A survey of cross-modal pre-training

Yitong Sun

Institute for Artificial Intelligence, Beihang University, Beijing, China, 100191

yt_sun@buaa.edu.cn

Abstract. Pre-training allows autonomous learning of samples' inherent semantics from unlabeled data in a large scale, and then obtains a task-independent general representation. Furthermore, the generalization of model encoding across different tasks can be improved. Inspired by early high-performance pre-trained models for a single modality of text or images, scholars began to explore image-text cross-modal pre-training methods, aiming to achieve the ability of understanding both visual information and textual semantics separately and achieve effective alignment between the two. This paper aims to summarize the development of image-text pre-training. Specifically, it summarizes the feature extraction and representation methods, pre-training tasks and extended downstream tasks used by various mainstream models. This paper also introduces and compares the mainstream models into two types of architectures: single-stream and dual-stream. Through the analysis, it can be concluded that the embedding representation determines the difficulty and performance of the subsequent fusion process, and the two kinds of architectures lead to different levels of fusion capabilities.

Keywords: cross-modality pre-training, image-text, single-stream, dual-stream.

1. Introduction

Pre-training process has attracted widespread attention due to its powerful generalization and efficient utilization of large-scale unlabeled data, and then the significant progress in the field of natural language processing (NLP) has further promoted the research of task-independent cross-modal pre-training. It is based on self-supervised learning and subsequently fine-tuned, making the model have a strong ability to adapt to different cross-modal downstream tasks. At present, the widely studied cross-modal pre-training model is the Vision-Language pre-training model (VLP), which makes use of the attention mechanism in the transformer architecture to learn the context relationship within single modality and the semantic correspondence between the two modalities by performing the designed pre-training tasks on large-scale unlabeled image-text pairs. Among them, the pre-processing and pre-training methods for text is relatively mature, and from various studies, it can be found that the research focus is on image feature representation and the fusion alignment of the two.

Image-text pairs need to be processed by embedding before entering into the network for learning. Due to continuous representation and higher semantics, visual data should be processed by appropriate design and requires deeper networks, while discrete text is generally processed by existing mature methods. In the early stage, convolutional neural networks (CNNs) for classification tasks are used to perform global feature extraction [1]. After that, many models use local features extracted by object detectors, such as ViLBERT [2], LXMERT [3], UNITER [4], VL-BERT [5] and so on, and then some models try to build a network based on Transformer to embed images like ViLT [6], MAE [7] and ALBEF [8]. With the improvement, their representation capabilities and cost have been improved. For learning, it builds association through designed tasks that utilize both text and images. Compared with training on target tasks with high-quality annotated datasets, it can learn closer semantic relationship between the sample pairs themselves, and the output will not be valid for a specific task.

This paper introduces the research advance of cross-modal pre-training from feature extraction and representation methods (Section2), pre-training tasks (Section3), mainstream model architectures (Section4) and downstream tasks (Section5), the details on mainstream models are shown in Table 1, so that other researchers can quickly learn mainstream methods and corresponding motivations, research priorities and possible future research directions in this field. This paper also explores corresponding problems by comparing different models and methods, which enables possible targeted improvement in follow-up work.

Table 1. The summary of mainstream image-text VLP models.

Model	**Architecture**	**Image**	**Pre-Training Tasks**	**Characteristics**
UNITER [4]	Single	RoI	MLM, MRM, ITM, WRA	conditional masking strategy
VL-BERT [5]	Single	RoI	MLM, MRC	be universal to different input formats
Pixel-BERT [9]	Single	Pixel(CNN)	MLM, ITM	random pixel sampling mechanism
InterBERT [10]	Single	RoI	MSM, MRM, ITM-hn	keep single-modal’s performance; all-attention
ViLT [6]	Single	ViT	MLM, ITM, WPA	lightweight; full-word masking
MAE [7]	Single	ViT	MRFR-hard	clever designs of masking and embedding
ViLBERT [2]	Dual	RoI	MLM, MRM, ITM	set extra text transformers before fusion
LXMERT [3]	Dual	RoI	MLM, MRM, ITM, VQA	set different depth layers for both modalities
CLIP [11]	Dual	CNN/ViT	ITM	achieve zero-shot migration on retrieval
ALBEF [8]	Dual	ViT	MLM, ITM, ITM-hn	the MoD improve learning on noisy data; use the contrast loss before interaction

2. Feature extraction and representation methods

Cross-modal pre-training requires an extraction representation method that is effective for the two modalities and easy for later-stage integration as a joint input to Transformer.

