M6: A Chinese Multimodal Pretrainer

Junyang Lin\textsuperscript{1}\textsuperscript{*}, Rui Men\textsuperscript{1}\textsuperscript{*}, An Yang\textsuperscript{1}, Chang Zhou\textsuperscript{1}, Ming Ding\textsuperscript{2}, Yichang Zhang\textsuperscript{1}, Peng Wang\textsuperscript{1}, Ang Wang\textsuperscript{1}, Le Jiang\textsuperscript{1}, Xianyan Jia\textsuperscript{1}, Jie Zhang\textsuperscript{1}, Jianwei Zhang\textsuperscript{1}, Xu Zou\textsuperscript{2}, Zhikang Li\textsuperscript{1}, Xiaodong Deng\textsuperscript{1}, Jie Liu\textsuperscript{1}, Jinbao Xue\textsuperscript{1}, Huiling Zhou\textsuperscript{1}, Jianxin Ma\textsuperscript{1}, Jin Yu\textsuperscript{1}, Yong Li\textsuperscript{1}, Wei Lin\textsuperscript{1}, Jingren Zhou\textsuperscript{1}, Jie Tang\textsuperscript{2}\textsuperscript{†}, Hongxia Yang\textsuperscript{1}\textsuperscript{†}

\textsuperscript{1}Alibaba Group, China

\textsuperscript{2}Tsinghua University, China

\{junyang.ljy,menrui.mr,ya235025,ericzhou.zc,yichang.zyc,zheluo.wp\}@alibaba-inc.com
\{wangang.wa,jiangle.jl,xianyan.xianyanjia,wanglin.zj,zhangjianwei.zjw\}@alibaba-inc.com
\{zhikang.lzk,xiaodongdeng.dx,d,sanshuai.lj,zhiji.xjb,zhule.zhl,jason.mjx,kola.yu\}@alibaba-inc.com
\{jiufeng.ly,weilin.lw,jingren.zhou,yang.yhx\}@alibaba-inc.com
\{dm18,zoux18\}@mails.tsinghua.edu.cn,jietang@tsinghua.edu.cn

ABSTRACT

In this work, we construct the largest dataset for multimodal pretraining in Chinese, which consists of over 1.9TB images and 292GB texts that cover a wide range of domains. We propose a cross-modal pretraining method called M6, referring to Multi-Modality to Multi-Modality Multitask Mega-transformer, for unified pretraining on the data of single modality and multiple modalities. We scale the model size up to 10 billion and 100 billion parameters, and build the largest pretrained model in Chinese. We apply the model to a series of downstream applications, and demonstrate its outstanding performance in comparison with strong baselines. Furthermore, we specifically design a downstream task of text-guided image generation, and show that the finetuned M6 can create high-quality images with high resolution and abundant details.

KEYWORDS

Multimodal Pretraining; Multitask; Text-to-Image Generation

1 INTRODUCTION

Pretraining has become a focus in the research in natural language processing (NLP) [1, 2, 7, 16, 18, 19, 27, 31, 37, 44, 49]. The recent GPT-3 with over 175 billion parameters demonstrates that large models trained on big data have extremely large capacity and it can outperform the state-of-the-arts in downstream tasks especially in the zero-shot setting. Also, the rapid development of pretraining in NLP sparks cross-modal pretraining. A number of studies [4, 11, 17, 22, 24, 25, 28, 29, 38, 51] have created new state-of-the-art performances for various cross-modal downstream tasks.

A pity is that most recent studies focus on the pretraining on English data. There are lack of both large-scale datasets in Chinese and large-scale models pretrained on the data of Chinese. Therefore, in this work, we develop a large-scale dataset M6-Corpus, which consists of over 1.9TB images and 292GB texts. To the best of our knowledge, this is the largest dataset in Chinese for pretraining in both multimodality and natural language. The dataset collected from the webpages consists of different types of data and covers a large scale of domains, including encyclopedia, question answering, forum discussion, product description, etc. Also, we design sophisticated cleaning procedures to ensure that the data are of high quality.

