss is defined as

$$L = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\{x,y\} \in \{\mathcal{X}_i, \mathcal{Y}_i\}} \left[-\log \frac{P(x,y)}{\sum_{z \in \mathcal{Y}_i} P(x,z)} \right], \tag{3}$$

where $\mathbb{E}$ is the mathematical expectation, and $\{x,y\} \in \{\mathcal{X}_i, \mathcal{Y}_i\}$ indicates a ground-truth input-target pair sampled from the dataset of the i-th task.

Our unified architecture is suitable for any task, as long as its input set $\mathcal{X}$ and target set $\mathcal{Y}$ are composed of images, texts, and videos. For example, the target set $\mathcal{Y}$ in classification tasks can be a set of class names, a set of class descriptions, or even a set of images with handwritten numbers representing class indexes. Detailed instances of $\mathcal{X}$ and $\mathcal{Y}$ will be introduced in the next subsection. Note that we currently focus on text, image, and video modalities, but more modalities are also applicable, as long as the corresponding tokenizers are applied.

Relation to Previous Perception Models. Our method shares the same goal of learning multi-modal representations as previous perception models. However, existing works follow a task-specific paradigm, while our method is designed for generic perceptual modeling. The main difference lies in two parts:

1) Previous works focus on inputs from certain combinations of modalities required by their target tasks, while our method handles arbitrary combinations of modalities with a unified architecture and shared parameters.

2) Previous works require prediction heads specifically designed and trained for each perception task, while our method models different tasks with the same formulation and processes them with unified modeling.

Therefore, when transferred to a new task, previous methods need to re-design their input formats and predic-

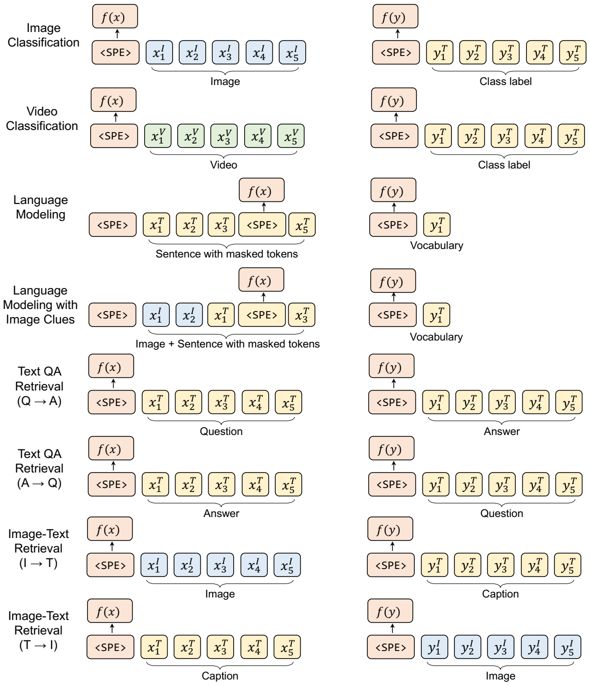
Image Classification
$f(x)$
<SPE>
x_1^I x_2^I x_3^I x_4^I x_5^I
Image
$f(y)$
<SPE>
y_1^T y_2^T y_3^T y_4^T y_5^T
Class label
Video Classification
x_1^V x_2^V x_3^V x_4^V x_5^V
Video
Class label
Language Modeling
x_1^T x_2^T x_3^T <SPE> x_5^T
Sentence with masked tokens
y_1^T
Vocabulary
Language Modeling with Image Clues
x_1^I x_2^I x_1^T <SPE> x_3^T
Image + Sentence with masked tokens
Vocabulary
Text QA Retrieval (Q → A)
Question
Answer
Text QA Retrieval (A → Q)
Answer
Question
Image-Text Retrieval (I → T)
Image
Caption
Image-Text Retrieval (T → I)
Caption
y_1^I y_2^I y_3^I y_4^I y_5^I
Image

Figure 3. Input and target formats of pre-training tasks. For each task, the left column represents the format of input sequence x, and the right column represents the format of the target sequence y. $f(x)$ and $f(y)$ indicate the representations used for calculating the joint probability distribution as in Eq. (2). Here, we have omitted the tokenizer and encoder for concision.

tion heads accordingly. Models require fine-tuning on sufficient task-specific data, resulting in remarkable human and computational costs. In contrast, our method can directly conduct zero-shot inference on novel tasks that do not appear in the pre-training stage. The performance can be further boosted with prompt tuning on few-shot downstream data and fine-tuning on sufficient downstream data.

