MULTI-MODAL PRE-TRAINING: A NEW PARADIGM FOR MULTI-MODALITY UNDERSTANDING

Bei Liu
Microsoft Research Asia
Beijing, China
bei.liu@microsoft.com

Jianlong Fu
Microsoft Research Asia
Beijing, China
jianf@microsoft.com

Abstract

Pre-training has been an emerging topic that provides a way to learn powerful representation for downstream tasks in many fields (e.g., natural language processing, computing vision). In the last few years, we have witnessed many research works on multi-modal pre-training, especially in the visionlanguage domain. Pre-training models achieve state-of-the-art performances in many downstream tasks. They outperform traditional models by a large margin with a very simple design and demonstrate the superiority of pre-training on a large scale of data. In this article, we will guide you to see the power of multi-modal pre-training and introduce our exploration in this direction.

I. WHAT IS MULTI-MODAL PRE-TRAINING?

A modality in the context of human-computer interaction is defined as the classification of a single independent channel of sensory input/output between the human and the computer [6]. Multi-modality focuses on studying the integration of multiple communicative modalities, such as vision, language, audition, depth, etc. Technologies in a single modality have achieved a high level and machines can even outperform human in some tasks (e.g., image classification, language translation, speech recognition). However, the real world is multi-modality and we humans interact with the world in multiple senses. To accelerate the progress of human-like AI, multi-modality learning becomes much more critical and attracts more attention from industrial scenarios in recent years. Using vision and language modalities for example, there are many tasks proposed, like automatically generate one or several sentence in natural language given visual signals (i.e., image/video captioning [4], visual storytelling [7], imagebased poem generation [9]) or vice versa, automatically generate visual signals guided by language [1], vision-language alignment (i.e., image/video-text retrieval, image language grounding, video temporal localization), visual question and answering [10], etc.

Pre-training has shown great potential in many domains, especially with the development of deep learning. Compared with traditional training with hand-crafted features, deep learning requires much more amount of data to learn the hidden features. While many tasks as mentioned above have limited supervised data in certain scenarios, a model pre-trained with a large scale of data provides a much better initial representation for faster convergence and higher performance. Forexample, many computer vision models use a backbone pretrained with ImageNet [2], and many recent natural language models utilize BERT [3] for initialization. Multi-modality pre-training is a new paradigm that provides better crossmodal representation for downstream multi-modality tasks by learning from a large scale of multi-modal data with welldesigned pre-training tasks. Figure 1 shows a general pipeline of a multi-modal pre-training. Each modality is fed to its own encoder for representation learning, and a multi-modal representation learning model is designed for joint learning of multiple modalities with designed general pre-training tasks. Pre-training model is used as initialization and then fine-tuned on different downstream tasks to achieve good performance.

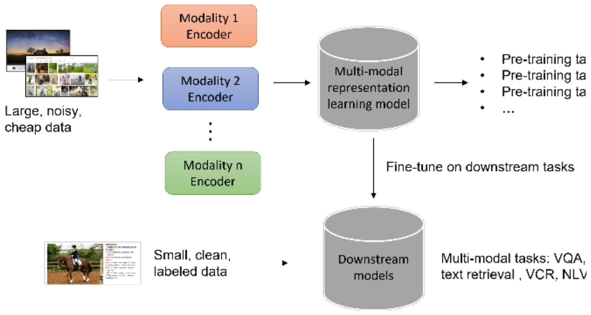

Large, noisy, cheap data
Modality 1 Encoder
Modality 2 Encoder
Modality n Encoder
Multi-modal representation learning model
Pre-training ta
Pre-training ta
Pre-training ta
...
Fine-tune on downstream tasks
Small, clean, labeled data
Downstream models
Multi-modal tasks: VQA, text retrieval , VCR, NLV

Fig. 1. Multi-modal pre-training pipeline for multi-modality tasks, using vision-language as an example.

