

M6: A Chinese Multimodal Pretrainer

Junyang Lin^{1*}, Rui Men^{1*}, An Yang^{1*}, Chang Zhou¹, Ming Ding², Yichang Zhang¹, Peng Wang¹, Ang Wang¹, Le Jiang¹, Xianyan Jia¹, Jie Zhang¹, Jianwei Zhang¹, Xu Zou², Zhikang Li¹, Xiaodong Deng¹, Jie Liu¹, Jinbao Xue¹, Huiling Zhou¹, Jianxin Ma¹, Jin Yu¹, Yong Li¹, Wei Lin¹, Jingren Zhou¹, Jie Tang^{2†}, Hongxia Yang^{1†}

¹Alibaba Group, China

²Tsinghua 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.dxd,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 pre-training 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 sparkles 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2021 Association for Computing Machinery.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

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 pre-training 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.

*Equal contribution.

†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 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 Wikipedia¹ 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[39] 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. [47] 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 models. 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 on 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 in Chinese. Specifically, while we unify our pretraining for both natural language and multimodalities, we construct large datasets of both plain texts and image-text pairs. We are interested in obtaining large-scale data that covers 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 the cross-modal pretraining. This raises the difficulty for the construction of a 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 a tremendous amount of text resources and images on the world wide web, the corpus for multimodal pretraining is assumed to be better when satisfying 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 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 multimodal pretraining tasks. Also, the data has a large coverage over domains, such as science, entertainment, sports, politics, common-sense 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.

¹<https://dumps.wikimedia.org/zhwiki/latest/>

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

Image	Source & Text
	<p><i>Source: Encyclopedia</i> 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.</p>
	<p><i>Source: Crawled Webpages</i> 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.</p>
	<p><i>Source: E-commerce</i> 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.</p>

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].

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: Statistics of the pretraining dataset. We demonstrate the sources of our data, and we calculate the number of images, tokens, and passages, as well as the size of images and texts.

Source (M)	Images(M)	Tokens (B)	Passages (M)	Image Size (TB)	Text Size (GB)
Encyclopedia	6.5	7.9	10.4	0.1	15.0
Webpages	46.0	9.1	106.0	1.5	70.0
E-commerce	8.0	0.5	8.5	0.3	12.2
Total	60.5	17.5	124.9	1.9	97.2

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.

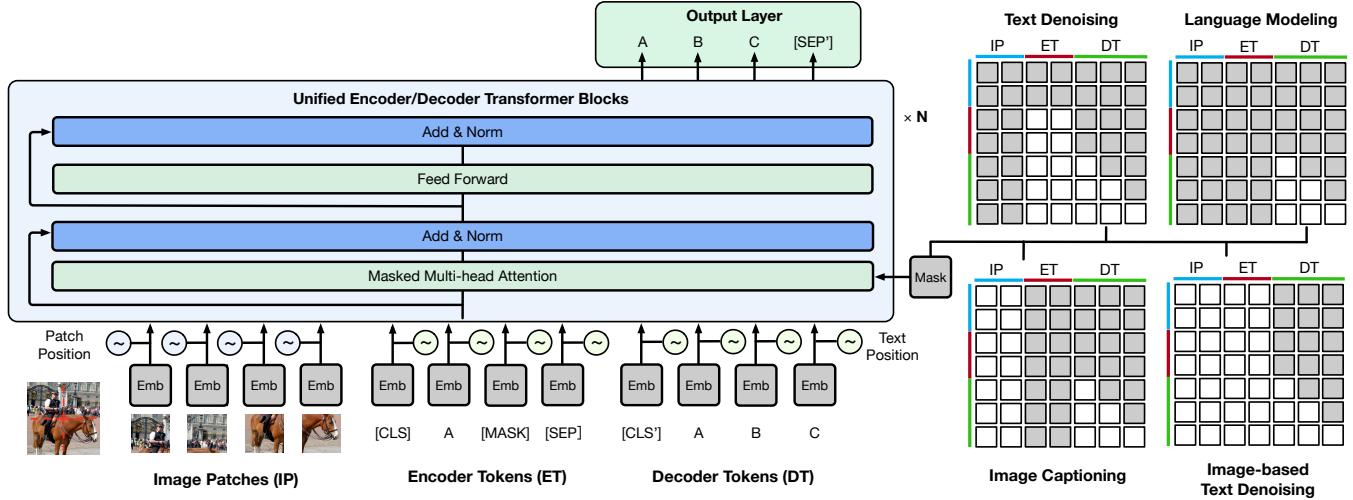


Figure 3: An overview of the pretraining tasks for M6. The design of masking strategies allows the learning of different tasks under the same framework. M6 is pretrained with image-based text denoising, image captioning, text denoising, and language modeling.