3.2. Pre-training on Multi-Modal Tasks

Our model is pre-trained on a variety of tasks simultaneously to learn the multi-modal generic representations. The pre-training tasks are illustrated in Fig. 3. Specifically, for uni-modal pre-training tasks, we adopt the most widely-used image classification, video classification, and language modeling tasks. To further enhance the relationships between different modalities, some cross-modal tasks are also employed, such as language modeling with image clues and image-text retrieval tasks. Note that for image and video classification tasks, we regard each class name (*e.g.*, `tigershark`) as a text sequence. This provides weak supervision for bridging the gap among the representations of images, videos, and texts.

Image and Video Classification. In image and video classification tasks, $\mathcal{X}$ denotes the set of all possible images or videos in the training dataset, and $\mathcal{Y}$ consists of candidate class labels in each dataset. Each class name is regarded as a text sequence to provide weak supervision of the relationship to texts. Both the input $x \in \mathcal{X}$ and target $y \in \mathcal{Y}$ start with an `<SPE>` token, whose feature at the encoder output represents the corresponding sequence.

Language Modeling with and without Image Clues. The language modeling task aims to predict the masked words according to the context. Both auto-regressive [8] and auto-encoding [18] language modeling are adopted. When inputs have no image, the auto-regressive and auto-encoding tasks correspond to the text generation and the masked language modeling tasks, respectively. When inputs have images, the auto-regressive and auto-encoding tasks correspond to the image caption and the masked language modeling with image clues tasks, respectively.

For auto-encoding language modeling, we follow the practice in BERT [18] to mask out 15% words from the text randomly. The model predicts each masked word based on all inputs. For auto-regressive language modeling, the model predicts each word based on its previous text and image (if any). Please refer to the Appendix for an efficient implementation of auto-regressive language modeling.

In this task, $\mathcal{X}$ consists of language sentences or image-text pairs. $\mathcal{Y}$ denotes the set of all words in the vocabulary, where each word is regarded as a single text sequence. Each word that needs to be predicted in $x \in \mathcal{X}$ is replaced by a `<SPE>` token, whose feature at the encoder output is used to match the words in the vocabulary $\mathcal{Y}$.

Image and Text Retrieval. For image-text retrieval, the input sets $\mathcal{X}$ and $\mathcal{Y}$ are composed of images and text sequences respectively, or vice versa. For text-only retrieval, the input sets $\mathcal{X}$ and $\mathcal{Y}$ are both text sequences. Each sequence in $\mathcal{X}$ and $\mathcal{Y}$ has a special token `<SPE>` at the beginning, whose feature at the output of the encoder serves as the final representation.

3.3. Zero-shot, Prompt Tuning and Fine-tuning

During the pre-training stage, our unified architecture learns to model the joint distribution of input and target sequences from arbitrary modalities. Thanks to the generic perceptual modeling, our pre-trained model can perform zero-shot inference on completely novel tasks that do not appear in the pre-training stage. Our model can be further adapted to downstream tasks with task-specific additional training data. For the few-shot scenario, we employ the prompt tuning [47] scheme, which only adds a few additional task-specific parameters to the model. The performance on specific tasks can be further improved by fine-tuning the whole model on sufficient downstream data.

Zero-shot Inference on Novel Tasks. Our model has the potential to perform zero-shot inference on any perception task that can be modeled by a joint probability distribu-

Dataset	#Images	#Videos	#Text
ImageNet-21k [17]	14.2M	0	21K
Kinetics-700 [32]	0	542K	700
Moments in Time [57]	0	792K	339
Books&Wiki [93]	0	0	101M
PAQ [41]	0	0	65M
CC3M [68]	3.0M	0	3.0M
CC12M [9]	11.1M	0	11.1M
COCO Caption [12]	113K	0	567K
Visual Genome [36]	108K	0	5.41M
SBU [58]	830K	0	830K
YFCC [31]	14.8M	0	14.8M

Table 1. Pre-training dataset statistics. #Images, #Videos and #Text represent the number of images, video clips, and textual sentences (or phrases), respectively.