Furthermore, in order to sufficiently leverage such a large amount of high-quality data, we propose to build an extremely large model that can process data of multiple modalities and adapt to different types of downstream tasks. Thus we propose a novel model called M6, referring to MultiModality-to-MultiModality Multitask Mega-transformer. The model is based on the transformer, and it is pretrained with multiple tasks. Pretraining endows the model with the capability of single-modality and multimodality understanding and generation. Based on the architecture of M6, we build M6-10B and M6-100B, which are scaled up to 10 billion and 100 billion parameters respectively. To be more specific, M6-100B is the recent largest model pretrained on Chinese data. We apply the model to a series of downstream applications, including product description generation, visual question answering, community question answering, Chinese poem generation, etc., and our experimental results show that M6 outperforms a series of strong baselines.

Another contribution of this work is that we first incorporate pretraining with text-to-image generation. Following Ramesh et al. [32], we leverage a two-stage framework for image generation. To be more specific, we apply a trained vector-quantized generative adversarial network to representing images with discrete image codes, and we then use the pretrained M6 to learn the relations between texts and codes. Such learning can bridge the two modalities and enables controllable text-to-image generation.

To summarize, the contributions of M6 are as follows:

• We collect and build the largest Chinese multi-modal pretraining data in industry, which includes 300GB texts and 2TB images.
• We propose M6 for multimodal pretraining in Chinese, and we scale the model size to up to 10 and 100 billion parameters.

\textsuperscript{*}Equal contribution.

\textsuperscript{†}Corresponding author.
Both M6-10B and M6-100B are the recent largest multimodal pretrained model.

- M6 is versatile and exceeds strong baselines by 11.8% in VQA, 18.4 in image captioning, and 10.3% in image-text matching. Furthermore M6 is able to generate high-quality images.
- With carefully designed large-scale distributed training optimizations, M6 has obvious advantages in training speed and greatly reduces training costs, creating the possibility for more widespread use of multi-modal pretraining.

2 DATASET

We collect and develop the largest multi-modality and text dataset in world knowledge of different fields. Also, we aim to collect data of world knowledge in Chinese, the dataset is highly required to work, the Chinese Wikipedia.

To perform large-scale multi-modal pretraining and learn complex (around 0.4B tokens) covering around 1M encyclopedia entries. An- vide both plain texts and image-text pairs on super large scale, [47] released a 100GB corpus named CLUECorpus2020, which is re-

2 DATASET

The image.

each pair the text provides a detailed description of a fraction of

difficulty for the construction of a large-scale dataset as the data

to preserve the linguistic acceptation of the texts, we implement a

For text data, we first remove HTML markups and duplicate punc-

tuations, and we only reserve characters and punctuation

tabulated marks, and we only reserve characters and punctuation

markups, duplicate punctuation marks, random combinations of

characters, etc.; (2). the images should be natural and realistic, and

the resolutions of the images need to be identifiable by humans; (3).

both the texts and images should not contain illegal content, such

as pornography, violence, etc.; (4). the images and texts should be

semantically relevant; (5). the datasets should cover a wide range of

fields, say sports, politics, science, etc., and therefore it can endow

the model with sufficient world knowledge.

2.3 Dataset Construction

Based on the requirements above, we collect data of both plain texts and image-text pairs. There are different types of data, including encyclopedia, crawled webpage, community question answering, forum, product description, etc. We present the details in Table 3. The collected corpus consists of both plain-texts and image-text pairs, which is compatible with the designed text-only and multi-modal pretraining tasks. Also, the data has a large coverage over domains, such as science, entertainment, sports, politics, commonsense of life, etc. We have also compared some characteristics of our corpus with existing datasets used for Chinese pretraining in Table 2. The size of our dataset is much larger than the previous ones. To our knowledge, this is the first large-scale, multimodal and multidomain corpus for Chinese pretraining.

We implement sophisticated preprocessing to obtain clean data.