2.1. Text feature extraction and representation

Text processing is based on the same approach in NLP. It applies the WordPiece word segmentation method [12] based on probability and a sentence is then represented as a sequence of subwords $\mathrm{w} = \{\mathrm{w}_1, \dots, \mathrm{w}_\mathrm{T}\}$ with a special label $[\mathrm{CLS}]$. Considering the disorder mechanism of self-attention layer, which is agnostic to the absolute position index and thus, the embedding that complements target position can break this limitation. Many methods project word w_i and its absolute position i in the sentence into a corresponding vector, and then send it into the norm layer:

$$h_i = LayerNorm(\widehat{w_i} + \widehat{u_i}) \tag{1}$$

Some methods also introduce paragraph embedding, which effectively determines the source of elements. In addition to providing spatial information, position embedding is also very necessary for masked-target prediction tasks and some image question answering tasks.

2.2. Image feature extraction and representation

Compared with discrete text, images are continuous with higher semantics, requiring more complex methods to extract features. Initially, models utilize CNNs of classification tasks to globally obtain a discretized grid representation [1]. With the introduction of “attention”, many methods use locally multi-target region of interest (RoI) features obtained by object detection models. The most commonly used detector is Faster R-CNN [13], which has bottom-up attention. For images, location information is also considered, and the final representation is represented in combination with RoI features and boundary coordinates as v_j, where layer normalization is applied to projected features before summing to balance the effects of different types of features:

$$\widehat{f_j} = \mathrm{LayerNorm}(W_F f_j + b_F)$$

$$\widehat{p_j} = \mathrm{LayerNorm}(W_P p_j + b_P)$$

$$v_j = (\widehat{f_j} + \widehat{p_j})/2 \tag{2}$$

Region-based feature extraction leads to some loss of visual semantics, such as specific behaviors of the target or the shape of the boundary, and the representation ability is limited to distinguishing given categories, where should have a wider range of semantics. To overcome this bias, Pixel-BERT [9] enhances the tightness of learning by pixel-level embedding and reduces the cost of bounding box labeling. Although the implementation also uses CNNs, it is different from the CNNs for classification tasks which directly extract global features, while the computation cost is similar.

Inspired by Transformer in NLP, pre-training models based on Vision Transformer (ViT) [14] directly transform image patches into vectors as visual words. Similar to text, the position information is usually added. The advantage of this method is that the image results can be sent directly with text representation which has a similar discrete structure to a network for subsequent processing.

In general, most cross-modal pre-training models take $I = \{[CLS], \widehat{w_1}, \dots, \widehat{w_T}, [SEP], \widehat{v_1}, \dots, \widehat{v_k}\}$ as the input to the Transformer for subsequent fusion.

3. Pre-training tasks

After obtaining inputs, the model also needs to learn semantics and alignment through proper pre -training tasks, and the task type and difficulty determine the generalization ability of the final model.

3.1. Cross-modal masked model prediction task

For language, Masked Language Model (MLM) uses a similar setup with BERT [15]. The difference is that the model can predict the target through unmasked tokens combining image information, which not only effectively learns text representations, but also helps to establish an association from visual to text modality. It should be noted that the embedding operation is on text input with masks. Assuming that the image sequence is $v = \{v_1, \dots, v_k\}$, and the text sequence is $w = \{w_1, \dots, w_T\}$, the task goal is to learn by minimizing the negative log-likelihood function based on $w_{\overline{m}}$ and all image regions v:

$$L_{MLM}(\theta) = -E_{(w,v)\backslash simD} \log P_\theta(w_m | w_{\overline{m}}, v). \tag{3}$$

Masked Region Model (MRM) is similar to MLM. However, it is more difficult to predict masked region only based on remaining known image areas, which may cause ambiguity, so it is more dependent on text. This task can establish an association from text to visual modality. Unlike discretized representation of text, visual features are high-dimensional and continuous, so learning cannot be supervised by log-likelihood in MLM tasks, and another loss function is proposed here to guide optimization, and the function f is different for the following three subtasks:

$$\mathrm{L}_{\mathrm{MRM}}(\theta) = \mathrm{E}_{(\mathrm{w},\mathrm{v})\backslash\mathrm{simD}}\mathrm{f}_{\theta}(\mathrm{v}_{\mathrm{m}}|\mathrm{v}_{\overline{\mathrm{m}}},\mathrm{w}). \tag{4}$$