Method	Pre-training Data		ImageNet-1k	Kinetics-400
	#Images	#Videos	Acc	Acc
DeiT [77]	1.28M	0	81.8	-
TimeSformer [7]	1.28M	650k	-	75.5
Ours $_{\text{w/o Tuning}}$	44.14M	1.33M	78.0	73.5
Ours $_{\text{PT (0.1\%)}}$	44.14M	1.33M	79.4	73.6
Ours $_{\text{FT (0.1\%)}}$	44.14M	1.33M	78.8	73.5
Ours $_{\text{PT (1\%)}}$	44.14M	1.33M	80.2	73.6
Ours $_{\text{FT (1\%)}}$	44.14M	1.33M	80.2	73.6
Ours $_{\text{FT (100\%)}}$	44.14M	1.33M	83.8	75.8

Table 2. Image and video classification performance under different tuning settings. PT means prompt-tuning, and FT means fine-tuning. The percentage of data used in tuning is noted. In addition, the data statistics for training or pre-training are also listed.

tion. For a task with input $x \in \mathcal{X}$ and a candidate target $y \in \mathcal{Y}$, we firstly tokenize x and y into two sequences. The joint probability $P(x, y)$ is then estimated following Eq. (2). Zero-shot inference can be conducted by maximum likelihood estimation, as described in Eq. (1). Performance can also be improved through prompt engineering, similar to the prompting [47] for language models such as GPT-3 [8], where network training is not required.

Prompt Tuning. For the few-shot scenario with limited training data, we adopt prompt tuning, which is memory-efficient and has been proved to be better than the fine-tuning scheme in few-shot NLP [25]. In prompt tuning, most pre-trained parameters are fixed, leaving only a small portion of task-specific parameters to be optimized. Specifically, following P-Tuning v2 [48], learnable prompt tokens with random initialization are added at each layer of the Transformer encoder, and class labels with linear heads are added for classification tasks. The <SPE> token and layer norm parameters are also tuned. We refer the readers to the Appendix for more details.

Fine-Tuning. For downstream tasks with sufficient training data, our model can also be fine-tuned to further improve its performance. During fine-tuning, our model can serve as a joint probability estimator (same as our proposed generic perceptual modeling), or a feature extractor (same as traditional pre-trained models). Under the setting of joint probability estimation, the downstream tasks are formulated in the same unified manner as in pre-training. On the other hand, similar to previous perception models, our model can also be used as a feature extractor by adding a task-specific head on the top of the encoder. We empirically find these two schemes achieve very similar performance, and hence the scheme of joint probability distribution estimator is used by default for consistency.

4. Experiments

4.1. Datasets

Our model is pre-trained on a variety of tasks, whose statistics are listed in Tab. 1. We pre-train image classification on ImageNet-21k [17]. For video classification, we pre-train on Kinetics-700 [32] and Moments in Time [57]. We pre-train language modeling on BookCorpora [93] & English Wikipedia (Books&Wiki) and PAQ [41]. For language modeling with image clues and image-text retrieval, we use a combination of COCO Caption [13], SBU Captions (SBU) [58], Visual Genome [36], CC3M [68], CC12M [9] and YFCC [31]. To evaluate the effectiveness of our method and verify the generalization of our pre-trained model, we also use several novel datasets that did not appear in pre-training, *i.e.*, Flickr30k [60], MSVD [11], VQA [26], and GLUE [81]. See Appendix for the details of datasets.

4.2. Implementation Details

The Transformer encoder used for experiments is of the same configuration with BERT$_{\text{BASE}}$ [18]. It is a 12-layer encoder with the embedding dimension of 768 and the attention head number of 12. The hidden dimension size in FFN is 3072. We pre-train the model with multiple tasks simultaneously. In each iteration, each GPU independently samples a single task and dataset. The gradients of different GPUs are synchronized after the gradient back-propagation. We use AdamW [34] optimizer with a base learning rate of 0.0002 and a weight decay of 0.05. Gradient clipping with 5.0 is used to stabilize training. We also use drop path [38] with a probability of 0.1 during training. The model is pre-trained on 128 Tesla V100 GPUs in a distributed fashion for 500k iterations. We use the cosine learning rate schedule with 50k iterations of linear warmup. See Appendix for more implementation details.

