Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese

Abstract The tremendous success of vision-language foundation models has promoted the research and application of computer vision and multimodal representation leearning. However, it is still difficult to effectively transfer such foundation models to language-specific scenarios. In this work, we propose Chinese CLIP with the two-stage pretraining method which trains the model with locked-image tuning in the first stage and contrastive tuning in the second one. Specifically, we have developed 5 Chinese CLIP models of multiple sizes, spanning from 77 to 958 million parameters, and we have pretrained them on a collected largescale dataset of Chinese image-text pairs. Our comprehensive experiments demonstrate that Chinese CLIP can achieve the state-of-the-art performance on MUGE, Flickr30K-CN, and COCO-CN in the setups of zero-shot learning and finetuning, and it is able to achieve competitive performance in zero-shot image classification based on the evaluation on the ELEVATER benchmark. We have released our codes, models, and demos.

1 Introduction Starting from the burst of pretraining in NLP, foundation models have attracted attention from multiple research communities. Foundation models that learn from large-scale unsupervised or weakly supervised data play as the basis of downstream models. A milestone of foundation models (Bommasani et al., 2021) in multimodal representation learning is CLIP (Radford et al., 2021). Different from the conventional generative pretraining, CLIP is a contrastive-learning-based model pretrained on a large-scale dataset of around 400 million image-text pair data collected from the web. Despite the simplicity of the method, CLIP not only achieved outstanding performance in visionlanguage retrieval but more importantly played as a vision foundation model and demonstrated state-ofthe-art performance in zero-shot image classification across a series of datasets. CLIP which builds a connection between vision and language has been transforming the research in both multimodal representation learning and computer vision. Be that as it may, it is difficult to efficiently transfer a cross-modal pretrained model to another language for several causes. First, learning to model the distribution of language-native vision and language data is significant for the transfer. Though CLIP performs as a strong foundation model in most scenarios, we find that it is hard for CLIP with machine translation to perform well on the Chinese-native cross-modal retrieval benchmark.

Figure 1 demonstrates large performance gaps between the original CLIP and our Chinese CLIP at all model scales. We assume that it is crucial for both encoders to learn from the language-native images and texts. Second, the performance of previous methods for Chinese multimodal pretraining has been inhibited by several factors. Pretraining from scratch requires collecting a large-scale quality language-specific image text pair dataset similar to Web Image Text (WIT) for OpenAI CLIP (Fei et al., 2021; Xie et al., 2022). Though the fast transfer of CLIP to Chinese data can be realized by using CLIP initialization and Locked-Image Tuning (Zhai et al., 2022), the vision encoder still cannot learn the information of images from the language-specific domains (Gu et al., 2022). Therefore, we propose Chinese CLIP, a language-specific vision-language foundation model pretrained on the publicly available Chinese image-text pair data. Additionally, we still use the same architecture as OpenAI CLIP. To realize efficient transfer of cross-modal foundation model to Chinese data, we develop a two-stage pretraining method, which is also adaptive to other vision-language foundation models, e.g., ALIGN, Florence, etc. Here in this work, we use CLIP as an example. To be specific, we first initialize both encoders with pretrained models, namely vision encoders from CLIP and text encoders from RoBERTa-wwm-Chinese (Cui et al., 2020). In Stage 1, we freeze the image encoder and only optimize the text encoder with LiT, and in Stage 2, we train both encoders with contrastive tuning 2. In this way, the new model can inherit from the foundation models through initialization and LiT, and effectively transfer to language-specific data through contrastive tuning. We evaluate Chinese CLIP on 3 Chinese cross-modal retrieval datasets, including MUGE2 , Flickr30K-CN (Lan et al., 2017), and COCOCN (Li et al., 2019c). Experimental results demonstrate that both the large-size and huge-size Chinese CLIP reach state-of-the-art performance on the 3 datasets in the setups of both zero-shot learning and finetuning. Additionally, we evaluate the capability of zero-shot image classification on the track “Image Classification in the Wild” of the ELEVATER benchmark (Li et al., 2022b). On the classification datasets, Chinese CLIP demonstrates competitive performance in comparison with state-of-the-artmethods and outperforms the Chinese baselines. Furthermore, we provide NVIDIA TensorRT and ONNX models for deployments, which run around 2 to 10 times faster than Pytorch models for inference. In brief, our contributions are: • We propose Chinese CLIP, a simple implementation of CLIP pretrained on our collected large-scale Chinese image-text pair data, and we propose a two-stage pretraining method to achieve high pretraining efficiency and improved downstream performance. • Chinese CLIP achieves state-of-the-art performance in cross-modal retrieval in the setups of zero-shot learning and finetuning, and competitive performance in zero-shot image classification. 2 Method CLIP (Radford et al., 2021) based on simple visionlanguage contrastive pretraining on large-scale weakly supervised data is a significant foundation model in multimodal representation learning. It can transfer to cross-modal retrieval directly, and its image encoder can play as a vision backbone. In this work, we propose to build a language-specific CLIP model by pretraining a vision-language model on large-scale Chinese multimodal data. In the following, we provide the details of method design and implementation of our Chinese CLIP.