Masked Region Feature Regression (MRFR) is a pixel-level regression task. It applies a fully connected (FC) layer to convert transformer's output into a vector $\mathrm{h}_{\theta}(\mathrm{v}_{\mathrm{m}})^{(\mathrm{i})}$ with same dimension of the RoI pooling feature $\mathrm{r}(\mathrm{v}_{\mathrm{m}}^{(\mathrm{i})})$, and then trains with L_2 regression function:

$$f_{\theta}(\mathrm{v}_{\mathrm{m}}|\mathrm{v}_{\overline{\mathrm{m}}},\mathrm{w}) = \sum_{\mathrm{i}=1}^{\mathrm{M}} \left\|\mathrm{h}_{\theta}\left(\mathrm{v}_{\mathrm{m}}^{(\mathrm{i})}\right) - \mathrm{r}\left(\mathrm{v}_{\mathrm{m}}^{(\mathrm{i})}\right)\right\|_2^2. \tag{5}$$

Masked Region Classification (MRC) predicts the target's class as a higher-dimensional task. It obtains the normalized distribution scores $\mathrm{g}_{\theta}(\mathrm{v}_{\mathrm{m}}^{(\mathrm{i})}) \in \mathrm{R}^{\mathrm{K}}$ through FC and softmax function. Since there is no ground truth, the output of Faster R-CNN with the highest confidence score is used here as a hard label in the form of a one-hot vector $\mathrm{c}(\mathrm{v}_{\mathrm{m}}^{(\mathrm{i})})$, then the network is trained by cross-entropy loss:

$$f_{\theta}(v_m|v_{\overline{m}},w) = \sum_{i=1}^{M} \mathrm{CE}(c(v_m^{(i)}) - g_{\theta}(v_m^{(i)})). \tag{6}$$

However, when there are no truly ground-truth labels available, it is inaccurate to select even the most likely one. To solve this problem, the improved task with KL divergence (MRC-kl) uses the detector's probability distribution of the regional category $\hat{\mathrm{c}}(\mathrm{v}_{\mathrm{m}}^{(\mathrm{i})})$ as a soft label:

$$f_{\theta}(v_m|v_{\overline{m}},w) = \sum_{i=1}^{M} D_{KL}(\hat{c}(v_m^{(i)})) - g_{\theta}(v_m^{(i)})). \tag{7}$$

3.2. Image-text matching

Unlike the masked-target prediction tasks, which regard the other modality as extra clues, Image-text Matching (ITM) directly establishes the overall alignment of the two modality samples. It creates negative samples by randomly replacing the image or text in the original pairs, and then trains a classifier to predict whether the sample pairs are homologous as a binary classification problem. It uses $[\mathrm{CLS}]$ to calculate the matching score $\mathrm{s}_{\theta}(\mathrm{w},\mathrm{v})$ through the FC layer and sigmoid function. This task converges fast, so it has been widely used. Besides, since $[\mathrm{CLS}]$ always represents a cross-modal union in most downstream tasks, the way of score calculation can alleviate the mismatch between pre-training and downstream tasks. During the training process, binary cross-entropy loss is used with an equal amount of positive and negative samples to avoid bias caused by sample imbalance:

$$L_{ILM}(\theta) = -\mathrm{E}_{(\mathrm{w},\mathrm{v})\backslash\mathrm{simD}}\left[\mathrm{y}\log\mathrm{s}_{\theta}(\mathrm{w},\mathrm{v}) + (1-\mathrm{y})\log\left(1 - \mathrm{s}_{\theta}(\mathrm{w},\mathrm{v})\right)\right]. \tag{8}$$