Table 4: 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. $n_{experts}$ is the number of experts. The M6-100B model employs multiple experts to scale up parameters to 100 billion. n_{param} is the number of all parameters.

Models	n_{layers}	d_{model}	n_{heads}	$n_{experts}$	n_{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.

3.4 Scaling up to 10 and 100 Billion Parameters

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.

To be more specific, we increase 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].

However, directly scaling up to M6-100B is much more difficult as there are more challenges for the computation resources. 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. We demonstrate the key statistics of the models of different scales in Table 4.

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:

$$p(x) = \frac{\exp(g(x)_i)}{\sum_j^N \exp(g(x)_j)}, \quad (1)$$

$$y = \sum_{i \in \mathcal{T}} p(x) E_i(x), \quad (2)$$

where $g(\cdot)$ refers to the sparse gating function, and \mathcal{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 TPUs 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 ($C = \frac{N \cdot c}{m}$, 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.² 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

²Note 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.

Table 5: 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
M6-10B	83.0	77.4	72.7	48.4	74.7

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. Compared with the images in the general domains, our collected data have the following features. The image and text are highly correlated as the text describes key features of the product, and there is no complex background in the images, which is easier to learn compared with the images in the public datasets such as MSCOCO [26].

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 pre-trained M6 to the VQA application.



Figure 4: Generated images for *sheep wool business casual suit* (绵羊毛商务休闲西服套装).

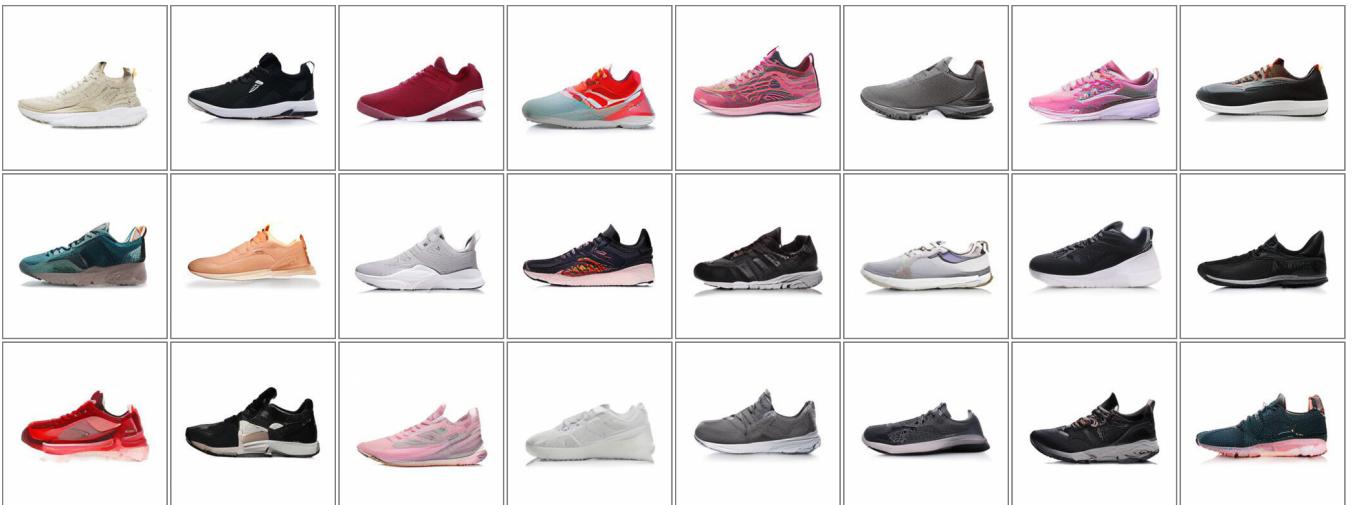


Figure 5: Generated images for *shock absorption and breathable running shoes* (减震透气跑鞋).

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* (军旅风迷彩高跟鞋).