2.1 Data One key to CLIP’s success should be the large-scale dataset for pretraining. Based on the experiments of a CLIP reimplementation (Ilharco et al., 2021), scaling up data and lengthening the training process can consistently improve the model performance in zero-shot learning. This year, the most recent multimodal pretrained models Wukong (Gu et al., 2022) and R2D2 (Xie et al., 2022) were pretrained on a public dataset of 100 million image-text pairs and an in-house dataset of 250 million samples, where only a subset of 23 million samples were released. For the facility in reimplementation, we aim at pretraining Chinese CLIP on as many publicly available data as possible, and thus we focus on collecting high-quality public datasets. We extract the Chinese data (with the mark “zh”) from the latest LAION-5B (Schuhmann et al., 2021), and collect the data from the Wukong dataset. However, due to the problems of unavailable links, we can only collect around 108 million samples and 72 million samples from LAION-5B and Wukong respectively. We additionally add the translated data from the classic English multimodal datasets, including Visual Genome (Krishna et al., 2017) and MSCOCO (Chen et al., 2015), where test sets are removed. Finally, we construct a dataset for Chinese multimodal pretraining with around 200 million image-text pairs.3 Below illustrates the procedure for data preprocessing. For the part of data from LAION-5B, we remove the samples with CLIP scores lower than 0.26 computed by mCLIP (Carlsson et al., 2022). Besides, we remove the samples with captions containing words in our internal blacklist. The blacklist contains words related to advertising, image filenames, etc. We remove those samples that are too short (fewer than 5 characters) or too long (more than 50 characters). For the images, we resize them to the resolution of 224 × 224 for most cases and 336 × 336 for the ViT-L/14@336px. 2.2 Pretraining Method There are multiple design choices for pretraining the Chinese CLIP models. One of the simplest methods should be pretraining from scratch, where both the image and text encoders are randomly initialized. However, we assume that its performance will be limited by the quantity and quality of the current pretraining data. To leverage the advantages of existent pretrained models, we initialize the models with weights from the pretrained checkpoints from the official release of CLIP 4 for the image encoder, and RoBERTa-wwm-ext and RBT3 5 for the text encoder. To adapt the model to the introduced pretraining data, it is available to pretrain it with “contrastive tuning”, similar to the way to transfer CLIP to downstream retrieval data. In comparison with contrastive tuning, Locked-image Tuning (LiT) (Zhai et al., 2022) demonstrated improved performance in downstream transfer. In this work, we propose a two-stage pretraining method, as shown in Figure 2. The core idea is to first utilize LiT to enable the text encoder to read out high-quality representations from the foundation vision model from OpenAI CLIP, and then transfer the whole model to the domain of the introduced pretraining data. It is not sufficient to pretrain Chinese CLIP with solely LiT, as the image encoder should learn the information of the images of the Chinese datasets and model the distribution of such data. Before the two-stage pretraining, we first initialize both encoders with pretrained models. In Stage 1, we “lock” the image encoder by freezing its parameters during pretraining. We only pretrain the text encoder for vision-language alignment, based on the assumption that the vision backbone with pretrained weights is already a powerful vision foundation model (Zhai et al., 2022; Gu et al., 2022). We pretrain it until there is no salient performance improvement in downstream tasks, even if we prolong the pretraining progress. Then we switch to Stage 2, where we “unlock” the image encoder by enabling its optimization. In Stage 2, we continue pretraining without any parameter frozen, so that the image encoder can learn to model the distribution of the image data from Chinese websites. In the ablation study, we discuss the influence of the initialization of the pretrained checkpoints and pretraining methods on the downstream performance. Experimental results show that the two-stage pretraining method can outperform either pretraining from scratch or directly finetuning from the pretrained models.