4.3. Evaluation on Pre-training Tasks

We first evaluate our pre-trained model on tasks that have been involved in the pre-training stage, while the datasets might be different. The widely used Imagenet-1k [17] and Kinetics-400 [32] are used for evaluating the image and video classification tasks, respectively. COCO Caption and Flickr30k are the typical datasets used to evaluate the performance on image caption and image-text retrieval.

Method	Pre-training Data			Text Retrieval						Image Retrieval					
				Flickr30k			COCO Caption			Flickr30k			COCO Caption		
	#Images	#Videos	#Text	R@1	R@5	R10	R@1	R@5	R10	R@1	R@5	R10	R@1	R@5	R10
ImageBERT [61] $_{\text{w/o Tuning}}$	6.0M	0	6.0M	70.7	90.2	94.0	44.0	71.2	80.4	54.3	79.6	87.5	32.3	59.0	70.2
UNITER-B [15] $_{\text{w/o Tuning}}$	4.2M	0	9.6M	80.7	95.7	98.0	-	-	-	66.2	88.4	92.9	-	-	-
ViLT [33] $_{\text{w/o Tuning}}$	4.2M	0	9.6M	73.2	93.6	96.5	56.5	82.6	89.6	55.0	82.5	89.8	40.4	70.0	81.1
Unicoder-VL [28]	3.8M	0	3.8M	86.2	96.3	99.0	62.3	87.1	92.8	71.5	91.2	95.2	48.4	76.7	85.9
UNITER-B	4.2M	0	9.6M	85.9	97.1	98.8	64.4	87.4	93.1	72.5	92.4	96.1	50.3	78.5	87.2
ViLT	4.2M	0	9.6M	83.5	96.7	98.6	61.5	86.3	92.7	64.4	88.7	93.8	42.7	72.9	83.1
Ours $_{\text{w/o Tuning}}$	44.14M	1.33M	201M	74.8	94.8	98.2	57.7	85.6	92.3	65.8	88.8	93.6	46.3	75.0	84.0
Ours $_{\text{PT (1\%)}}$	44.14M	1.33M	201M	84.4	97.8	99.2	61.4	86.7	93.2	71.1	91.6	95.1	47.0	75.3	84.3
Ours $_{\text{FT (1\%)}}$	44.14M	1.33M	201M	78.4	95.7	97.8	60.2	85.1	90.6	61.0	85.7	91.0	43.6	70.9	80.5
Ours $_{\text{PT (10\%)}}$	44.14M	1.33M	201M	86.4	98.2	99.5	61.6	87.0	93.2	72.5	92.3	95.7	47.2	75.4	84.3
Ours $_{\text{FT (10\%)}}$	44.14M	1.33M	201M	84.9	97.4	98.3	60.9	85.5	92.1	67.9	89.4	92.9	45.6	73.4	82.6
Ours $_{\text{FT (100\%)}}$	44.14M	1.33M	201M	87.9	98.2	99.1	64.7	87.8	93.7	74.9	93.5	96.0	48.3	75.9	84.5

Table 3. Image-text retrieval performance under different tuning settings. PT means prompt-tuning, and FT means fine-tuning. The percentage of data used in tuning is noted. In addition, the pre-training dataset statistics of competitive methods are also listed.

Method	COCO Caption				Flickr30k			
	B@4	M	C	S	B@4	M	C	S
Unified VLP [92]	36.5	28.4	116.9	21.2	30.1	23.0	67.4	17.0
Ours $_{\text{w/o Tuning}}$	33.6	27.0	109.8	20.3	17.0	16.2	41.2	11.2
Ours $_{\text{PT (1\%)}}$	34.3	27.2	109.6	21.2	28.1	21.6	59.1	15.6
Ours $_{\text{FT (1\%)}}$	28.0	26.8	100.1	20.2	18.9	19.7	45.3	14.3
Ours $_{\text{PT (10\%)}}$	35.0	27.9	114.1	21.3	28.8	22.1	61.7	16.8
Ours $_{\text{FT (10\%)}}$	32.7	27.5	109.0	21.1	26.9	21.6	52.1	14.5
Ours $_{\text{FT (100\%)}}$	35.6	28.1	116.5	21.5	30.1	24.5	72.7	18.2

Table 4. Image caption performance under different tuning settings. B@4, M, C, S stand for BLEU-4, METEOR, CIDEr, and SPICE scores, respectively. Additionally, Unified VLP [92] conducted pre-training with around 3.0M image-text pairs.

