M6: Multi-Modality-to-Multi-Modality Multitask Mega-transformer for Unified Pretraining

Junyang Lin[1*], Rui Men[1*], An Yang[1*], Chang Zhou[1], Yichang Zhang[1], Peng Wang[1],
Jingren Zhou[1], Jie Tang[2†], Hongxia Yang[1†]

[1]DAMO Academy, Alibaba Group

[2]Tsinghua University

{junyang.ljy,menrui.mr,ya235025,ericzhou.zc,yichang.zyc,zheluo.wp,jingren.zhou,yang.yhx}@alibaba-inc.com

jietang@tsinghua.edu.cn

ABSTRACT

Multimodal pretraining has demonstrated success in the downstream tasks of cross-modal representation learning. However, it is limited to the English data, and there is still a lack of large-scale dataset for multimodal pretraining in Chinese. In this work, we propose the largest dataset for pretraining in Chinese, which consists of over $1.9TB$ images and $292GB$ texts. The dataset has large coverage over domains, including encyclopedia, question answering, forum discussion, etc. Besides, we propose a method called M6, referring to **M**ulti-**M**odality-to-**M**ulti-**M**odality **M**ultitask **M**ega-transformer, for unified pretraining on the data of single modality and multiple modalities. The model is pretrained with our proposed tasks, including text-to-text transfer, image-to-text transfer, as well as multi-modality-to-text transfer. The tasks endow the model with strong capability of understanding and generation. We scale the model to 10 billion parameters, and build the largest pretrained model in Chinese. Experimental results show that our proposed M6 outperforms the baseline in a number of downstream tasks concerning both single modality and multiple modalities, and the $10B$-parameter pretrained model demonstrates strong potential in the setting of zero-shot learning.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision**.

KEYWORDS

Multi-modal pretraining; Large-scale pretraining; Cross-modal understanding and generation

1 INTRODUCTION

Pretraining has recently greatly promoted the development of natural language processing (NLP). A series of studies in pretraining [1, 2, 8, 15, 17, 18, 24, 29, 35, 41, 46] have gradually pushed the limit of model performance in natural language understanding and even natural language generation. Besides, pretraining leveraging large-scale data enables the building of extremely large models with large capacity. The recent GPT-3 with over 175 billion parameters demonstrates that large models can achieve outstanding performance even in the setting of few-shot or zero-shot learning. The rapid development of pretraining in NLP sparkles cross-modal pretraining. Similar to the pretrained models for natural language processing, those models for mutlimodal understanding and generation also incorporate the architecture of self-attention-based transformer [39]. A number of studies [4, 10, 16, 19, 21, 22, 26, 27, 36, 50] have created new state-of-the-art performance for various cross-modal downstream tasks. The developments reflect the significance of pretraining in today's research and application in natural language processing and cross-modal representation learning.

However, most studies, especially multimodal pretraining, focus on the pretraining on English data. There is a lack of large-scale dataset in Chinese for multimodal pretraining, and it is even hard to find a popular cross-modal downstream task with Chinese benchmark datasets and strong baselines. Therefore, in this work, we first propose 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 wide range 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 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 model called **M6**, which refers to **M**ulti-**M**odality-to-**M**ulti-**M**odality **M**ultitask

---

Mega-transformer. The model is based on the self-attention-based transformer, and it is pretrained with our proposed tasks. Pretraining endows the model with the capability of single-modal and multimodal understanding and generation.

Based on the architecture of M6, we scale the model to **10B** parameters, which is the **largest** model in Chinese pretraining. We conduct experiments on the existing Chinese multimodal datasets and the Chinese natural language benchmarks. We show that M6 outperforms the baselines in multimodal downstream tasks, and the large M6 with $10B$ parameters can reach a better performance. In the evaluation of natural language processing, M6 outperforms the strong baseline with $2.6B$ parameters in natural language downstream tasks especially in the setting of zero-shot learning. This demonstrates the model's transferability and generality.

In brief, our contributions are:

- We propose the **largest** dataset for multimodal and NLP pretraining in Chinese, which consists of over $1.9TB$ multimodal data and $292GB$ plaintext data.
- We propose M6 that is able to perform single-modal and cross-modal understanding and generation, and we scale the model up to **10B** parameters and build the **largest** pretrained model in Chinese.
- Experimental results show that M6 outperforms the baselines in both multimodal and NLP downstream tasks, and the $10B$-parameter model has strong capability of zero-shot learning.

2 DATASET

We collect and develop the largest multi-modality and text dataset in Chinese for now, which is one of the key contributions of this paper. In this section, we first identify the limitations of existing datasets and then describe the construction and preprocessing procedure of our proposed dataset.