For text data, we first remove HTML markups and duplicate punctuation marks, and we only reserve characters and punctuation marks that are in Chinese and English. We remove the topics that are shorter than 5 characters and contents shorter than 15 characters. We further apply in-house spam detection to remove sentences that contain words related to certain political issues, pornography, or words in the list of dirty, naughty, and other bad words. In order to preserve the linguistic acceptance of the texts, we implement a language model to evaluate their perplexities, and sentences with high perplexities are discarded. Only images with at least 5000 pixels are reserved for pretraining. A sequence of classifiers and heuristic rules are applied to filter out images containing illegal content. We also use a pretrained image scorer to evaluate the qualities of images. For images and texts in crawled webpages, we only consider images and their surrounding text as relevant image-text pairs. Other sentences in the webpages are discarded.

3 M6 FRAMEWORK

Multimodal pretraining leverages both the power of self-attention-based transformer architecture and pretraining on large-scale data. We endeavor to endow the model with strong capability of cross-modal understanding and generation. In this section, we describe the details of our proposed pretrained model M6, which refers to Multi-Modality-to-Multi-Modality Multitask Mega-transformer.
Table 1: Statistics of our pretraining dataset. We demonstrate the sources of our data, and we calculate the number of images, tokens, and passages, the average length, as well as the size of images and texts.

| Source              | Modality          | Images (M) | Tokens (B) | Passages (M) | Avg. Length | Image Size (TB) | Text Size (GB) |
|---------------------|-------------------|------------|------------|--------------|-------------|----------------|----------------|
| Encyclopedia        | Plain-text        | -          | 31.4       | 34.0         | 923.5       | -              | 65.1           |
| Community QA        | Plain-text        | -          | 13.9       | 113.0        | 123.0       | -              | 28.8           |
| Forum discussion    | Plain-text        | -          | 8.7        | 39.0         | 223.1       | -              | 18.0           |
| Common Crawl        | Plain-text        | -          | 40.3       | 108.7        | 370.7       | -              | 83.3           |
| Encyclopedia        | Image & Text      | 6.5        | 7.9        | 10.4         | 759.6       | 0.1            | 15.0           |
| Crawled Webpages    | Image & Text      | 46.0       | 9.1        | 106.0        | 85.8        | 1.5            | 70.0           |
| E-commerce          | Image & Text      | 8.0        | 0.5        | 8.5          | 62.1        | 0.3            | 12.2           |
| **Total**           |                   | 60.5       | 111.8      | 419.6        | 266.4       | 1.9            | 292.4          |

Figure 1: Examples of the multimodal data of M6-Corpus. We demonstrate three cases that belong to different categories, including encyclopedia, crawled webpages, and product description.

Table 2: Comparison with the existing large-scale Chinese corpora for pretraining. Our dataset is the largest dataset for Chinese pretraining. The size of texts is larger than that of the existing datasets, and the size of images is even larger than that of ImageNet.

| Dataset            | Text Size (GB) | Image Size (GB) | Multidomain |
|--------------------|----------------|-----------------|-------------|
| CN-Wikipedia       | 1.6            | ×               | ×           |
| THUCTC             | 2.2            | ×               | ×           |
| HFL                | 21.6           | ×               | ✓           |
| CLUE Corpus        | 100.0          | ×               | ✓           |
| ImageNet           | ×              | 1000            | ✓           |
| M6-Corpus          | 292.4          | 1900            | ✓           |

3.1 Visual and Linguistic Inputs

The mainstream multimodal pretraining methods transform images to feature sequences via object detection. However, the performance of the object detectors as well as the expressivity of their backbones strongly impact the final performance of the pretrained models in the downstream tasks. We observe that a large proportion of the images contain only a few objects. Take the images of the data of e-commerce as an example. We randomly sample 1M images and perform object detection on the images. The results show that over 90% of the images contain fewer than 5 objects. Also, the objects have high overlapping with each other. To alleviate such influence, we turn to a simple but effective solution following Gao et al. [12] and Dosovitskiy et al. [8]. In general, we split an image into patches and extract features of the 2D patches with a trained feature extractor, say ResNet-50. Then we line up the representations to a sequence by their positions.