Results. Tab. 2, Tab. 3, and Tab. 4 present the evaluation results of our models on four pre-training tasks, *i.e.*, image classification, video classification, image-text retrieval, and image caption. We compare our model with task-specific SOTA methods having the similar model size.

Results show that without any tuning, our pre-trained model reaches reasonable performance on these tasks. Although the performance is slightly worse than the SOTA methods. We speculate that the performance gap is due to the limited capacity of our model, which may have a negative impact on the representation ability. Note that our method shares a similar model size with other methods, but need to simultaneously process much more pre-training tasks from various datasets and modalities.

By conducting prompt tuning on each task with 1% downstream data, the performance is boosted to a level close to SOTA performance. It's worth noting that all parameters of other methods are specifically trained on the target tasks. While for our prompt tuning, only a small amount of parameters are tuned, and the encoder is still fixed and shared among different tasks, indicating that our method can handle different tasks with low marginal cost.

Method	Pre-training Data			MSVD				
	#Images	#Videos	#Text	B@4	M	R	C	S
ORG-TRL [88]	1.4M	650k	-	54.3	36.4	73.9	95.2	-
Ours $_{\text{w/o Tuning}}$	44.14M	1.33M	201M	20.3	25.8	52.1	45.7	6.5
Ours $_{\text{PT (1\%)}}$	44.14M	1.33M	201M	54.8	38.9	74.7	104.8	6.6
Ours $_{\text{FT (1\%)}}$	44.14M	1.33M	201M	47.3	35.8	66.2	80.1	6.2
Ours $_{\text{PT (10\%)}}$	44.14M	1.33M	201M	57.2	39.1	75.6	112.1	6.8
Ours $_{\text{FT (10\%)}}$	44.14M	1.33M	201M	56.7	38.7	70.0	88.2	6.7
Ours $_{\text{FT (100\%)}}$	44.14M	1.33M	201M	61.5	42.3	79.0	131.0	7.7

Table 5. Video caption (novel task) performance under different tuning settings. Note that this task did not appear in our pre-training. The only task related to video modality in our pre-training is video classification. In addition, the pre-training statistics of competitive methods are also listed.

Method	Pre-training Data			Text Retrieval MSVD			Video Retrieval MSVD		
	#Images	#Videos	#Text	R@1	R@5	R10	R@1	R@5	R@10
CLIP4clip [56]	400M	380k	400M	56.6	79.7	84.3	46.2	76.1	84.6
CLIP2video [21]	400M	0	400M	58.7	85.6	91.6	47.0	76.8	85.9
Ours $_{\text{w/o Tuning}}$	44.14M	1.33M	201M	42.7	69.1	79.6	34.6	64.5	75.4
Ours $_{\text{PT (1\%)}}$	44.14M	1.33M	201M	61.2	83.7	89.0	42.6	73.3	82.5
Ours $_{\text{FT (1\%)}}$	44.14M	1.33M	201M	49.6	75.8	83.7	37.5	68.2	79.3
Ours $_{\text{PT (10\%)}}$)	44.14M	1.33M	201M	61.3	84.8	90.9	43.1	74.2	83.4
Ours $_{\text{FT (10\%)}}$)	44.14M	1.33M	201M	59.1	81.9	87.4	41.7	71.6	81.3
Ours $_{\text{FT (100\%)}}$	44.14M	1.33M	201M	61.5	83.5	90.2	45.4	75.8	85.0

Table 6. Video-text retrieval (novel task) performance under different tuning settings. Note that this task did not appear in our pre-training. In addition, the pre-training statistics of competitive methods are also listed.

We further fine-tune the pre-trained model with 100% of the downstream data. With full-data fine-tuning, our model achieves performance on-par with or better than the SOTA methods on all these tasks, which proves our model has learned high-quality representations. We also compare the performance of prompt tuning and fine-tuning in the scenario of few-shot learning. On all of these tasks, prompt tuning shows a consistently better performance than fine-tuning with the same amount of data, which demonstrates its superiority under few-shot scenarios.

