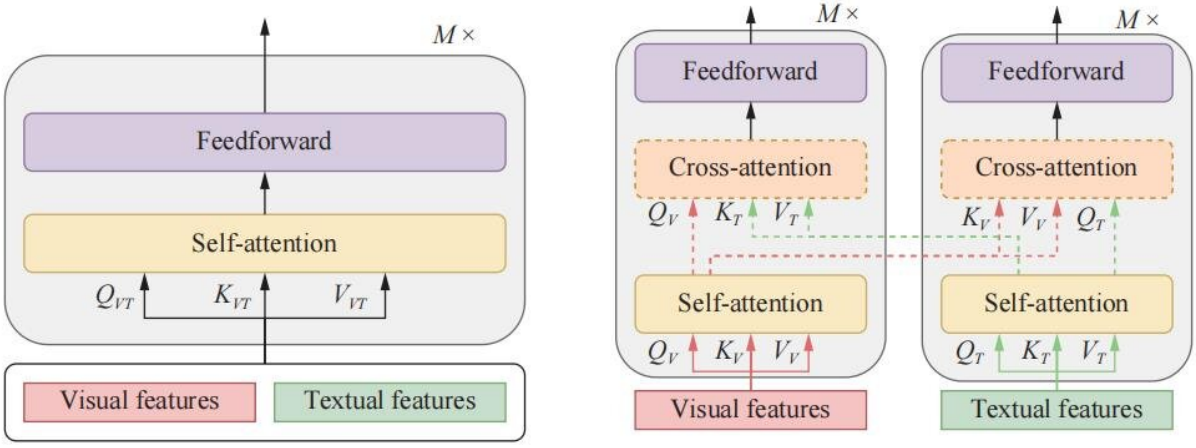
VLP: A survey on vision-language pre-training

June 20 2023



(a) Single-stream architecture (b) Dual-stream architecture

The single-stream architecture refers to that the text and visual features are concatenated together, then fed into a single transformer block, as shown in Fig. 1(a). The dual-stream architecture refers to that the text and visual features are not concatenated together but sent to two different transformer blocks independently, as shown in Fig. 1(b). Credit: Beijing Zhongke Journal Publising Co. Ltd.

In a paper published in *Machine Intelligence Research*, a team of researchers explored the problem of whether pre-trained models can be applied to multi-modal tasks and made significant progress. This paper surveys recent advances and new frontiers in vision-language pre-

training (VLP), including image-text and video-text pre-training.

To give readers a better overall grasp of VLP, researchers first review its recent advances in five aspects: feature extraction, model architecture, pre-training objectives, pre-training datasets, and downstream tasks. Then, they summarize the specific VLP models in detail. Finally, they discuss the new frontiers in VLP.

Making machines respond in ways similar to humans has been a relentless goal of AI researchers. To enable machines to perceive and think, researchers propose a series of related tasks, such as face recognition, reading comprehension, and human-machine dialog, to train and evaluate the intelligence of machines in a particular aspect. Specifically, domain experts manually construct standard datasets and then train and evaluate relevant models on them.

However, due to the limitations of related technologies, it is often necessary to train on a large amount of labeled data to obtain a better and more capable model. The recent emergence of pre-training models based on the transformer structure has alleviated this problem. They are first pre-trained via self-supervised learning that typically exploits auxiliary tasks (pre-training objectives) to mine supervision signals from large-scale unlabeled data to train the model, thereby learning universal representations.

Then, they can achieve surprising effectiveness by fine-tuning with only a tiny amount of manually-labeled data on downstream tasks. Since the advent of BERT in natural language processing (NLP), various pre-training models have sprung up in the uni-modal field. Substantial works have shown they are beneficial for downstream uni-modal tasks and avoid training a new model from scratch.

Similar to the uni-modal field, there is also a problem of less high-

quality labeled data in the multi-modal field. The natural question is whether the above pre-training method can be applied to multi-modal tasks. Researchers have explored this problem and made significant progress.

In this paper, researchers focus on mainstream vision-language pre-training (VLP), including image-text and video-text pre-training. VLP mainly learns the semantic correspondence between different modalities by pre-training on large-scale data. For example, in image-text pre-training, researchers expect the model to associate "dog" in text with what "dog" looks like in images.

In video-text pre-training, they expect the model to map objects/actions in the text to objects/actions in the video. To achieve this goal, the VLP objects and model architecture need to be cleverly designed to allow the model to mine the associations between different modalities.

To give readers a better global grasp of VLP, researchers first comprehensively review its recent advances and focus on five significant aspects: feature extraction, model architecture, pre-training objectives, pre-training datasets and downstream tasks. Then they summarize the specific state-of-the-art (SOTA) VLP models in detail. Finally, they conclude the paper and have broad discussions on new frontiers in VLP.

This paper reviews recent advances of VLP from five aspects.

Firstly, researchers describe how VLP models preprocess and represent an image, video, and text to obtain counterpart features, different models are introduced.

Secondly, they introduce the architecture of the VLP models from two different perspectives: one is single-stream versus dual-stream from a multi-modal fusion perspective, and the other one is encoder-only versus

encoder-decoder from the overall architectural design perspective.

Thirdly, an introduction on how researchers pre-train VLP models by using different pre-training objectives is given, which are crucial for learning the universal representation of vision-language. The pre-training objectives are summarized into four categories: completion, matching, temporal, and particular types.

Fourthly, researchers divide pre-training datasets into two main categories: image-language pre-training and video-language pre-training. And they provide details about representative pre-training datasets for each category. Finally, they introduce the fundamental details and goals of downstream tasks in VLP.

Then researchers summarize the specific SOTA VLP models in detail. They present the summary of mainstream image-text VLP models and mainstream video-text VLP models in tables. After that, researchers propose the future development of VLP. They suggest that based on existing works, VLP can be further developed from the following aspects: incorporating acoustic information, knowledgeable and cognitive learning, prompt tuning, model compression and acceleration, out-of-domain pre-training and advanced model architecture. Researchers hope that their survey can help others understand VLP better and inspire new work to advance this field.

More information: Fei-Long Chen et al, VLP: A Survey on Vision-language Pre-training, *Machine Intelligence Research* (2023). DOI: 10.1007/s11633-022-1369-5

Provided by Beijing Zhongke Journal Publising Co.