The processing of the input word sequence is much simpler. We follow the similar preprocessing procedures in the previous work [4, 11, 24]. We apply WordPiece [34, 45] and masking to the word sequence and embed them with an embedding layer, following BERT [6].
### 3.2 Unified Encoder-Decoder

We integrate the image embeddings $e^i$ and the word embeddings $e^t$ into the cross-modal embedding sequence $e = \{e^i, e^t\}$. We send the sequence to the transformer backbone for high-level feature extraction. To differ their representations, we add corresponding segment embeddings for different modalities. Specifically, we leverage the self-attention-based transformer blocks for our unified cross-modal representation learning. To be more specific, the building block is identical to that of BERT or GPT, which consists of self attention and point-wise feed-forward network (FFN). On top of the transformer backbone, we add an output layer for word prediction, and thus we tie its weights to those of the embedding layer.

In the unified framework, we use different masking strategies to enable encoding and decoding. The input is segmented into three parts, including visual inputs, masked linguistic inputs, and complete linguistic inputs. We apply bidirectional masking to both the visual inputs and masked linguistic inputs, and we apply causal masking to the complete linguistic inputs. Thus the model is allowed to encode and decode in the same framework.

### 3.3 Pretraining Methods

We pretrain the model with the multitask setup, including text-to-text transfer, image-to-text transfer, and multimodality-to-text transfer. Thus the model can process information of different modalities and perform both single-modal and cross-modal understanding and generation.

**Text-to-text Transfer** As demonstrated in Figure 3, the model learns to perform text denoising and language modeling in the setting of text-to-text transfer. In text denoising, we mask the input text by a proportion, which is 15% in practice following BERT [6]. Specifically, we mask a continuous span of text with a single mask, and the model should learn to decode the whole sequence. This encourages the model to learn both recovering and length predicting. Besides, in order to improve the model ability in generation, we add a setup of language modeling, where the encoder receives no inputs and the decoder learns to generate words based on the previous context.
We scale up the model size to 10 billion parameters and 100 billion parameters, which are named M6-10B and M6-100B. The increase in model size provides a much larger capacity for the model that it can learn knowledge from more data. For the construction of M6-10B, we simply scale up the model by hyperparameter tuning. Alternatively, inspired by the recent progress in sparse activations [10, 20, 35], we combine Mixture-of-Experts (MoE) with M6 to build the version of 100 billion parameters. Note that the original MoE requires mesh-tensorflow as well as TPUs. This sets limits for a number of researchers without such resources. Thus we implement the M6-100B with MoE with our in-house framework Whale [43] to perform model parallelism with GPUs following the implementation of Megatron-LM [36].

However, directly scaling up to M6-100B is much more difficult as there are more challenges for the computation resources. Specifically, different from the conventional FFN layer, the MoE layer is a parallel combination of multiple FFN layers, each of which acts as an expert. This is also called expert parallelism. The model first learns a sparse gating network to route the tokens to specific experts. Thus each token is only sent to a small set of experts and the computation can be much less compared with that in dense models. This kind of model is highly efficient as it realizes data parallelism and expert parallelism across workers. The computation of MoE layer for a specific token $x$ can be described as below:

$$
p(x) = \frac{\exp(g(x)_i)}{\sum_j^N \exp(g(x)_j)},
$$

$$
y = \sum_{i \in T} p(x)E_i(x),
$$

To be more specific, we increment the size of hidden states and the number of layers. To better leverage GPU memory, we apply mixed-precision training and activation checkpointing to save memory. Still, the model cannot be fit into one single GPU, and thus we use model parallelism to split the feed-forward networks and attention heads to multiple GPUs following the implementation of Megatron-LM [36].