3 Evaluation To comprehensively probe the effects of Chinese CLIP, we follow the conventional practice that we first evaluate its basic capabilities of cross-modal retrieval, i.e. text-to-image retrieval and imageto-text retrieval, in different domains, including e-commerce and the general domain. Additionally, as the contrastive-learning-based pretraining builds a foundation vision model that is semantically connected with natural language, we follow Radford et al. (2021) and evaluate its capabilities of zero-shot classification. Specifically, we validate Chinese CLIP on the classification datasets of the ELEVATER benchmark (Li et al., 2022b),which is known as “Image Classification in the Wild (ICinW)”. 3.1 Cross-modal Retrieval 3.1.1 Datasets and Metrics We validate Chinese CLIP on 3 cross-modal retrieval datasets, namely MUGE-Retrieval, Flickr30K-CN (Lan et al., 2017), and COCOCN (Li et al., 2019c). MUGE-Retrieval is an image-text retrieval dataset, where data are extracted from Chinese E-commerce websites. Flickr30K-CN and COCO-CN are built from the classical datasets Flickr30K and MSCOCO-1K whose texts are translated into Chinese. Ourevaluation includes setups of zero-shot learning and finetuning. For zero-shot learning, we use Chinese CLIP models to compute the similarity scores between images and texts and return the top-K most similar candidates. For finetuning, we finetune the Chinese CLIP models for cross-modal retrieval with contrastive tuning. The evaluation is the same as that in zero-shot learning. The evaluation metrics are Recall@K, where K = {1, 5, 10}, and Mean Recall (MR, i.e., the average of Recall@K). For comparison, we choose the base-size and large-size Wukong and R2D2 as the baselines, which are the previous SOTA models in Chinese multimodal representation learning. Following these baselines, we report validation performance on MUGE and test performance on Flickr30K-CN and COCO-CN. Note that in the setup of finetuning, R2D26 is essentially an end-to-end model of retrieval and ranking. 3.1.2 Results Table 1 reports the model performance on MUGE-Retrieval. For the base-size model, CN-CLIPViT-B/16 outperforms the baselines on all metrics and in both setups of zero-shot learning and finetuning. Specifically, for the base-size models, CN-CLIPViT-B/16 surpasses WukongViT-B/32 by 17.0 MR in zero-shot learning and surpasses

R2D2ViT-B by 8.7 MR in finetuning. Besides, the tiny model CN-CLIPRN50 can outperform the basesize WukongViT-B/32 by 8.9 MR in zero-shot learning and 8.0 MR in finetuning. For the large-size models, CN-CLIPViT-L/14 can outperform both baselines in all metrics and CN-CLIPViT-L/14@336px pretrained on images of a larger resolution can achieve the state-of-theart performance. CN-CLIPViT-L/14@336px outperforms R2D2ViT-L/14 by 6.6 MR in zero-shot learning and 3.8 MR in finetuning. When scaling to CN-CLIPViT-H/14, the performance is further improved. Compared with the best large-size model CN-CLIPViT-L/14@336px, CN-CLIPViT-H/14 surpasses it by 2.7 MR in zero-shot learning and 2.3 MR in finetuning. Table 2 and 3 report the model performance on Flickr30K-CN and COCO-CN. We focus on the evaluation of R@1. In both datasets, CN-CLIP achieves better performance than the baselines. For the base-size models, in the setup of zero-shot learning of Flickr30K-CN, CN-CLIPViT-B/16 surpasses WukongViT-B/32 by 17.0 R@1 in text-to-image retrieval and 8.4 R@1 in image-to-text retrieval, and in the finetuning setup, CN-CLIPViT-B/16 surpasses R2D2ViT-B by 0.8 R@1 in image retrieval and 0.9 R@1 in text retrieval. Similarly, in the finetuning setup of COCOCN, CN-CLIPViT-B/16 surpasses R2D2ViT-B by 1.9 R@1 in image retrieval and 1.3 R@1 in text