2.1 Existing Datasets

The construction of large-scale corpus with high quality and domain coverage is crucial to Chinese pretraining. In early previous works, the Chinese Wikipedia[1] is one of the most frequently used datasets to train Chinese language models. It contains $1.6GB$ texts (around $0.4B$ tokens) covering around $1M$ encyclopedia entries. Another corpus with a comparable size is the THUCTC[37] dataset, which includes $740K$ news articles. However, with the rapidly increasing capacity of recent language models, the scale of these existing datasets is clearly insufficient. Recently, Cui et al. [5] employs unreleased extended data that are 10 times larger than the CN-Wikipedia to pretrain their Chinese language model. Xu et al. [45] released a $100GB$ corpus named CLUECorpus2020, which is retrieved from the multilingual Common Crawl dataset. However, the scale of the datasets is still insufficient to facilitate super large-scale pretraining compared with existing English pretrained datasets. For example, GPT-3 contains 175B parameters and is trained on $570GB$ texts. Meanwhile, the dataset should contain image-text pairs rather than plain texts for multi-modal pretraining.

2.2 Standards for a High-quality Dataset

To perform large-scale multi-modal pretraining and learn complex world knowledge in Chinese, the dataset is highly required to provide both plain texts and image-text pairs in super large scale, covering a wide range of domains. In order to perform large-scale multi-modal pretraining in Chinese, we focus on the construction of large-scale datasets. Specifically, while we unify our pretraining for both natural language and multimodality, we construct large datasets of both plain texts and image-text pairs. We are interested in obtaining large-scale data that cover a wide range of domains, so that it is possible for the model to learn the complex world knowledge of different fields. Also, we aim to collect data of multiple modalities for multimodal pretraining. This raises the difficulty for the construction of large-scale dataset as the data for multimodal pretraining are usually image-text pairs, where in each pair the text provides a detailed description of a fraction of the image.

Though there are tremendous texts and images on the world wide web, a high-quality corpus for multimodal pretraining should satisfy the following properties: (1). the sentences should be fluent natural language within a normal length, and should not contain meaningless tokens, such as 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 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 above requirements, we collect data of both plain texts and image-text pairs. There are different types of data, including encyclopedia, crawled webpages, community question answering, forum discussion, product description, etc. We present the details in Table 1. The collected corpus consists of both plain-texts and image-text pairs, which is compatible with the designed text-only and multimodal 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

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 image and text.

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

Image	Source & Text
	Source: Encyclopedia 广东草龟是属于曲颈龟亚目龟科的一种草龟。又称黑颈乌龟。 The Guangdong tortoise is a kind of tortoise belonging to Cryptodira. It is also known as black-necked turtle.
	Source: Crawled Webpages 根据之前信息，马斯克称Cybertruck将配备三种动力版本，其中包括单电机后驱，双电机后驱和三电机全驱版本。 According to the previous news, Elon Musk said that Cybertruck will be equipped with three versions of power, including a single-motor rear drive, a dual-motor rear drive and a three-motor full-drive version.
	Source: E-commerce 柔软的针织面料就能给人一种舒服的感觉，大篇幅的印花以点缀的作用让整体显得更加青春阳光，宽松简约落肩尽显时尚风范，十分适合日常穿搭。 The softly knitted fabric can give people a comfortable feeling. The large-length prints make the whole look youthful and sunny. Its loose and simple extended sleeves look fashionable, and it is very suitable for daily wear.

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	✓

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 texts as relevant image-text pairs. Other sentences in the webpages are discarded.

3 APPROACH

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 **M**ulti-**M**odality-to-**M**ulti-**M**odality **M**ultitask **M**ega-transformer.

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. Besides, we believe that object features are not the best choice of image representation in this

Source & Text
Source:Encyclopedia 神经网络是一种运算模型，由大量的节点（或称神经元）之间相互连接构成，其在模式识别、智能机器人等领域已经成功解决了许多实际问题。 Neural network is a computational model, which is composed of a large number of nodes (or neurons) connected to each other. It has successfully solved many practical problems in the fields of pattern recognition and intelligent robots.
Source:Community QA 宽带连接不上、本地连接不见了、是不是网卡坏了？ 回答：这个问题很简单，最大的可能就是你把驱动误删了。 The broadband connection is not available, the local connection is missing, is the network card broken? Answer: This problem is very simple. The most likely reason is that you deleted the driver by mistake.
Source:Forum discussion 如何评价1700亿参数的GPT-3？ 回答：GPT-3依旧延续自己的单向语言模型训练方式，不过这次的训练数据有570GB。 How to evaluate the 170 billion parameter GPT-3? Answer: GPT-3 continues its single-direction language model training method, but this time the size of its training dataset is 570GB.
Source:Common Crawl 北京市互联网金融行业协会的前身为北京市网贷行业协会，成立于2014年12月，是中国第一个网贷行业协会组织。 The predecessor of the Beijing Internet Finance Industry Association was the Beijing Internet Loan Industry Association. It was established in December 2014 and is the first online loan industry association in China.

Figure 2: Examples of the plain text data of M6-Corpus. We demonstrate three cases that belong to different categories, including encyclopedia, community QA, forum discussion, and common crawl.