Table 4: Model sizes of M6. $n_{\text{layers}}$ is the number of transformer layers. $d_{\text{model}}$ is the dimension of hidden states in each layer. $n_{\text{heads}}$ is the number of attention heads in each layer. $n_{\text{experts}}$ is the number of experts. The M6-100B model employs multiple experts to scale up parameters to 100 billion. $n_{\text{param}}$ is the number of all parameters.

| Models   | $n_{\text{layers}}$ | $d_{\text{model}}$ | $n_{\text{heads}}$ | $n_{\text{experts}}$ | $n_{\text{param}}$ |
|----------|----------------------|---------------------|---------------------|---------------------|-------------------|
| M6-base  | 24                   | 1024                | 16                  | 1                    | 327M              |
| M6-10B   | 50                   | 4096                | 128                 | 1                    | 10B               |
| M6-100B  | 24                   | 1024                | 16                  | 1024                 | 100B              |

**Image-to-text transfer** Image-to-text transfer is similar to image captioning, where the model receives the visual information as the input, and learns to generate a corresponding description. In this setting, we add the aforementioned patch feature sequence to the input and leave the masked input blank. The model encodes the patch features, and decodes the corresponding text.

**Multimodality-to-text transfer** Based on the setup of image-to-text transfer, we additionally add masked linguistic inputs, and thus the model should learn to generate the target text based on both the visual information and the noised linguistic information. This task allows the model to adapt to the downstream tasks with both visual and linguistic inputs.
where \( g(\cdot) \) refers to the sparse gating function, and \( T \) refers to the indices of top-\( k \) values of \( g(\cdot) \). The output of MoE is a linear combination of the computation of selected expert FFNs \( f(\cdot) \).

In expert parallelism, the parameters of experts do not share across workers; while those of other parts are identical across workers. Therefore, it is necessary to perform all-to-all communication across workers at the MoE layers in order to dispatch tokens to selected experts and combine them to their original experts. While Lepikhin et al. [20] and Fedus et al. [10] implement the MoE on TPU s with one expert in each MoE layer on a TPU, we implement our model on Nvidia GPUs where there are several experts in each MoE layer on a GPU so as to fully utilize the memory. As all-to-all communication takes up a large amount of time, the optimization to improve efficiency is highly significant. We implement a series of optimization, including half-precision communication. A key problem is load balancing, which denotes that tokens can gather to only a few experts due to dynamic routing. Following Fedus et al. [10], we apply expert capacity, which refers to the number of tokens for an expert, \( C = \frac{N}{m} \), where \( C \) refers to expert capacity, \( N \) refers to the number of tokens in a batch, \( c \) refers to capacity factor (which is a hyperparameter usually larger than 1.0) and \( m \) refers to the number of experts), to alleviate this problem. Tokens out of the capacity of an expert are dropped from the computation and they are sent to next layers through residual connections. We find that the overloading problem can be severe, and this issue can be a significant one in the future research of expert models [21].

Besides the optimization in all-to-all communication, we compare the top-2 gating and top-1 gating and find that they can achieve similar model performance in perplexity, while the latter converges slightly slower. The effectiveness of top-1 gating enables faster computation. Besides, we also apply methods of memory optimization for higher efficiency. We find that gradient clipping globally can increase costs on all-to-all communication as it computes norms across all experts, and thus we apply local clipping for memory saving. We implement M6-100B with around 100 billion parameters on 128 Nvidia A100s and the speed of pretraining achieves 1440 samples/s (for samples of the sequence length of 272).

We demonstrate that using MoE structure for model size scaling is effective and it can achieve similar performance to that of M6-10B, the largest dense model, within 2-3 times shorter time. The negative log perplexity of M6-100B reaches \(-2.297\), in comparison with M6-10B that reaches \(-2.253\) but with twice of time. This shows that the MoE-based M6 model has advantages on the time basis compared with dense models with many more FLOPs.

4 APPLICATIONS

4.1 Text-to-Image Generation

Text-to-image generation has been an open problem for a long time. Previous studies mainly focused on generation on a limited domain, among which Generative Adversarial Nets (GANs) [14, 48] are dominated methods. Following Ramesh et al. [32], we leverage a two-stage framework for text-to-image generation, including discrete representation learning and language modeling.