4.4. Generalization to Novel Tasks

Thanks to the generic perceptual modeling, our pre-trained model can generalize to novel tasks by converting the tasks into our unified task formulation. We evaluate zero-shot inference on tasks that did not appear in

Method	Pre-training Data			VQA v2 test-dev		
	#Images	#Videos	#Text	Yes/No	Numbers	Others
Unified VLP [92]	3.1M	0	-	87.2	52.1	60.3
Ours$_{\text{w/o Tuning}}$	44.14M	1.33M	201M	0.9	3.0	25.5
Ours$_{\text{PT (0.1\%)}}$	44.14M	1.33M	201M	63.0	31.8	49.6
Ours$_{\text{FT (0.1\%)}}$	44.14M	1.33M	201M	63.0	30.6	49.1
Ours$_{\text{PT (1\%)}}$	44.14M	1.33M	201M	70.8	41.3	57.7
Ours$_{\text{FT (1\%)}}$	44.14M	1.33M	201M	71.0	42.4	57.5
Ours$_{\text{FT (100\%)}}$	44.14M	1.33M	201M	84.8	47.4	61.8

Table 7. Visual question answering (novel task) performance under different tuning settings. Note that this task did not appear in our pre-training.

Method	GLUE					
	MNLI (Acc)	QNLI (Acc)	QQP (F1)	RTE (Acc)	SST-2 (Acc)	MRPC (F1)
PLM [3]$_{\text{w/o tuning}}$	49.4	50.7	46.6	53.8	70.6	44.2
BERT$_{\text{BASE}}$ [81]	84.6	92.7	71.2	66.4	93.5	88.9
RoBERTa$_{\text{BASE}}$[50]	87.6	92.8	91.9	78.7	94.8	90.2
Ours$_{\text{w/o Tuning}}$	49.6	51.0	53.6	55.6	70.6	76.1
Ours$_{\text{PT (1\%)}}$	60.1	76.0	70.2	56.3	80.9	80.3
Ours$_{\text{FT (1\%)}}$	47.3	60.6	68.9	49.1	69.7	72.3
Ours$_{\text{PT (10\%)}}$	68.5	83.2	77.0	58.2	83.4	83.2
Ours$_{\text{FT (10\%)}}$	60.5	71.5	71.4	50.5	79.1	80.6
Ours$_{\text{FT (100\%)}}$	81.7	89.9	87.1	64.3	90.2	86.6

Table 8. Natural language understanding (novel task) performance under different tuning settings. Note that this task did not appear in our pre-training.

pre-training, *i.e.*, video caption, video-text retrieval, visual question answering, and natural language understanding.

Video Caption and Video-Text Retrieval. Our pre-trained model is evaluated on MSVD [11] dataset. Specifically, for video caption, $\mathcal{X}_1$ consists of the concatenation of video and language sequences that have been predicted, and $\mathcal{X}_2$ denotes the set of all words in the vocabulary. For the video-text retrieval, the input sets $\mathcal{X}_1$ and $\mathcal{X}_2$ consist of possible video and text sequences, or vice versa.

Visual Question Answering. In visual question answering, the model is asked to answer a question w.r.t a reference image from a list of answer candidates. We evaluate our pre-trained model on VQA [26] dataset. $\mathcal{X}_1$ is a set of image-text sequence, where the text is the question tokens followed by a `<SPE>` token used to predict the answers. Each $x_2 \in \mathcal{X}_2$ is an answer sequence beginning with `<SPE>`. Inference is achieved by computing the similarity between output features of `<SPE>` in x_1 and x_2.

Natural Language Understanding. Six language-only tasks are chosen from GLUE benchmark [81] to evaluate the natural language understanding ability of our pre-trained model. These tasks are either single sentence classification or sentence-pair classification tasks. We follow [24] to construct the textual class labels for each dataset. Here, the input sequence $x_1 \in \mathcal{X}_1$ denotes the input single sentence or the sentence-pair, and the sequence $x_2 \in \mathcal{X}_2$ represents the class labels in each dataset.