retrieval. For the tiny-size CN-CLIPRN50, itagain achieves or surpasses the performance of WukongViT-B/32 in several metrics of Flickr30KCN and COCO-CN. Specifically, CN-CLIPRN50 surpasses WukongViT-B/32 by 3.1 R@1 in the zeroshot learning of Flickr30K-CN image retrieval and by 2.6 R@1 in the finetuning of COCO-CN text retrieval. For the large-size models, in the zero-shot setup of Flickr30K-CN, CN-CLIPViT-L/14 surpasses WukongViT-L/14 by 16.3 R@1 in text-toimage retrieval and 4.1 R@1 in image-to-text retrieval. CN-CLIPViT-L/14@336px further improves over CN-CLIPViT-L/14 by 1.0 R@1 in image retrieval and 3.1 R@1 in text retrieval. In the finetuning setup, CN-CLIPViT-L/14 surpasses R2D2ViT-L/14 by 0.5 R@1 in text retrieval. CN-CLIPViT-L/14@336px achieves equal performance with R2D2ViT-L/14 in image retrieval and surpasses it by 1.0 R@1 in text retrieval. Similarly, in the finetuning setup of COCO-CN, CN-CLIPViT-L/14 surpasses R2D2ViT-L/14 by 0.9 R@1 in text retrieval. CN-CLIPViT-L/14@336px further surpasses R2D2ViT-L/14 by 1.0 in image retrieval and 1.9 R@1 in text retrieval. On Flickr30K-CN and COCO-CN, scaling from CN-CLIPViT-L/14 to CN-CLIPViT-H/14 improves the performance in almost all the metrics. Specifically, in the zero-shot setup of Flickr30K-CN, CN-CLIPViT-H/14 surpasses CN-CLIPViT-L/14 by 3.2 R@1 in image retrieval and 1.4 R@1 in text retrieval. Moreover, in the finetuning setup of COCO-CN, CN-CLIPViT-H/14 even surpasses CN-CLIPViT-L/14@336px with larger image resolution by 1.4 R@1 in image retrieval and 2.3 R@1 in text retrieval. We also compare our CN-CLIPViT-H/14 with a huge CLIP-like model, T-Bletchley7 . This model has 2.5 billion parameters and is pretrained on billions of multilingual image-caption pairs. In the finetuning setup of COCO-CN, with smaller sizes of model parameters and pretrain dataset, CN-CLIPViT-H/14 still surpasses T-Bletchley by 3.9 MR. 3.1.3 Ablation Study Here we provide an ablation study on our proposed two-stage training methods. To validate its significance and effectiveness, we design several setups for the ablation study. Our experiments are conducted on CN-CLIPViT-B/16. To evaluate the。influence of initialization, we pretrain a model from scratch, and to examine the influence of LiT, we pretrain a model without freezing the image encoder. For a better demonstration, we report the curves of model performance of zero-shot retrieval on different datasets in terms of pretraining progress, indicated by the number of processed samples. Figure 3 shows the performance on different tasks, namely MUGE text-to-image retrieval, textto-image and image-to-text retrieval on Flicrk30KCN and COCO-CN. In comparison with pretraining with pretrained model initialization, pretraining from scratch performs much worse though it shows consistent performance improvement in terms of pretraining progress. As to the importance of LiT, we observe different phenomena on different datasets. On MUGE, a dataset of samples originally collected from Chinese websites, we find that pretraining without LiT might be the best solution though its performance gap with two-stage pretraining is quite small. However, on the other datasets, i.e., Flickr30K-CN and COCO-CN, whose samples are translated from the English datasets, we find that our two-stage pretraining performs significantly better than pretraining without LiT. Furthermore, we observe a common phenomenon that in the two-stage pretraining, switching from Stage 1 to Stage 2 can effectively boost the model performance to a higher level. This reflects the importance of adapting the pretrained model to the data distribution of the Chinese multimodal data, especially those concerned with visual information.