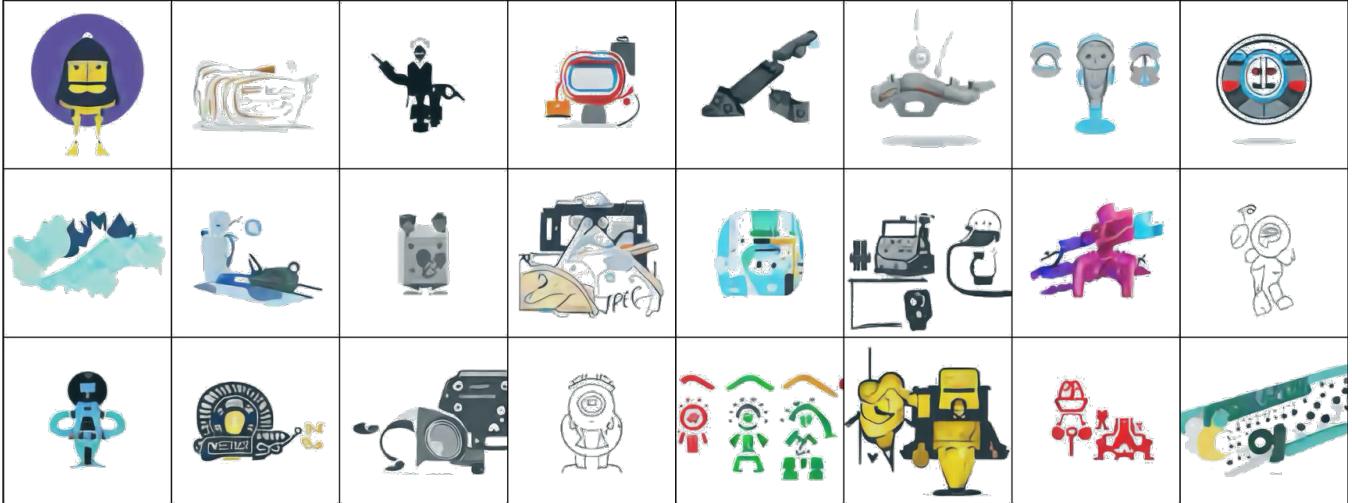


Figure 7: Generated images for *a clip art of robots* (机器人矢量插图)).

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

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

+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.



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 8: Several 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.

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.