Table 3: Model sizes of M6. n_{layers} **is the number of transformer layers.** d_{model} **is the dimension of hidden states in each layer.** n_{heads} **is the number of attention heads in each layer.** n_{param} **is the number of all parameters.**

Models	n_{layers}	d_{model}	n_{heads}	n_{param}
M6-base	24	1024	16	327M
M6-large	50	4096	128	10B

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 perform encoding and decoding 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 [7]. 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.

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.

3.4 Scaling up to 10 Billion Parameters

In order to sufficiently leverage the large-scale dataset that we build, we endeavor to build an extremely large model based on the architecture of M6. We choose a simple solution that we increase

context as we observe that a large proportion of the images contain only a few objects. Take the images of the e-commerce data 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 and effective solution following Gao et al. [11] and Dosovitskiy et al. [9]. In general, we split an image into patches and extract features of the 2D patches with a trained feature extractor, e.g. 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 apply WordPiece [34, 42] and masking to the word sequence and embed them with an embedding layer, following BERT [7].

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
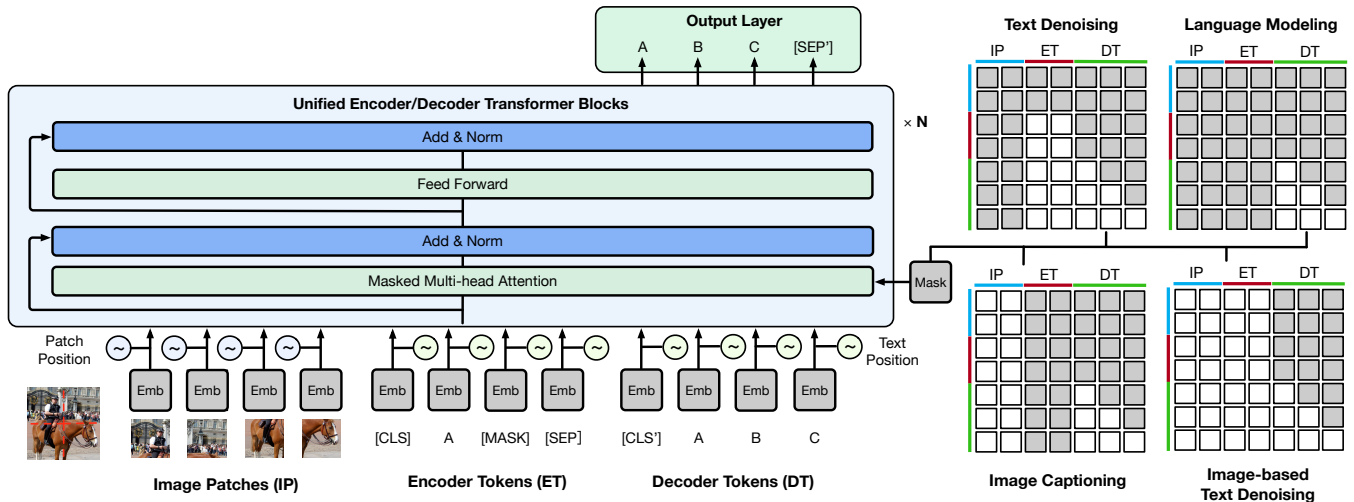
Figure 3: An overview of the pretraining tasks for M6. The design of masking strategies allows the learning of different tasks under the same framework. M6 is pretrained with image-based text denoising, image captioning, text denoising, and language modeling.

Table 4: Results on the FMIQA dataset. We report both the overall accuracy and the accuracy on specific question types.

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-large	**83.0**	**77.4**	**72.7**	**48.4**	**74.7**

Table 5: Results on the E-Commerce ITM dataset. We report the accuracy on the test set.

Model	Accuracy	Delta
InterBert	81.8	-
M6-base	**90.2**	8.4

the model size by hyperparameter tuning. We present the details of deployment in Section 5.

3.5 Pretraining Details

We implement both M6-base with $327M$ parameters and M6-large with $10B$ parameters, whose detailed hyperparameters are listed in Table 3. All the weight parameters are randomly initialized based on the Gaussian distribution of zero mean and standard deviation of 0.02, following Devlin et al. [7]. For the consistency between word embedding and object representation, we transform the patch features of 2048 dimensions to d_{model} through linear projection. The M6-base has been pretrained on all samples for over an epoch, while the M6-large has only been pretrained for around $30M$ samples due to the limitation of hardware resources. More details are presented in the appendix.

4 EXPERIMENTS

In this section, we introduce our experiment settings and results on both multimodal and NLP downstream tasks. We also provide some analyses and case studies.

4.1 Downstream Tasks

4.1.1 Multimodal Downstream Tasks. We compare M6 with competitive baselines on several multimodal downstream tasks, including visual QA, image-text matching and image captioning. These

tasks evaluate the model's ability on cross-modal understanding from various perspectives.

Visual Question Answering We leverage the FMIQA dataset [12] 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 4. 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-large further brings 5.2% improvement.