3.2 Zero-shot Image Classification 3.2.1 Open-Domain Image Classification Benchmark in Chinese Contrastive pretraining on image-text pairs builds a connection between vision and natural language. Natural language supervision instead of crowdsourced labeling endows models with the capability of zero-shot image classification by computing similarities between the given image and the text descriptions of the labels in the candidate set. Recent progress in this field is the ELEVATER benchmark (Li et al., 2022b). The track ICinW for open-domain image classification consists of a series of image classification datasets, including ImageNet (Deng et al., 2009), CIFAR (Krizhevsky et al., 2009), MNIST (Deng, 2012), etc. In order to evaluate Chinese CLIP on the datasets, we first transform the datasets for Chinese models by translating labels and prompts into Chinese. 3.2.2 Experimental Results Table 4 reports the performance of both English models and Chinese models. The baselines pretrained on English data include DeCLIP (Li et al., 2021b), ALIGN (Jia et al., 2021), CLIP (Radford et al., 2021), and OpenCLIP (Ilharco et al., 2021), and the baselines pretrained on Chinese data include BriVL (Fei et al., 2021) and Wukong (Gu et al., 2022). We report the results of the variant with the best downstream performance for the models. We first focus on the comparison with the Chinese baselines. To be specific, on all datasets including ImageNet classification, Chinese CLIP surpasses both baselines significantly, and the relative achievements on some datasets are over 100%. Besides, we also compare Chinese CLIP with the foundation models, e.g., CLIP and ALIGN, which are pretrained on English data. It can be found that Chinese CLIP outperforms CLIP or ALIGN on CIFAR-10, CIFAR-100, FER-2013, KITTIDistance, MNIST, PatchCamelyon, and PascalVOC-2007. Also, on the classification datasets for general concepts or objects common in both western and eastern culture, Chinese CLIP consistently achieves better performance. This indicates that Chinese CLIP is capable of categorizing images to general prototypes. However, as to the classification concerned with proper nouns, e.g., FGVC-Aircraft, it is difficult for all models to achieve high accuracy. We assume that the related images and texts are not common in the pretraining datasets, and it is also hard for the models to understand the names of airplanes without finetuning. Specifically, for Chinese models, the translation or even transliteration can significantly affect the performance of Chinese CLIP. It encourages building a benchmark of “Image Classification in the Wild for Chinese Models”. 3.2.3 Analysis Sensitivity to Handcrafted Prompts While the benchmark ELEVATER provides specific prompts for each dataset, we find that this is not always the best option, in comparison with our baseline, translation of the prompts provided by OpenAI CLIP. The baseline with around 90 prompts performs the best on average. However, for some datasets, specific prompts designed with human knowledge can boost the performance significantly. A typical case is the classification of airplanes. We test CN-CLIPViT-L/14 with our specified prompts that are related to the knowledge of aircraft, e.g.,“label, a photo of an airplane", “label, a zoomed image of a fighter”, etc., and the translation of the OpenAI prompts. Experimental results show that the model can achieve an accuracy of 16.0 with the specified prompts but only 13.8 with the OpenAI prompts. Inability to Understand Negation Previous studies (Khandelwal and Sawant, 2019; Hosseini et al., 2021) demonstrate that even strong NLP pretrained models often make mistakes in negation problems. We explore CLIP’s capability to understand negation by conducting experiments on KITTI-Distance (Fritsch et al., 2013) and PatchCamelyon (Veeling et al., 2018). KITTI-Distance provides 4 options for models to judge, including “next to a car”, “near a car”, “at a distance away from a car”, and “no car”. The last one is concerned with negation. We compare the model performance using the text “no cars” and “others” for the last label. We observe that it is hard for the model to understand negation. By changing the label from “others” to “no cars”, the performance drops by 48.1% in accuracy (49.9 vs. 25.9). Similarly, in the experiments on PatchCamelyon, the performance drops from 63.5 to 50.2 by changing labels from “mainly red” and “green block in the middle” to “no green block in the middle” and “green block in the middle”. This shows the limitation of the training of CLIP in learning negation. The texts in the pretraining datasets are mostly descriptions of the images, which indicate their objects or features but often do not indicate the absence of objects. 