II. WHY WE NEED MULTI-MODAL PRE-TRAINING?

In the multi-modality domain, the amount of high-quality data is limited and the annotation of multi-modal data is very costly. For example, MSCOCO [8], which is a widely-used image-description dataset for image-text retrieval and image captioning, costs 108,000 dollars to label 5 sentences for all 120,000 images. On the other side, there are a large amount of weakly paired data available on the Internet, such as imagetext pairs, and video-transcript-audio data. The power of pretraining is intuitive. A machine that has seen a lot of data in advance and performs well at some pre-defined multi-modal tasks can better perform multi-modal tasks compared to the one trained from scratch.

III. HOW MULTI–MODAL PRE-TRAINING WORKS?

The key to multi-modality learning is the alignment between different modalities. This is challenging due to many aspects. First, the representation of each modality is different which makes the alignment difficult to learn. For example, the representation of images (i.e., RGB) is real-valued and dense while language (i.e., word token) is represented in discrete and sparse form. Second, different modalities are not exactly matched which makes the learning of alignment even harder. For example, a sentence for an image can only indicate part of the information in the image and we cannot picture a whole video with only its audio. Third, it is difficult to directly evaluate the goodness of the alignment learned by a pre-trained model.

In this section, we will introduce three research works that we have done to tackle the above challenges.

A. End-to-end image-language pre-training

In early works of image-language pre-training, image representation is usually fixed by using region-based image features following previous works on image-language tasks (e.g., image captioning, image question and answering). Having an image, we first extract regions of objects in the image and use the visual features of these regions as input for multi-modal learning. However, there are three drawbacks to using regionbased features. First, region-based features only focus on the foreground objects in the images while neglecting the context in the background of images. Context is not that critical for object-centered tasks (i.e., image classification and object detection) while for language-related tasks, context is much more important. Second, the visual representation of images is limited to the pre-defined categories while the semantics in the language domain is much larger. Third, as the object detector used to extract visual features is too heavy to be jointly optimized with multi-modality learning, the extracted region features cannot be optimized for target cross-modal tasks.

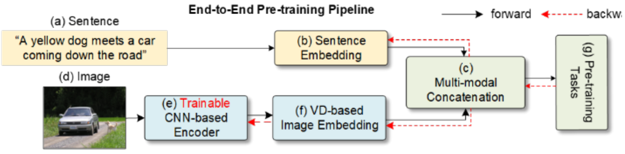

End-to-End Pre-training Pipeline
forward
backw
(a) Sentence
"A yellow dog meets a car coming down the road"
(b) Sentence Embedding
(c) Multi-modal Concatenation
(g) Pre-training Tasks
(d) Image
(e) Trainable CNN-based Encoder
(f) VD-based Image Embedding

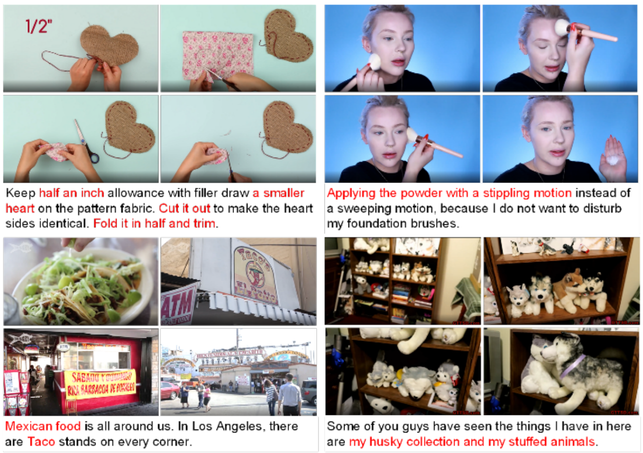
Fig. 2. SOHO: the first end-to-end vision-language pre-training framework[5].

To overcome the above shortages of pre-extracted regional features, we propose the first end-to-end image-language pretraining model to See Out of tHe bOx (SOHO) [5] for better cross-modality learning in CVPR 2021. As illustrated in Figure 2, for an image-text pair, we use Transformerbased text embedding as a language encoder and a trainable CNN-based visual encoder to extract visual representation. In this design, we do not need an object detection model and the information we can learn from images is not limited to pre-defined categories. The visual backbone is optimized in an end-to-end fashion and visual features can be updated in alignment with language.