To carefully inspect the benefit of pretraining, we also report the accuracy scores on different types of questions. We find that M6-base greatly improves the baseline by 10.1% on the questions about the spatial relations between the objects. Meanwhile, M6-base surpasses the baseline by 7.9% on the counting questions, indicating that the pretraining may facilitate numeric reasoning. The pretraining achieves 6.8% improvement on the questions related to object and action detection. Interestingly, we find the improvement on color recognition questions is relatively marginal (+2.8%), which may reflect that our pretraining contributes little to the ability of color recognition.

Image-text Matching We evaluate the model's ability in cross-modal retrieval. Specifically, we construct a dataset (named E-Commerce

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-large	**4.70**	**3.50**	**3.82**

Table 7: Zero-shot templates for NLP downstream tasks. Words in bold denote the input text.

Datasets	Templates
TNEWS	标题:**title**; 关键词:**keywords**; 分类:**label** Title:**title**;Keywords:**keywords**;Label:**label**
ChID	成语填空:... **prefix** [MASK] **postfix**... Idiom cloze:... **prefix** [MASK] **postfix**...
CMRC2018	**paragraph** 问题: **question**? 回答: **paragraph** Question: **question**? Answer:

Table 8: Results of the models on the three NLP downstream tasks with zero-shot settings. The results of the baselines are those reported in their original papers. *Random* denotes randomly choosing from the candidate lists.

Models	n_{param}	TNEWS Acc.	CMRC F1	CHID Acc.
Random	-	25.0	-	10.0
CPM-medium	334M	61.8	8.60	52.4
CPM-large	2.6B	70.3	13.37	**68.5**
M6-base	327M	65.5	13.58	61.7
M6-large	10B	**72.7**	**13.88**	65.3

ITM) containing pairs of texts and images from the mobile Taobao. Each pair belongs to a single item. More details of the data statistics and data processing are referred to the appendix. We require the model to perform binary classification to discriminate positive and negative samples. We compare our model with InterBert [22], 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 5. It should be noted that the InterBert and M6-base are both implemented with transformer-based architecture and have similar model scale. However, M6-base still outperforms InterBert by 9.3%. 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.

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. More details of the construction of the finetuning dataset are referred to the appendix. We finetune the model with the image-to-text transfer task. At the stage of inference, we apply beam search with a beam size of 5. We compare our model with a baseline of transformer in the human evaluation. We ask several annotators with 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 on 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 10B parameters further improves the correctness and richness (about 14.7% and 7.0%).

4.1.2 NLP Downstream Tasks. We focus on the evaluation of the zero-shot learning of our model on several NLP downstream tasks, including text classification, reading comprehension, and cloze. M6 is compared with the recent largest published Chinese Pretrained Language Model [48], which consists of around 2.6*B* parameters and has achieved strong performance on many NLP downstream

tasks in the settings of zero-shot. The templates we used for zero-shot settings are shown in Table 7. Results on these NLP downstream tasks are shown in Table 8.

Text Classification We first evaluate the model's zero-shot ability on text classification tasks. This task requires the model to categorize text (i.e., sentences, paragraphs, or sentence pairs) into organized groups. We conduct our experiments on TouTiao News Titles Classification (TNEWS) [44]. We calculate the perplexity of each candidate sentence-label pair of the validation dataset. Pairs with the lowest perplexities are treated as the predictions. Following Zhang et al. [48], we convert the original tasks to the task of 4-way classification tasks with 3 negative samples for better efficiency as there are more than 10 kinds of labels. We repeat this procedure three times to make the result more stable and report the average accuracy.

Experimental results show that M6-base outperforms CPM-medium by 4.6% on TNEWS. M6-large further surpasses CPM-large by 3.3% and achieves a new state-of-the-art, even if it is pretrained with much fewer samples. This indicates that our models have better understanding of the semantic differences of labels in different categories.

Reading Comprehension A span-extraction dataset for Chinese Machine Reading Comprehension (CMRC2018) is chosen to evaluate the model's ability of single-modal understanding [6]. This task requires the model to extract spans from the given passages to answer several questions. We concatenate the given passage and one of the following questions as the input of M6 pretrained models. Slightly different from the original extracting task, we ask the model to generate an answer according to the input text.
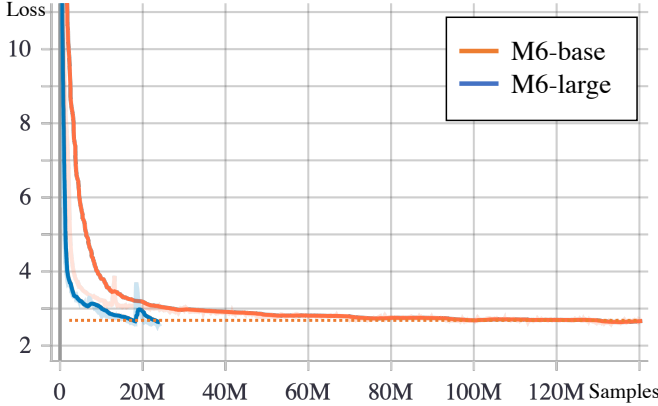
Figure 4: The comparison of loss of M6-base and M6-large. It can be found that M6-large can reach the same loss with much fewer pretraining steps.