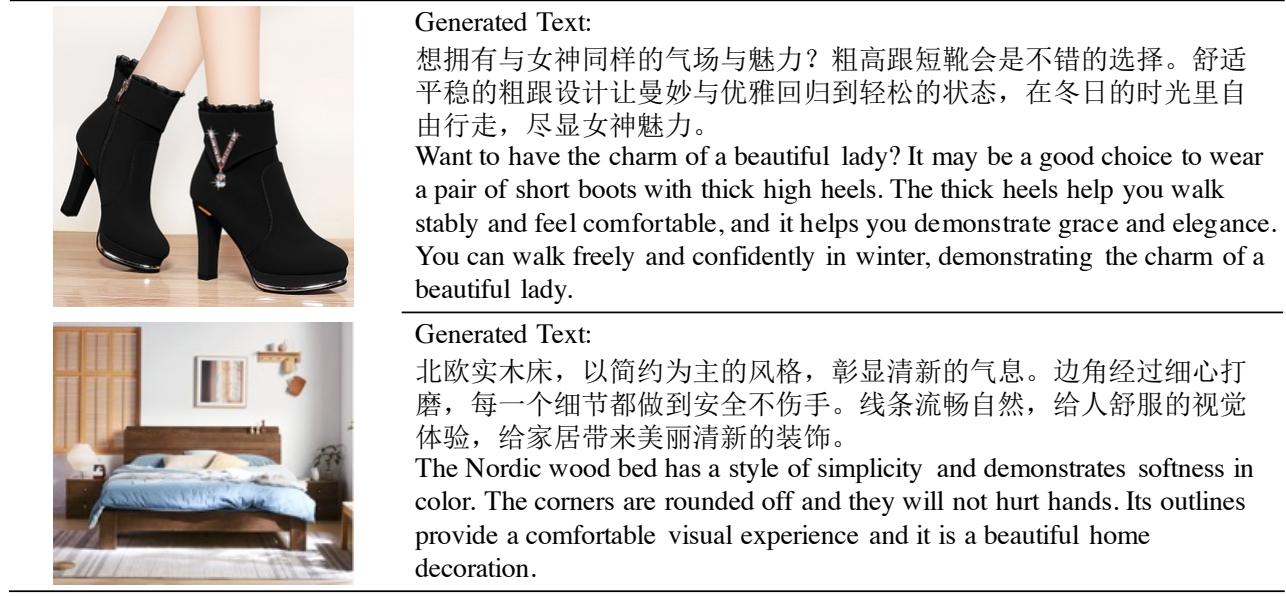


Figure 9: Two examples of image caption. We only use the image feature as input and provide no prompt.

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 10: One example of general question answering. The prompt which includes the question successfully triggers the model to generate long texts in the style of forum discussion.

Ancient Chinese poetry has various specific formats. We adopt the simplest constraints that

- The poem shall be consisted of at least 4 lines.
- The total number of lines shall be even.
- Each line must have exactly 5 or 7 words.
- All lines shall have the same number of words.

Text generation under format constraint is done in a search framework that we generate short sentences ending with punctuation until the number of words meets the constraint. We repeat this process until the model generates an "*<eos>*" token, or the number

of lines exceeds a limit of 16. Figure 11 illustrates an example of a generated poem.

4.6 Image-Text Matching

We evaluate the model's ability in cross-modal retrieval. Specifically, we construct a dataset (named E-Commerce ITM) containing pairs of texts and images from the mobile Taobao. Each pair belongs to a single item. 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

Prompt:
见情郎 作者: 唐 李白 正文:
Title: Dating Author: Bai Li (Tang Dynasty) Text:

Generated Text:
上马出东门，摧车欲问道。
却顾所来径，苍苍横翠微。
相去既远离，衣带日以疏。
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。
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. Despite prolonged separation, we don't have specific words when we finally meet each other, only discussing about everyday life. Oh! Where can I find such a lover, that we both grow old together?

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 discriminate 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 architecture [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 image 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 pretraining 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], LXBERT [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 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. 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 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.

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] 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.
- [7] 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.
- [8] 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).
- [9] Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming Transformers for High-Resolution Image Synthesis. *arXiv:2012.09841 [cs.CV]*
- [10] William Fedus, Barret Zoph, and Noam Shazeer. 2021. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *CoRR abs/2101.03961* (2021). <https://arxiv.org/abs/2101.03961>
- [11] 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*.
- [12] 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.
- [13] 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).
- [14] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *arXiv preprint arXiv:1406.2661* (2014).
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR 2016*. 770–778.
- [16] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654* (2020).
- [17] 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).
- [18] 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).
- [19] 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).
- [20] Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. 2020. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* (2020).
- [21] Mike Lewis, Shruti Bhosale, Tim Dettmers, Naman Goyal, and Luke Zettlemoyer. 2021. BASE Layers: Simplifying Training of Large, Sparse Models. *CoRR abs/2103.16716* (2021).
- [22] 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).
- [23] 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).
- [24] 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).
- [25] 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).
- [26] 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.
- [27] Yinhai 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).
- [28] 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.
- [29] 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).
- [30] 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).
- [31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. [n.d.]. Language models are unsupervised multitask learners. ([n. d.]).
- [32] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. *arXiv:2102.12092 [cs.CV]*
- [33] Shaogang 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.
- [34] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL 2016*.
- [35] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538* (2017).
- [36] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *arXiv preprint arXiv:1909.08053* (2019).
- [37] 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.
- [38] 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*.
- [39] Maosong Sun, Jingyang Li, Zhipeng Guo, Z Yu, Y Zheng, X Si, and Z Liu. 2016. Thuctc: an efficient chinese text classifier. *GithHub Repository* (2016).
- [40] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP 2019*. 5099–5110.
- [41] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural Discrete Representation Learning. In *NIPS*.
- [42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS 2017*. 5998–6008.
- [43] 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).
- [44] 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).
- [45] 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).
- [46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR 2017*. 1492–1500.
- [47] Liang Xu, Xuanwei Zhang, and Qianqian Dong. 2020. CLUECorpus2020: A Large-scale Chinese Corpus for Pre-trainingLanguage Model. *arXiv preprint arXiv:2003.01355* (2020).
- [48] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.
- [49] 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.
- [50] 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).
- [51] 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.