Different from language modality where each word has its own particular meaning, pixels in images often share the same semantics. To better align image and text in the semantic level, we group pixels at the feature level with similar features into one item to indicate a consistent semantic. This is achieved by applying a visual dictionary (VD)-based image embedding to the image encoder outputs. Text embedding and VD-based embedding are then concatenated for three designed pre-training tasks: image-text matching, masked language modeling, and masked vision modeling. Through this work, we find that end-to-end learning can result in a better representation of multimodality.

B. High-resolution video-language pre-training

Data is one of the main factors in deep learning-based models. In the video-language domain, the datasets are limited in either scale or scope. Early datasets that use videos and annotated descriptions are limited in scale due to the heavy cost of annotation. The most used large-scale video-language dataset for pre-training (i.e., HowTo100M) consists of only instructional videos with their transcripts. Thus in a videolanguage pre-training work accepted by CVPR 2022 [11], we collect a video-language dataset (HD-VILA-100M) to overcome both limitations.



Keep half an inch allowance with filler draw a smaller heart on the pattern fabric. Cut it out to make the heart sides identical. Fold it in half and trim.

Applying the powder with a stippling motion instead of a sweeping motion, because I do not want to disturb my foundation brushes.

Mexican food is all around us. In Los Angeles, there are Taco stands on every corner.

Some of you guys have seen the things I have in here are my husky collection and my stuffed animals.

Fig. 3. One example of video-language pair in HD-VILA-100M [11].

We use transcripts along with videos from YouTube as the source of our dataset. Figure 3 shows an example in HDVILA- 100M dataset. HD-VILA-100M has three key properties. First, it is one of the largest video-language datasets. It includes 100 million video clips and transcript pairs from 3.3 million videos. It covers 371.5K hours in total which is 2.8 times than HowTo100M dataset in duration. The average length of each sentence is 13.4 which is about 8 times longer than HowTo100M. This ensures the richness of semantic in language. Second, all the videos are in high resolution with 720p. The quality of videos is much higher than most video datasets that are 240p or 360p. Third, the dataset is diverse and balanced in consideration of topics as shown in Figure 4. It covers 15 popular categories on YouTube and the number of video clips in each category is balanced.

To efficiently utilize the high-resolution videos in pretraining, we propose to form a hybrid image sequence which consists of one high-resolution frame and several surrounding low-resolution frames from a video clip. The hybrid image sequence is then fed into a novel hybrid video encoder that learns spatiotemporal information with a hybrid Transformer. Since the alignment between videos and transcripts is not as high as video-description pairs, we adopt contrastive learning to ensure paired data are close to each other while unpaired ones are far from each other.

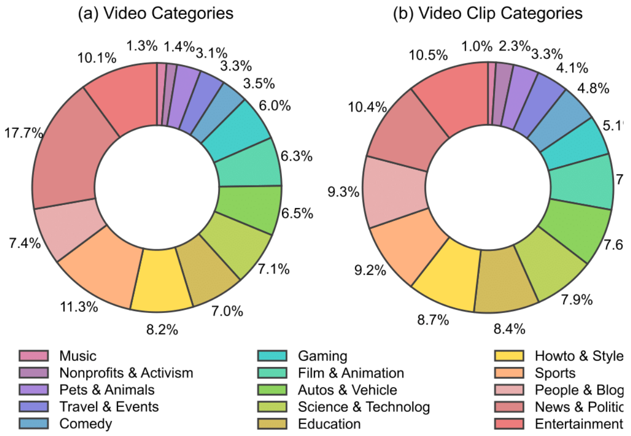


Fig. 4. Distribution of categories in HD-VILA-100M.