Our M6 models show comparable performance with CPM-large on Reading Comprehension. However, without finetuning, M6 models are more likely to generate a long sentence to answer the given question which may limit the improvement in F1-score when scaling up the model to M6-large.

Cloze This task requires the model to choose the correct word from the given candidate lists to fill in the blanks. We conduct experiments on the Chinese IDiom cloze test dataset (ChID)[49]. In this task, we utilize a pattern similar to the pretraining task of text denoising. The blank in the passage is replaced by the token "[MASK]" in the encoder input and is replaced by each idiom from the given candidate list in the decoder input. A similar procedure as mentioned above is applied to get predictions.

With the help of text denoising, M6-base outperforms CPM-medium by a large margin (17.7%). Our model can have a better view of the whole passage with the bi-directional attention compared with CPM which only uses single-directional attention. However, limited by hardware conditions, passages longer than 256 tokens are truncated when we evaluate our models on CHID. This makes the performance of the M6-large not further improved compared with that of the M6-base. Another possibility is that the insufficient pretraining affects the performance of the large model. We have found that the more samples that it has been pretrained, the higher performance it can achieve. We demonstrate the relevant details in the appendix.

4.2 Analysis

4.2.1 Loss Evaluation. To evaluate the difference between the base version and the large version, we first take a look at their changes in cross-entropy loss to learn about their perplexities of generation. We demonstrate the results in Figure 4. From the figure, it can be found that the large model with 10 billion parameters has lower training loss in the whole process, and its loss degrades much faster than the base model.

4.2.2 Prompt Design for Different Tasks. The pretrained model has strong capability of performing different types of text generation

or visual-information-based text generation. Recent studies [2, 13, 33] show that the design of prompt can significantly affect the model performance in the tasks of fewshot or zero-shot learning. Therefore, we attempt to design different prompts to evaluate whether our pretrained model is capable of performing different tasks without finetuning. A common requirement for such ubiquitous model is whether it can answer questions. We thus conduct analyses of visual question answering and general question answering. In visual question answering, as the model should answer questions based on the image, we design relatively simple questions, such as "What/Where is this", and add a question mark and tokens "Answer:" for the model to better recognize it as a question. For general question answering, we make the questions more complex that they can trigger a discussion for the generation of longer and more informative texts. Similarly, we also add a question mark and prompt "Answer:" at the end.

From Figure 5, it can be found that the model can recognize the main feature of the image, which refers to "The Great Wall", and provide a brief description, which is consistent with the truth. However, when the model meets questions about specific features of the image, the model finds it difficult to answer. We assume that this results from the data where most of the images and texts are loosely connected, while the caption may not describe a specific feature, such as color, number, etc. This leaves an issue about how to pretrain a model that can effectively provide answers about these kinds of features.

Figure 6 demonstrates an example of general question answering. The case shows that the model is capable of generating long texts. This designed prompt as question triggers the model to answer in the style of forum discussion. 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 not only has the ability of long text generation but it can also answer general questions.

5 DEPLOYMENTS

In order to evaluate the effects of larger models with more parameters, we choose the simplest solution by scaling up model with hyperparameter tuning. By adding layers and increasing the size of hidden layers, we start from the base version M6 to the large version with 10B parameters. However, it cannot be fit into a single GPU as its requirements for memory is incredible. To be more specific, even if we apply FP16 with loss in precision, the model parameters take up $20GB$. Besides, the activation as well as the optimizer states can take up much more space. In concern of this problem, we have attempted a series of methods for efficient implementation. The first step to save memory is the application of mixed precision training, where we use O2 for higher efficiency that only the master weights are in FP32. We further apply activation checkpointing to save memory, and thus much activation need to be recomputed during backpropagation while this saves memory. While this only allows training a model of around $1B$ parameters, there is still more work to be done. We have implemented two solutions that can remove the obstacle. One is that we apply ZeRO optimizer [30]. ZeRO, referring to Zero Redundancy

Prompt:
这是什么？ 回答：
What is this? *Answer:*

Generated Text:
长城。长城是中国古代劳动人民智慧的结晶，是中华民族精神的象征，是世界文明的标志。它不仅仅是一种建筑形式，更是一个国家、一个民族的精神象征。
The Great Wall. The Great Wall reflects the great talent of Chinese people, and it is an important representation of the spirit of Chinese, and it is also a logo of human civilization. It is a form of architecture, and it is even a spiritual representation of a country and a people.

Prompt:
这里是哪儿？ 回答：
Where is this? Answer:

Generated Text:
故宫博物院，位于北京市中心，是中国明清两代的皇家宫殿，也是世界上现存规模最大，保存最为完整的木质结构古建筑之一。
The Forbidden City. The Museum of the Forbidden City is located at the center of Beijing, and it is the palace of the Ming and Qing Dynasty. It is one of the largest and the most well-preserved wooden ancient buildings.