Word-Region Alignment (WRA) is first proposed by UNITER [4]. ViT-based models also design a similar word-patch matching (WPA). It is distinguished from ITM which aligns globally, encouraging fine-grained local alignment. WRA optimizes the relationship by learning the transport matrix T with the goal of minimizing the cost of transmitting the distribution of one modality to another, which makes the alignment robust. Since it is difficult to directly solve the minimization of T, the IPOT algorithm [16] is used for estimation. Then the OT distance is used as the loss to update the parameter:

$$L_{WRA}(\theta) = \mathrm{D}_{\mathrm{ot}}(\mu,\mathrm{v}) = \min_{\mathrm{T}\in\Pi(\mathrm{a},\mathrm{b})} \sum_{\mathrm{i}=1}^{\mathrm{T}}\sum_{\mathrm{j}=1}^{\mathrm{K}} T_{ij}\cdot c\left(w_i, v_j\right), \tag{9}$$

where $\Pi(\mathrm{a},\mathrm{b}) = \{\mathrm{T} \in \mathrm{R}_{+}^{\mathrm{T}\times\mathrm{K}} | T1_m = a, T^{\top}1_n = b\}$, c is the cosine similarity to assess the cost.

Besides, InterBERT [10] innovatively proposes two harder versions of MLM and ITM: Masked Group Prediction (MGM) and ITM on hard negative samples (ITM-hn). MGM consists of original MRM and Masked Segment Prediction Model (MSM), which masks not tokens but continuous fragments of text. It can encourage the model to learn semantic interaction. The core of ITM-hn is the construction of negative samples. It is not based on random replacing that is easy to distinguish for models, but filters out sentences that overlap more words with different semantics. This task requires the model to establish stronger semantic connections, so as to improve the matching ability.

3.3. Image-text contrast learning

This task is similar to matching tasks with broader options, requiring models to predict truly matched N positive sample pairs from a larger $\mathrm{N} \times \mathrm{N}$ sample combination. The two modalities are represented by label $[\mathrm{CLS_V}]$ and label $[\mathrm{CLS_W}]$, and then the similarity is calculated by the dot product operation. The model uses the cross-entropy loss of the two normalized similarities to guide the learning process:

$$L_{VLC} = \frac{1}{2}\mathrm{E}_{(\mathrm{I},\mathrm{T})\sim\mathrm{D}}\left[\mathrm{CE}\left(\mathrm{y}^{\mathrm{v2w}}, \mathrm{p}^{\mathrm{v2w}}(\mathrm{I})\right) + \mathrm{CE}\left(\mathrm{y}^{\mathrm{w2v}}, \mathrm{p}^{\mathrm{w2v}}(\mathrm{T})\right)\right], \tag{10}$$

$\mathrm{p}(\cdot)$ is similarity, $\mathrm{y}^{\mathrm{v2w}}$ and y^{w2v} are labels based on image-to-text and text-to-image retrieval.

4. Model architecture

4.1. Single-stream models

The single-stream models refer to a unified transformer architecture, which can learn the context information of one modality and the connection between the two with the attention mechanism. Single -stream models are more efficient because the sharing settings of parameters for both modalities.

UNITER [4] is one early study of the general cross-modal models, which breaks the barrier of task-specific. Unlike previous models like ViLBERT [2] and VL-BERT [5] applying a joint random mask of the two modalities, UNITER proposes a conditional masking strategy, which limits only one modality data to be randomly masked at a time, and then predicts under the complete observation of another modality to avoid weak connections caused by the simultaneous absence of important regions.

Given VQA's input is <Question, Answer, Image>, while VC's input is <Caption, Image>, VL-BERT [5] solves the input format problem of different cross-modal tasks by designing representation, a sum of four embeddings with special settings for images. $[IMG]$ is used as a token part, and paragraph embedding sets three flags to distinguish elements. For visual features, non-RoI regions are also considered with the use of global features of that area. Finally, position embedding is set the same since any permutation of the input sequence should achieve the same result. VL-BERT is also pre-trained on text corpus to improve the model's generalization on long and difficult sentences.

Pixel-BERT [9] achieves the highest fine-grained alignment learning, establishing a thorough connection between images and text pixel by pixel to learn richer visual semantics. Inspired by Dropout [17], Pixel-BERT proposes a random pixel sampling mechanism, which only randomly passes some of the pixel samples to the cross-modal layer during training to make up for the difficulty of predicting pixel-level features as well as reducing operation cost. More importantly, it allows the model to learn semantics through incomplete input and thus alleviates the overfitting problem.