C. Probing inter-modality in vision-language pre-training

It is essential to learn the relation between different modalities in multi-modality tasks. In the vision-language domain, learning the inter-modal alignment between visual information and language semantics is very important. For language, the structured text with grammar makes it easy to learn the intrarelation of words. While for images, image features with CNN-based backbones (e.g., grid or regional feature) lacks the global relationship learning between different semantics. The inconsistent representation of visual and language modalities adds the burden of intra-modal learning on the visual side and inter-modal learning of both modalities encapsulated in the multi-modal module. This makes the learning of alignment even hard.

In our paper published in NeurIPS 2021 [12], we propose the first fully Transformer-based image-language pre-training model as shown in Figure 5. By adopting self-attention for visual feature learning, the spatial inductive bias is not introduced and we can learn long-range global relations of visual semantics before joint learning. This ensures the multi-modal Transformer is more specialized for cross-modal joint learning. To further measure the fusion quality of inter-modality learning, we propose the Inter-Modality Flow (IMF) metric to compute the information flow between two modalities.

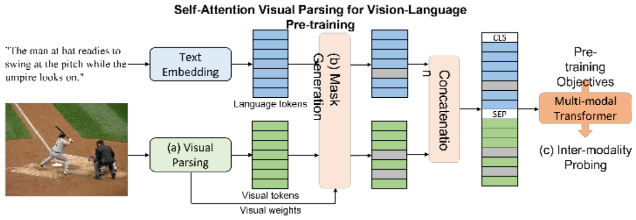

Self-Attention Visual Parsing for Vision-Language Pre-training
"The man at bat readies to swing at the pitch while the umpire looks on."
Text Embedding
Language tokens
(a) Visual Parsing
Visual tokens
Visual weights
(b) Mask Generation
Concatenation
CLS
SEP
Pre-training Objectives
Multi-modal Transformer
(c) Inter-modality Probing

Fig. 5. The first fully Transformer-based image-language pre-training model[12].

IV. WHAT IS THE NEXT?

We can see many research works in multi-modality pretraining in the past few years, especially in the vision-language domain. However, how to efficiently use pre-trained models in industry scenarios still faces many challenges. First of all, pre-training models are usually too heavy while real-time computing is often required in real applications. Secondly, how to bridge the domain gap between pre-training data and the real data in the wild (e.g., health care, navigation, digital human) remains a problem. Moreover, as we have claimed above, the real world is about much more modalities than vision and language. How to effectively learn the joint representation and alignment between more than two modalities is worth studying. For example, in the domain of embodied AI, more modalities (such as depth, action, segmentation, etc.) are involved. How to learn the alignment between different modalities when it is more complex will be a problem. No matter how difficult the path ahead is, we still believe multimodality pre-training is the right way we need to walk towards the real AI.

REFERENCES

[1] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin,Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In ACM MM, pages 2236–2244, 2019.
[2] 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] 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.
[4] Yupan Huang, Hongwei Xue, Bei Liu, and Yutong Lu. Unifying multimodal transformer for bi-directional image and text generation. In ACM MM, pages 1138–1147, 2021.
[5] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In CVPR, pages 12976–12985, 2021.
[6] Fakhreddine Karray, Milad Alemzadeh, Jamil Abou Saleh, and Mo Nours Arab. Human-computer interaction: Overview on state of the art. International Journal on Smart Sensing and Intelligent Systems, 1(1), 2017.
[7] Nanxing Li, Bei Liu, Zhizhong Han, Yu-Shen Liu, and Jianlong Fu. Emotion reinforced visual storytelling. In ICMR, pages 297–305, 2019.
[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll ́ar, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, pages 740–755. Springer, 2014.
[9] Bei Liu, Jianlong Fu, Makoto P Kato, and Masatoshi Yoshikawa. Beyond narrative description: Generating poetry from images by multiadversarial training. In ACM MM, pages 783–791, 2018.
[10] Bei Liu, Zhicheng Huang, Zhaoyang Zeng, Zheyu Chen, and Jianlong Fu. Learning rich image region representation for visual question answering. arXiv preprint arXiv:1910.13077, 2019.
[11] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In CVPR, 2022.
[12] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. NeurIPS, 34, 2021.