Prompt:
他的衣服是什么颜色的？ 回答：
What color is his clothes? *Answer:*

Generated Text:
黑色。他的鞋子是白色。
Black. His shoes are white.

Prompt:
这个人的头盔是什么颜色的？ 回答：
What color is this man's helmet? Answer:

Generated Text:
蓝色的。
Blue.

Figure 5: Examples of general visual question answering without finetuning. We turn the origin questions to the designed pattern , with typical tokens such as "?" and "Answer:". The pretrained model can recognize the question and provide the answer as well as some further description. However, when meets questions about specific features of the image, such as color, our M6 fails to generate correct answers.

Prompt:
出国留学的理由是什么？ 回答：
What is the reason for studying abroad? *Answer:*

Generated Text:
谢邀。这个问题我觉得我有资格回答一下，因为我就是一个留学生。首先说一下我的家庭背景吧，父母都是普通工薪阶层，没有什么特别的背景，但是我从小就很聪明，也很努力，所以从小学到高中都是班里前几名，虽然成绩一直不是特别拔尖，但还是可以保持在年级前五十的水平。大学的时候，我选择了出国读研究生，现在已经毕业两年多了，回想起来，当初选择出国的原因真的很简单，就是想去看看外面的世界，体验一下不一样的生活。
Thanks for your invitation. I think I can answer this question, as I am an international student. For starters, I'd like to talk about my family. My parents are all working-class people, and they have nothing special. Yet since I was a little child I have been very clever and hard-working, and thus I have always been the top students in my class. Although I was not the best, I could keep myself in the top 50 of the students of our grade. When I was in college, I chose to study abroad for graduate school, and now I have graduated from school for two years. Actually the reason for my studying abroad is quite simple. I just wanted to go outside and experience different life.

Figure 6: One example of general question answering. We use the similar prompt as the question triggers the model to generate long texts in the style of forum discussion.

Optimizer, uses the computation and memory of data parallelism to reduce memory and compute requirements of each GPU by partitioning weight parameters, gradients, and optimizer states across GPUs and CPUs in the distributed environment. Furthermore, it is available to apply ZeRO-Offload [31] to save more memory as possible by offloading the states to CPUs. The other solution is pipeline parallelism [40]. Pipeline parallelism partitions the layers to stages across the available GPUs and the model can be trained in parallel. To further improve the training efficiency, we apply gradient accumulation to speed up.

6 RELATED WORK

The tremendous success of NLP pretraining, including BERT [7], GPT [2, 28, 29], and also some other related studies [1, 8, 18, 24, 46], inspires the research in cross-modal representation learning. Also, recent studies show that the ubiquitous Transformer architecture [39] can be extended to different fields, including computer vision [3, 9]. Therefore, the simplest solution to incoporate 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 image features. Basically, they are mainly pretrained masked language modeling, masked object classification, and image-text matching. VisualBERT [20] and Unicoder-VL [19] simply uses BERT and pretrains with the aforementioned tasks. UNITER [4] pretrains the model with an additional task of word-region alignment. Oscar [21] enhances the alignment between objects and their corresponding words or phrases. VILLA [10] further improves model performance by adding their proposed adversarial learning methods to pretraining and finetuning. Except for pretraining tasks, some studies focus on the features of image. Most pretraining methods for multimodal representation learning utilize the features generated by a trained object detector, say Faster R-CNN [32]. PixelBERT [16] accepts raw images as input and extract their latent representations with a learnable ResNet [14] or ResNext [43]. FashionBERT [11] 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 [26], LXMERT [38] and InterBERT [22]. ViLBERT-MT [27] enhances model performance with multi-task finetuning. ERNIE-ViL [47] enhances the model with the application of scene graph information.

7 CONCLUSION

In this work, we propose the largest dataset M6-Corpus for pretraining in Chinese, which consists of over $1.9TB$ images and $292GB$ texts. The dataset has large coverage over domains, including encyclopedia, question answering, forum discussion, common crawl, etc. In order to sufficiently leverage the large-scale data, we endeavor to construct an extremely large model that has great capacity of learning big data. We propose a method called M6 that is able to process information of multiple modalities and perform both single-modal and cross-modal understanding and generation. The model is scaled to large model with 10 billion parameters with sophisticated deployment, and the $10B$-parameter M6-large is the largest pretrained model in Chinese. Experimental results show that our proposed M6 outperforms the baseline in a number of downstream tasks concerning both single modality and multiple modalities. The 10B-parameter pretrained model has greater capacity, and it has outstanding performance in the setting of zero-shot learning. In the future, we will continue the pretraining of extremely large models by increasing data to explore the limit of its performance.

REFERENCES

[1] Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. In International Conference on Machine Learning. PMLR, 642–652.

[2] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. arXiv preprint arXiv:2005.14165 (2020).

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In European Conference on Computer Vision. Springer, 213–229.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In ECCV 2020. 104–120.

[5] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. arXiv preprint arXiv:2004.13922 (2020).

[6] Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2018. A span-extraction dataset for chinese machine reading comprehension. arXiv preprint arXiv:1810.07366 (2018).