InterBERT [10] ensures stability on both cross-modal and single-modal tasks. The two high-level representations generated by independent modules can further form a joint one by feed-forward networks. Given ideal attention should focus on the overall situation, it proposes all-attention instead of co-attention which only notices the other. The same embedding space facilitates attention on broader cross-modal context meanwhile, realizing the combination of self-attention and co-attention.

ViLT [6] achieves lightweight image embedding based on ViT instead of CNNs, using Transformer to encode both rather than different modules to process them separately. Compared with Pixel-BERT, ViLT breaks the limitation of regionalized semantic representation alike while greatly reducing computational cost. Besides, it proposes a full-word masking setting just like the conditional strategy, reducing the dependence of unmasked subwords in order to make full use of visual information. On this basis, MAE [7] requires only one percent of the unlabeled data to achieve the same effect. The core idea is to randomly mask most image regions rather than the 15% regular setting. Similar to the full-word masking idea for text of ViLT, large-scale masking of image patches can also enhance the use of text information. In particular, MAE only encodes the visible area, combines visual embeddings with the text features of the masked area together, and finally forms a long vector sequence according to absolute position to represent the entire image, which greatly reduces the amount of computation.

Unlike models that strive to improve versatility, Uni-Perceiver-MoE [18] aims to solve the performance degradation on specific tasks due to universal learning. It introduces conditional multi -expert models to adjust the shared parameters according to inputs, and solves the inconsistent optimization direction of different tasks and modalities. To measure interference, the change of task i's loss before and after task j guiding the optimization of a shared parameter θ is calculated:

$$\triangle_j L_i(x_i) \doteq E_{x_j}(L_i(x_i;\theta) - L_i(x_i;\theta - \lambda \frac{\nabla_\theta L_j(x_j)}{||\nabla_\theta L_j(x_j)||})) \approx \lambda E_{x_j}\left(\frac{\nabla_\theta L_j(x_j)}{|\nabla_\theta L_j(x_j)|}^T \nabla_\theta L_i(x_i)\right). \quad (11)$$

Further, the interference of task j on task i can be quantified as:

$$I_{i,j} = E_{x_i}\left(\frac{\triangle_j L_i(x_i)}{\triangle_i L_i(x_i)}\right). \quad (12)$$

4.2. Dual-stream models

Dual-stream models learn intra-modal feature representation through two independent transformers without shared parameters, and then use a co-attention mechanism to achieve cross-modal interaction.

For ViLBERT [2], text is processed additionally before fusion, in view of visual features are already extracted high-level features, while words need to be aggregated to express semantics, which is in line with our instincts. By stacking multiple cross-modal and single-modal layers, learning within and between modalities are continuously optimized. The main difference between LXMERT [3] and ViLBERT is that LXMERT has set different depth layers for both modalities for intra-modal learning.

As one of the most popular models, CLIP [11] use a large amount of raw data to train from scratch. Since text has more semantic possibilities, learning from the text associated with images is a good idea. Experiments show that CLIP's zero-shot migration effect on retrieval can even exceeds traditional supervised models, while this excellent performance does not guarantee the same effect on other tasks.

As noises in datasets usually cause overfitting which reduces generality, ALBEF [8] proposes momentum distillation (MoD) to generate pseudo-targets with similar semantics as additional supervision, so that the target options are more flexible. For alignment, ALBEF also applies the contrast loss in CLIP before interaction. On the one hand, it can improve the understanding of single modality. On the other hand, it can align ahead of time, which is convenient for further integration.

5. Downstream tasks

Pre-training models need to be migrated to new downstream tasks with little fine-tuning to preserve the pre-trained generalization. Some models directly select downstream tasks as pre-training tasks to learn representation capabilities that are closer to practical applications. Visual question answering (VQA), which requires models to select the correct answer from the answer bank, is a basic classification task. Compared with VQA, visual common sense reasoning (VCR) [19] not only selects the answer, but also needs to indicate the reason for choosing the answer through semantic reasoning. Natural language visual reasoning (NLVR) [20] is a broader task of VCR, which processes long difficult text sequences covering various linguistic phenomena, which is much closer to the real world.

In addition to multiple-choice question answering, there are also tasks of retrievial. Image-text retrieval consists of two subtasks: image-retrieval-text (TR) and text-retrieval-image (IR). CLIP is the best for the tasks because the consistent objective with contrast learning. For more difficult text tasks than matching, the model needs to generate text sentences with appropriate semantics and syntax for a given visual sample. Adding explanatory captions requires not only a wealth of linguistic knowledge, but also an accurate and consistent understanding of the scenes, entities, and interactions that appear in visual inputs. The common difficulty of the above tasks is that the similarity and alignment relationship between different modal feature spaces of complex samples cannot be directly measured.