Result. Tab. 5, Tab. 6 and Tab. 7 show the results on video caption and video-text retrieval and visual question answering, respectively. Our pre-trained model can obtain reasonable zero-shot performance on these novel tasks. Note that none of previous works can perform this type of zero-shot inference at all. From Tab. 7, we note that our model shows unsatisfactory zero-shot performance on "Yes/No" and "Number" subsets in VQA. We speculate that it may be due to the distribution difference between those answers and our pre-training corpora. We futher conduct prompt tuning on these tasks with only 1% data, which brings our model to a level close to the SOTA results. By further fine-tuning with 100% downstream data, our model can achieve results on par with or better than the SOTA methods.

On the GLUE benchmark, our model can achieve comparable performance with [3] in zero-shot evaluation. When fine-tuning the pre-trained model with 100% downstream data, our model performs slightly worse than BERT$_{\text{BASE}}$. Since our model has the same number of parameters as BERT$_{\text{BASE}}$, but need to process much more tasks from various datasets and modalities, we speculate that the performance drop is due to the limited capacity of the model.

5. Conclusion

In this paper, we propose a unified perception architecture that processes various modalities and tasks with a single model and shared parameters. With pre-training on unimodal and multi-modal tasks, our model shows the ability of zero-shot inference on novel tasks, and can reach the performance close to SOTA results by prompt tuning with only a small amount of downstream data. The performance can be further improved to be on par with or superior to SOTA results by full-data fine-tuning.

Limitations. Our method is currently only applicable when the target set is discrete, such as classification and retrieval. Whether our model can be extended to regression tasks is still questionable. Future work may explore the unified perception model of both discrete and continuous target sets.

Potential Negative Societal Impact. This work may share the common negative impacts of large-scale training, which may consume lots of electricity and result in increased carbon emissions. This method also learns from a large number of datasets that may contain data biases. Future work may seek for more efficient and unbiased training.

Acknowledge. The work is supported by the National Key R&D Program of China (2020AAA0105200), Beijing Academy of Artificial Intelligence, in part by the General Research Fund through the Research Grants Council of Hong Kong under Grants (Nos. 14204021, 14207319, 14203118, 14208619), in part by Research Impact Fund Grant No. R5001-18, in part by CUHK Strategic Fund.

References

[1] Hassan Akbari, Li Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021. 1, 2

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019. 3

[3] Anonymous. Are BERT families zero-shot learners? a study on their potential and limitations. In *Submitted to ICLR*, 2022. under review. 8

[4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*, 2021. 1, 2

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3

[6] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, July 2021. 1, 2

[7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 4, 6, 13, 15

[8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 2, 3, 5, 6

[9] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *CVPR*, 2021. 3, 6, 15

[10] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv preprint arXiv:2103.14899*, 2021. 1, 2

[11] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, pages 190–200, 2011. 6, 8

[12] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6, 15

[13] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 6

[14] Xinyang Chen, Sinan Wang, Bo Fu, Mingsheng Long, and Jianmin Wang. Catastrophic forgetting meets negative transfer: Batch spectral shrinkage for safe transfer learning. In *NeurIPS*, 2019. 2

[15] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020. 2, 3, 7

[16] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, pages 702–703, 2020. 15

[17] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 3, 6, 15

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3, 5, 6

[19] William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. 14

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 4, 13

[21] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 7

[22] Inc. Flickr. Flickr terms & conditions of use. https://www.flickr.com/help/terms. 15

[23] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 3

[24] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020. 8

[25] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *ACL*, 2021. 3, 6

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*, 2017. 6, 8

[27] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021. 1, 2

[28] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. *arXiv preprint arXiv:1909.00964*, 2019. 7

[29] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[30] Zhengbao Jiang, Frank F. Xu, J. Araki, and Graham Neubig. How can we know what language models know? *TACL*,

8:423–438, 2020. 3

[31] Sebastian Kalkowski, Christian Schulze, Andreas Dengel, and Damian Borth. Real-time analysis and visualization of the yfcc100m dataset. In *Proceedings of the 2015 workshop on community-organized multimodal mining: opportunities for novel solutions*, pages 25–30, 2015. 3, 6, 15