In the first stage, we focus on transforming images into sequences of discrete codes. There are a number of alternatives for discrete code generation, including VQVAE [41] and VQGAN [9]. In the second stage, it is necessary to build a language model to learn to generate text and code sequence. In the finetuning, we add code embedding and output layers to the pretrained M6. We concat the word sequence and the aforementioned generated code sequence as the input, and we set the objective of autoregressive language modeling for the training. At the stage of inference, we input the text sequence, and the model generates codes autoregressively with top-k sampling. The last step is to transform the code sequence to an image with the generator from the first stage.

We construct a dataset for text-to-image generation in E-commerce. Specifically, we collect over 50 million product titles and images from the mobile Taobao. We apply a series of processing methods on the images to filter the unqualified. We filter the images with complex background features (characters, patterns, etc.) with the in-house white-background image detector and OCR model. We then filter the images with over 3 objects with our in-house object detector based on Faster R-CNN [33]. We finally obtain 1.8m high-quality product image-text pairs for finetuning.

We demonstrate two examples in Figure 4 and Figure 5. It can be found that the generated images have high quality and the generated objects resemble the real ones. Furthermore, in Figure 6, we find that the model is able to imagine items according to the query *military style camouflage high heels* (军旅风迷彩高跟鞋), which do not exist in the real world. The imagination ability provides room for creative design in real-world industrial scenarios, such as clothing design, shoe design, etc.

We also finetune M6 under our proposed framework on another dataset which contains 3 million images crawled from the Internet, which cover more general domains. And we find that the model can adapt to different domains. As shown in Figure 7, the model is able to generate clip arts of robots. This reveals the versatility of the framework in text-to-image generation.

4.2 Visual Question Answering

We demonstrate our experimental results on a visual question answering dataset, and we illustrate how we directly apply the pretrained M6 to the VQA application.

| Model    | Detection | Relation | Color | Number | Overall |
|----------|-----------|----------|-------|--------|---------|
| baseline | 74.0      | 64.5     | 69.0  | 41.9   | 66.8    |
| M6-base  | 79.0      | 71.0     | 70.9  | 45.2   | 71.0    |
| M6-10B   | 83.0      | 77.4     | 72.7  | 48.4   | 74.7    |

Note that the M6-10B trained on multimodal data has first been trained on plain text data, and it can actually start with much lower cross-entropy loss (around 1/3 of the loss of the one trained from random initialization). We will make a more comprehensive comparison in order to fairly evaluate the effect and efficiency of the MoE scaling.
We leverage the FMIQA dataset [13] as the Chinese visual QA benchmark, which requires the model to generate the answer given an image and a question. We implement a transformer-based model as our baseline. For the evaluation, we split the test set manually by random sampling 200 from the dataset as there is no official release of the test set, and we evaluate the overall accuracy by human evaluation. The results are demonstrated in Table 5. The pretrained M6-base outperforms the baseline by a large margin (+6.2%), which indicates the effectiveness of multimodal pretraining. Scaling up the model to M6-10B further brings 5.2% improvement.

Furthermore, we show that simply finetuning on such a small VQA dataset may limit the potential of M6. Therefore, we directly leverage M6 for the VQA application. We find that the model is able to recognize general features and provide more related knowledge based on its understanding. Though the model pretrained on pseudo-parallel image-text pairs cannot directly answer questions about detailed features, such as color, number, etc., it is able to answer questions related to background knowledge. We demonstrate some examples in Figure 8.

4.3 Image Captioning

Image captioning requires the model to generate a caption that describes the given image, which examines the model ability of cross-modal generation. We construct a dataset (named E-Commerce IC) containing pairs of product descriptions and product images from Taobao. Since too long or too short descriptions may be noisy, we discard pairs with a description longer than 100 words or less than 10 words. To avoid dirty generations, we further use an in-house tool to filter descriptions that may contain dirty words (i.e., pornographic or violent words). Finally, E-Commerce IC contains about
Figure 6: Generated images for military style camouflage high heels.

Figure 7: Generated images for a clip art of robots.