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT 2019. 4171–4186.

[8] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified Language Model Pre-training for Natural Language Understanding and Generation. In NeurIPS 2019. 13042–13054.

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

[10] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-Scale Adversarial Training for Vision-and-Language Representation Learning. In NeurIPS 2020.

[11] Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In SIGIR 2020. 2251–2260.

[12] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. arXiv preprint arXiv:1505.05612 (2015).

[13] Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making Pre-trained Language Models Better Few-shot Learners. arXiv preprint arXiv:2012.15723 (2020).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In CVPR 2016. 770–778.

[15] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. arXiv preprint arXiv:2006.03654 (2020).

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

[17] Zihang Jiang, Weihao Yu, Daquan Zhou, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. 2020. Convbert: Improving bert with span-based dynamic convolution. arXiv preprint arXiv:2008.02496 (2020).

[18] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. CoRR abs/1909.11942 (2019).

[19] Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2019. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training. CoRR abs/1908.06066 (2019).

[20] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. CoRR abs/1908.03557 (2019).

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. CoRR abs/2004.06165 (2020).

[22] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. Interbert: Vision-and-language interaction for multi-modal pretraining. arXiv preprint arXiv:2003.13198 (2020).

[23] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In ECCV 2014. 740–755.

[24] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019).

[25] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In ICLR 2019.

[26] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In NeurIPS 2019. 13–23.

[27] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2019. 12-in-1: Multi-Task Vision and Language Representation Learning. CoRR abs/1912.02315 (2019).

[28] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf (2018).

[29] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. [n.d.]. Language models are unsupervised multitask learners. ([n. d.]).

[30] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2019. Zero: Memory optimization towards training a trillion parameter models. arXiv preprint arXiv:1910.02054 (2019).

[31] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. 2021. ZeRO-Offload: Democratizing Billion-Scale Model Training. arXiv preprint arXiv:2101.06840 (2021).

[32] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In NeurIPS 2015. 91–99.

[33] Timo Schick and Hinrich Schütze. 2020. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. arXiv preprint arXiv:2009.07118 (2020).

[34] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In ACL 2016.

[35] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In ICML 2019. 5926–5936.

[36] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In ICLR 2020.

[37] Maosong Sun, Jingyang Li, Zhipeng Guo, Z Yu, Y Zheng, X Si, and Z Liu. 2016. Thuctc: an efficient chinese text classifier. GitHub Repository (2016).

[38] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In EMNLP-IJCNLP 2019. 5100–5110.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In NeurIPS 2017. 5998–6008.

[40] Ang Wang, Xianyan Jia, Le Jiang, Jie Zhang, Yong Li, and Wei Lin. 2020. Whale: A Unified Distributed Training Framework. arXiv preprint arXiv:2011.09208 (2020).

[41] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Jiangnan Xia, Liwei Peng, and Luo Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. arXiv preprint arXiv:1908.04577 (2019).

[42] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016).

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In CVPR 2017. 1492–1500.

[44] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. CLUE: A Chinese Language Understanding Evaluation Benchmark. In COLING 2020. 4762–4772.

[45] Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. CLUECorpus2020: A Large-scale Chinese Corpus for Pre-trainingLanguage Model. arXiv preprint arXiv:2003.01355 (2020).

[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In NeurIPS 2019. 5754–5764.

[47] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. arXiv preprint arXiv:2006.16934 (2020).

[48] Zhengyan Zhang, Xu Han, Hao Zhou, Pei Ke, Yuxian Gu, Deming Ye, Yujia Qin, Yusheng Su, Haozhe Ji, Jian Guan, et al. 2020. CPM: A Large-scale Generative Chinese Pre-trained Language Model. arXiv preprint arXiv:2012.00413 (2020).

[49] Chujie Zheng, Minlie Huang, and Aixin Sun. 2019. ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In ACL 2019. 778–787.

[50] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified Vision-Language Pre-Training for Image Captioning and VQA. In AAAI 2020. 13041–13049.

A IMPLEMENTATION

A.1 Pretraining Details

For the processing of the visual inputs, we resize the raw images to the resolution of 224 * 224 and then split them into 4 * 4 patches. For each patch, we obtain its feature of a 2048-dimensional vector with a trained ResNet-50. For the text processing, we tokenize the texts with BERT's tokenizer based on WordPiece [42] and directly use BERT-base's embedding layer for word embedding. The vocabulary size is 21128. We control the length of the input for different tasks. The length of the patch feature sequence is fixed to 16, and the length of text, including masked linguistic input and complete linguistic input, is set to 512 by truncating or padding. Specifically, in the tasks where there are both masked text and complete text, we truncate or pad each part to the length of 256.

We implement both M6-base with 327M parameters and M6-large with 10B parameters, whose detailed hyperparameters are listed in Table 3. All the weight parameters are randomly initialized based on the Gaussian distribution of zero mean and standard deviation of 0.02, following Devlin et al. [7]. For the consistency between word embedding and object representation, we transform the patch features of 2048 dimensions to d_{model} through linear projection.