6. Conclusion

This paper mainly summarizes image-text cross-modal pre-training and some meaningful ideas. It has been found that the current field focuses on image embedding and fusion alignment methods. For image features, the current performance is better with fine-grained pixel-level embedding, or directly using linear projection without CNNs, where the semantic representation is stronger and discretization is easy to align with discrete text features. For fusion, dual-stream interaction is generally shallow with insufficient fusion, but the exchange of space for time is suitable for timeliness retrieval tasks based on similarity. While single-stream models learn in the co-embedding space, building deeper semantic relationships. Models use self-attention to learn single-modal semantics, co-attention for cross-modal alignment, and all-attention to further breaks the limitation of only focusing on the other modality. Some models usually enhance intra-modal learning through self-attention layers after interaction.

In the future, image-text pre-trainings will be extended to video-audio tasks with a wider range of scenarios. For more diverse needs, how to design effective pre-training tasks is worth thinking about. It is also challenging to grasp essential semantic information from the added elements such as special effects. Besides, current methods have realized fine-grained token-region alignment, while achieving effective alignment between words with semantics and objects in images still needs further research.

References

[1] Li L H, Yatskar M, Yin D, et al. Visualbert: A simple and performant baseline for vision and language [J]. arXiv preprint arXiv:1908. 03557, 2019.

[2] Lu J, Batra D, Parikh D, et al. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks [J]. Advances in neural information processing systems, 32, 2019.

[3] Tan H, Bansal M. Lxmert: Learning cross-modality encoder representations from transformers [J]. arXiv preprint arXiv:1908. 07490, 2019.

[4] Chen Y C, Li L, Yu L, et al. Uniter: Universal image-text representation learning [C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX. Cham: Springer International Publishing: 104-120, 2020.

[5] Su W, Zhu X, Cao Y, et al. Vl-bert: Pre-training of generic visual-linguistic representations [J]. arXiv preprint arXiv:1908.08530, 2019.

[6] Kim W, Son B, Kim I. Vilt: Vision-and-language transformer without convolution or region supervision [C]//International Conference on Machine Learning. PMLR: 5583-5594, 2021.

[7] He K, Chen X, Xie S, et al. Masked autoencoders are scalable vision learners [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 16000-16009, 2022.

[8] Li J, Selvaraju R, Gotmare A, et al. Align before fuse: Vision and language representation learning with momentum distillation [J]. Advances in neural information processing systems, 34: 9694-9705, 2021.

[9] Huang Z, Zeng Z, Liu B, et al. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers [J]. arXiv preprint arXiv:2004.00849, 2020.

[10] Lin J, Yang A, Zhang Y, et al. Interbert: Vision-and-language interaction for multi-modal pretraining [J]. arXiv preprint arXiv:2003.13198, 2020.

[11] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [C]//International conference on machine learning. PMLR: 8748-8763, 2021.

[12] Wu Y, Schuster M, Chen Z, et al. Google's neural machine translation system: Bridging the gap between human and machine translation [J]. arXiv preprint arXiv:1609.0814 4, 2016.

[13] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks [J]. Advances in neural information processing systems, 28, 2015.

[14] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[15] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.

[16] Xie Y, Wang X, Wang R, et al. A fast proximal point method for computing exact wasserstein distance [C]//Uncertainty in artificial intelligence. PMLR: 433-453, 2020.

[17] Srivastava N, Hinton G, Krizhevsky A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The journal of machine learning research, 15(1): 1929-1958, 2014.

[18] Zhu J, Zhu X, Wang W, et al. Uni-perceiver-moe: Learning sparse generalist models with conditional moes [J]. arXiv preprint arXiv:2206.04674, 2022.

[19] Zellers R, Bisk Y, Farhadi A, et al. From recognition to cognition: Visual commonsense reasoning [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 6720-6731, 2019.

[20] Suhr A, Zhou S, Zhang A, et al. A corpus for reasoning about natural language grounded in photographs [J]. arXiv preprint arXiv:1811.00491, 2018.