260k text-image pairs. We finetune the model with the image-to-text transfer task on E-Commerce IC.

We compare our model with a baseline of transformer in the human evaluation. We ask several annotators with the linguistic background to evaluate from three perspectives: grammar (whether a text is fluent without grammatical error), correctness (whether a text is faithful to the image), richness (whether a text is informative and attractive). During the evaluation, we randomly sample 100 images from the test set. For each image, an annotator is asked to score the text generated by different models. The scores are within the range of [0, 5].

The results in Table 6 show that M6-base outperforms the baseline in all of the metrics. We find that all models achieve high scores in grammar. However, in both correctness and richness, M6-base outperforms the baseline model by a large margin (+18.2% and +14.4%), indicating that multimodal pretraining helps to generate more faithful, informative and attractive texts. Scaling up the model to M6-10B further improves the correctness and richness (about 14.7% and 7.0%). Figure 9 illustrates two examples of image caption.

Table 6: Results on the E-Commerce IC dataset.

| Model    | Grammar | Correctness | Richness |
|----------|---------|-------------|----------|
| baseline | 4.45    | 2.58        | 3.12     |
| M6-base  | 4.61    | 3.05        | 3.57     |
| M6-10B   | 4.70    | 3.50        | 3.82     |
4.4 Question Answering

To demonstrate the potential availability in the applications of intelligent chatbots, we further employ the M6 model to generate long answers in the style of forum discussion. Human-generated questions are collected from various Chinese forums, which are input to the model to generate the answer. At the stage of inference, we append a question mark and a token "Answer:" in the prompt, which better triggers the model to generate an answer. To facilitate the generation of longer and more informative texts, we pick more complex questions.

Figure 10 demonstrates an example of general question answering. The model can illustrate a man’s own experiences that are related to the question and also point out the answer at the end. This generated text confused human annotators and passed the Turing Test. It shows that the model can not only answer general questions but also generate long fluency text.

4.5 Poem Generation

We apply the pretrained model to Chinese poem generation. The model is able to generate genres with format constraints.
想拥有与女神同样的气场与魅力？粗高跟短靴会是不错的选择。舒适平稳的粗跟设计让曼妙与优雅回归到轻松的状态，在冬日的时光里自由行走，尽显女神魅力。

North America实木床，以简约为主的风格，彰显清新的气息。边角经过细心打磨，每一个细节都做到安全不伤手。线条流畅自然，给人舒服的视觉体验，给家居带来美丽清新的装饰。

图9：两个实例的图像描述。我们仅使用图像特征作为输入，并提供无提示。

图10：一个实例的一般问题回答。包含问题的提示成功触发模型生成论坛讨论风格的长文本。

4.6 图像-文本匹配

我们评估模型在跨模态检索中的能力。具体来说，我们构建一个数据集（命名为E-Commerce ITM）包含来自移动淘宝的文本和图像对。每一对都属于一个单个商品。我们从淘宝收集235K件服装商品。对于每个产品，除了产品图片外，我们还通过重写产品标题获取一个查询。具体来说，我们进行4.6.1 图像-文本匹配。
both grow old together?

everyday life. Oh! Where can I find such a lover, that we
when we finally meet each other, only discussing about

Despite prolonged separation, we don’t have specific words
larger. Cloud
other by a long distance, and my clothes
are green trees standing by. We have been apart from each
look larger and

I ride on a horse through the east gate, and get out of my cart
to ask the way. Looking back at the way I come from, there
are green trees standing by. We have been apart from each
other by a long distance, and my clothes look larger and
larger. Clouds gather water to come, going here and there.

Generating Text:

Title: Dating Author: Bai Li (Tang Dynasty) Text:

Generated Text:

上马出东门，催车欲问道。
却顾所来径，苍苍横翠微。
浮云卷水来，纷纷予还往。
相见无杂言，但道桑麻长。
安得此中人，与之共同老。

Figure 11: One example of a generated poem, the prompt and the constraint mask work together to generate a poem based on the given title.