We implement M6-base on 16 NVIDIA V100-32G, and we implement M6-large on 96 NVIDIA V100-32G. We pretrain the model with AdamW [25] optimizer with $\beta_1, \beta_2 = (0.9, 0.999)$ and $\epsilon = 1e-8$. We set the peak learning rate to 1.5e-4, and we apply a scheduler with cosine decay with a warmup ratio of 0.01 to control the learning rate. The batch size set for M6-base is 1024. and the batch size set for M6-large is 6144 with gradient accumulation set to 8. For natural language pretraining, we only pretrain the models with text-to-text transfer, and for multimodal pretraining, we pretrain the model with both image-to-text transfer and multimodality-to-text transfer. The M6-base for natural language pretraining has been pretrained for 146K steps, and the M6-large for natural language pretraining has been pretrained for 4500 steps. The M6-base for multimodal pretraining has been pretrained for 320K steps, and the M6-large for multimodal pretraining has been pretrained for 8000 steps. Note that M6-large has been pretrained for fewer steps due to the limitation of computation resource, and thus this may influence the final performance of the M6-large model. We demonstrate the results on Figure 7. It can be found that the increase in pretrained samples gradually improve the model performance of the M6-large on ChID in the setting of zero-shot learning. This shows that the large model has greater capacity and the increase in data can further improve its performance.

A.2 In-house Datasets

Image-Text Matching To construct the E-Commerce ITM dataset, we collect 235K products in the clothing industry from Taobao. For each product, aside from the product image, we obtain a query by rewriting the product title. Specifically, we conduct named entity recognition on the title using a in-house tool, which extracts the

---

Table 9: Data statistics of the datasets of the downstream tasks.

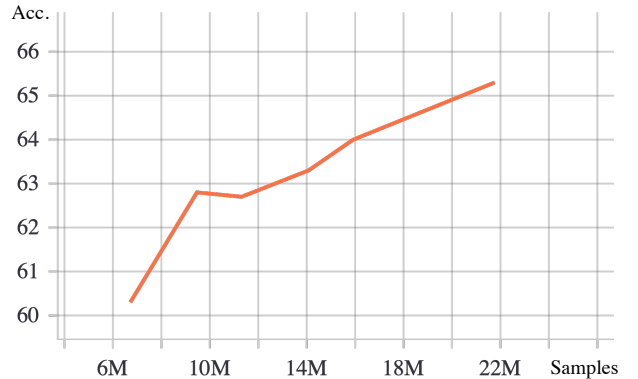| Datasets	Training	Validation	Testing
FMIQA[2]	239K	0.2K	0.2K
E-Commerce ITM	460K	5K	5K
E-Commerce IC	250K	5K	5K
TNEWS	53K	10K	10K
CHID	84K	3K	3K
CMRC	10K	3K	1K

Figure 7: The performance of M6-large on ChID. The more samples M6 has been pretrained on, the higher performance it can achieve.

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. The whole dataset is randomly split into 460K, 5K, 5K image-text pairs for the training, validation and testing sets, respectively. The images in the validation and testing sets are ensured not appearing in the training set. The ratio of positive and negative pairs in each data split is around 1:1.

Image Captioning We construct the E-Commerce IC dataset by collecting pairs of product descriptions and product images from Taobao. To obtain a clean dataset, we remove the products with less than 100 user clicks. 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 a in-house tool to filter descriptions that may contain dirty words (i.e., pornographic or violent words). Finally, the dataset is consists of 260K image-text pairs that covers four categories: clothing, food, furniture and electronic products. The average length of descriptions is about 73 words. We random split the dataset into 250K, 5K, 5K instances for training, validation and testing, respectively.

A.3 Public Datasets

Visual Question Answering We leverage the FMIQA dataset [12] as the Chinese visual QA benchmark, which requires the model to generate the answer given an image and a question. The images of FMIQA are derived from MS-COCO[23] and the question-answer pairs are annotated in Chinese by crowd-sourcing. The released part of FMIQA includes 165K training and 75K validation samples. However, there is no official release of the original test set. In our experiments, we re-split the original validation dataset into 3 pieces: 200 samples for validation, 200 for human testing and the remaining samples merging into the training set.

Text Classification We conduct our experiment on TouTiao News Titles Classification (TNEWS) [44]. The news data were published by TouTiao and the dataset consists of 73360 news titles. Each title is labeled with a category and the total number of categories is 15.

Cloze We conduct experiments on the Chinese IDiom cloze test dataset (ChID)[49]. The datasets consists of around 498,611 passages with 623,377 blanks covered from news, novels, and essays. It consists of 3,848 candidate Chinese idioms. For each blank in the passage, the dataset provides ten options, with one golden idiom, several similar ones, and some other randomly chosen ones.

Reading comprehension A span-extraction dataset for Chinese Machine Reading Comprehension (CMRC2018) is chosen to evaluate the model's ability of single-modal understanding [6]. It consists of 19071 human-annotated questions from Wikipedia. Each sample consists of a question, a context, and related answers.