[32] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 6, 15

[33] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021. 1, 2, 7

[34] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[35] Ranjay Krishna. Visual genome terms & conditions of use. `https://visualgenome.org/about`. 15

[36] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 2, 3, 6, 15

[37] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019. 2, 3

[38] Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. *arXiv preprint arXiv:1605.07648*, 2017. 6

[39] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 3

[40] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019. 2, 3

[41] Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. Paq: 65 million probably-asked questions and what you can do with them. *arXiv preprint arXiv:2102.07033*, 2021. 6, 15

[42] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, 2020. 2

[43] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2, 3

[44] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021. 3

[45] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2021. 1, 2

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[47] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 2, 3, 5, 6

[48] Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021. 3, 6

[49] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. Gpt understands, too. *arXiv preprint arXiv:2103.10385*, 2021. 3

[50] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2, 3, 8

[51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *ICCV*, 2021. 1, 2

[52] Google LLC. Conceptual 12m terms & conditions of use. `https://github.com/google-research-datasets/conceptual-12m/blob/main/LICENSE`. 15

[53] Google LLC. Conceptual captions terms & conditions of use. `https://github.com/google-research-datasets/conceptual-captions/blob/master/LICENSE`. 15

[54] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019. 2

[55] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019. 3

[56] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 7

[57] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 42(2):502–508, 2019. 3, 6, 15

[58] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *NeurIPS*, 24:1143–1151, 2011. 3, 6, 15

[59] Gao Peng, Geng Shijie, Zhang Renrui, Ma Teli, Fang Rongyao, Zhang Yongfeng, Li Hongsheng, and Qiao Yu. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 3

[60] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase corre-

spondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 6

[61] Di Qi, Lin Su, Jianwei Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020. 2, 7

[62] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26, 2020. 3

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 3

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1, 2

[65] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28:91–99, 2015. 2

[66] Zhang Renrui, Fang Rongyao, Gao Peng, Zhang Wei, Li Kunchang, Dai Jifeng, Qiao Yu, and Li Hongsheng. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 3

[67] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *ACL*, 2016. 4, 13

[68] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 3, 6, 15

[69] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*, 2020. 3

[70] Lucas Smaira, João Carreira, Eric Noland, Ellen Clancy, Amy Wu, and Andrew Zisserman. A short note on the kinetics-700-2020 human action dataset. *arXiv preprint arXiv:2010.10864*, 2020. 15

[71] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013. 14

[72] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2, 3

[73] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 3

[74] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *CVPR*, pages 7464–7473, 2019. 3

[75] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2, 3

[76] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. Training data-efficient image transformers & distillation through attention. In *ICML*, volume 139, pages 10347–10357, July 2021. 1, 2

[77] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, pages 10347–10357. PMLR, 2021. 6, 15

[78] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. *arXiv preprint arXiv:2103.17239*, 2021. 1, 2

[79] Princeton University and Stanford University. Imagenet terms & conditions of use. `https://image-net.org/download`. 15

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1, 2

[81] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018. 6, 8

[82] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 1, 2

[83] Han Xu, Zhang Zhengyan, Ding Ning, Gu Yuxian, Liu Xiao, Huo Yuqi, Qiu Jiezhong, Zhang Liang, Han Wentao, Huang Minlie, et al. Pre-trained models: Past, present and future. *arXiv preprint arXiv:2106.07139*, 2021. 3

[84] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis E.H. Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *ICCV*, pages 558–567, October 2021. 1, 2

[85] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6023–6032, 2019. 15

[86] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 15

[87] Yanyi Zhang, Xinyu Li, Chunhui Liu, Bing Shuai, Yi Zhu, Biagio Brattoli, Hao Chen, Ivan Marsic, and Joseph Tighe. Vidtr: Video transformer without convolutions. In *ICCV*, pages 13577–13587, October 2021. 1, 2

[88] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *CVPR*, pages 13278–13288, 2020. 7

[89] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and

Yi Yang. Random erasing data augmentation. In *AAAI*, volume 34, pages 13001–13008, 2020. 15

[90] Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Qibin Hou, and Jiashi Feng. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*, 2021. 1, 2

[91] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*, 2021. 3

[92] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pretraining for image captioning and vqa. In *AAAI*, 2020. 2, 7, 8

[93] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, pages 19–27, 2015. 6, 15