Table 7: Results on the E-Commerce ITM dataset. We report the accuracy on the test set.

| Model     | Accuracy | Improvement |
|-----------|----------|-------------|
| InterBert | 81.8     |             |
| M6-base   | 90.2     | 10.3%       |

named entity recognition on the title using an in-house tool, which
extracts the terms describing the style, color, category and texture
of the product. These terms are then concatenated into a natural
language query, which is used in image-text matching. The length
of each query is between 6 to 12 words. The pairs of the query
and corresponding product image are labeled as positive samples.
The negative samples are constructed by randomly substituting the
query in the original pairs.

We require the model to perform binary classification to dis-
criminate positive and negative samples. We compare our model
with InterBert [25], which is also a Chinese multi-modal pretrained
model effective in cross-modal classification downstream tasks. The
InterBert utilizes object-based features and has been pretrained on
Taobao product image-text data as well.

The results are shown in Table 7. It should be noted that the
InterBert and M6-base are both implemented with transformer-
based architecture and have similar model scales. However, M6-base
still outperforms InterBert by 10.3%. In experiments, we find the
product images generally contain relatively fewer detected objects,
which may harm the performance on this task. In contrast, M6
avoids this problem by employing the patch features and achieves
much better performance.

5 RELATED WORK

The tremendous success of NLP pretraining, including BERT [6],
GPT [2, 30, 31], and also some other related studies [1, 7, 19, 27,
49], inspires the research in cross-modal representation learning.
Also, recent studies show that the ubiquitous Transformer architec-
ture [42] can be extended to different fields, including computer
vision [3, 8]. Therefore, the simplest solution to incorporate recent
pretraining methods and cross-modal representation learning is the
extension of BERT. From the perspective of architecture, there are
mainly two types, including single-stream model and dual stream
model. Specifically, single-stream model is simple and it gradually
becomes the mainstream architecture. These models mostly differ
in their designs of pretraining tasks or the construction of input im-
age features. Basically, they are mainly pretrained masked language
modeling, masked object classification, and image-text matching.

VisualBERT [23] and Unicoder-VL [22] simply use BERT and are
pretrained with the aforementioned tasks. UNITER [4] pretrains the
model with an additional task of word-region alignment. Oscar [24]
enhances the alignment between objects and their corresponding
words or phrases. VILLA [11] further improves model performance
by adding their proposed adversarial learning methods to pretrain-
ing and finetuning. Except for pretraining tasks, some studies focus
on the features of images. Most pretraining methods for multimodal
representation learning utilize the features generated by a trained
object detector, say Faster R-CNN [33]. PixelBERT [17] accepts
raw images as input and extract their latent representations with
a learnable ResNet [15] or ResNext [46]. FashionBERT [12] splits
the images into patches with a trained ResNet without co-training.
Besides single-stream models, dual-stream models also can achieve
outstanding performance, such as VilBERT [28], LXMERT [40] and
InterBERT [25]. ViLBERT-MT [29] enhances model performance
with multi-task finetuning. ERNIE-VIL [50] enhances the model
with the application of scene graph information. In spite of these
successful cases, it still requires further researches to unmask the
success of multimodal pretraining.

6 CONCLUSIONS

In this work, we propose the largest dataset M6-Corpus for pre-
training in Chinese, which consists of over 1.9TB images and 292GB
texts. The dataset has large coverage over domains, including en-
cyclopedia, question answering, forum discussion, common crawl,
etc. We propose a method called M6 that is able to process infor-
mation of multiple modalities and perform both single-modal and
cross-modal understanding and generation. The model is scaled
to large model with 10B and 100B parameters with sophisticated
deployment, and both models are the largest multimodal pretrained
models. We apply the model to a series of downstream applications,
showing its versatility. More specifically, we design a downstream
task of text-guided image generation, and the finetuned M6 can
reach superior performance by producing images of high quality.

In the future, we will continue the pretraining of extremely large
models by increasing the scale of data and models to explore the
limit of performance, and we also endeavor to search for more
downstream applications for further generalization.