3.3 Deployment For deployment, we develop ONNX-based and TensorRT-based models based on our Pytorchbased pretrained Chinese CLIP models. As expected, we observe that the inference efficiency increases significantly while there is almost no performance sacrifice. Specifically, the inference efficiency of TensorRT-based models is around 2 to 10 times faster than the Pytorch-based models. More statistics are listed in Appendix A.6. 4 Related Work Previous vision-language pretrained models are mostly BERT/T5-style (Devlin et al., 2019; Raffel et al., 2020), which involves cross-modal fusion (Chen et al., 2020; Li et al., 2019a,b; Lu et al., 2019; Lin et al., 2020; Li et al., 2020; Huang et al., 2020; Xu et al., 2021; Zhang et al., 2021; Shen et al., 2021; Wang et al., 2021b, 2022a; Li et al., 2021a, 2022c,a; Wang et al., 2021a, 2022b). CLIP (Radford et al., 2021), instead, is a contrastive-learning-based two-tower model, which can serve as a vision foundation model. Following CLIP, a series of similar contrastivelearning-based multimodal pretrained models were proposed and reached new SOTAs in cross-modal retrieval and zero-shot classification (Jia et al., 2021; Yao et al., 2021; Yuan et al., 2021). Furthermore, CLIP can be adaptive to other models. A typical case is that CLIP is essential to many image generation models, e.g., DALL-E (Ramesh et al., 2021), DALL-E 2 (Ramesh et al., 2022), Stable Diffusion (Rombach et al., 2022), etc. The success of multimodal pretraining encouraged the transfer of the existing methods to Chinese pretraining, including generative pretrained models (Lin et al., 2021a; Fei et al., 2021; Yang et al., 2021; Lin et al., 2021b; Wang et al., 2022a) and contrastive pretrained models (Fei et al., 2021; Gu et al., 2022; Xie et al., 2022; Chen et al., 2022b). 5 Conclusion In this work, we propose Chinese CLIP, a Chinesespecific vision-language foundation model. Specifically, we construct a pretraining dataset of around 200 million samples, and pretrain a series of Chinese CLIP models with the proposed two-stage pretraining method, which improves both pretraining efficiency and effectiveness. Our comprehensive evaluation shows that Chinese CLIP can reach stateof-the-art performance on multiple cross-modal retrieval datasets in zero-shot learning and finetuning. Furthermore, we demonstrate that Chinese CLIP models can also achieve competitive performance in zero-shot image classification across 10 datasets. Limitations A number of issues reflect the limitations of this work but also point out some directions for our future research. In this section, we generally discuss some limitations about the scale of data and model. Data The core of CLIP pretraining is the simple but effective large-scale contrastive pretraining on extremely large-scale data. Though we have utilized around 200 million samples, compared with recent studies (Yuan et al., 2021; Chen et al., 2022a) the scale of our pretraining data is relatively small. Thus one of our next-step studies is scaling up the quantity of the pretraining data to evaluate the performance improvement with data scaling. Furthermore, we still find it hard to decide what a “high-quality” dataset for CLIP is. In the previous studies (Jia et al., 2021; Li et al., 2021b), the preprocessing methods are mostly simple to avoid the loss of data. However, there are still many samples where the image and text are not matched properly, which may provide negative information to the pretraining. In our future research, we plan to use the pretrained Chinese CLIP model to compute a score for each image-text pair in a larger dataset, filter those whose scores are under the specified threshold, and pretrain the new models with the new data. This is one of the possible solutions to explore the relationship between data quality and pretraining effectiveness. Also, such cycling might bring continuous performance enhancement in downstream tasks. Model Recently we have witnessed that in many domains the scaling of model size can lead to consistent performance improvement (Gordon et al., 2021; Wei et al., 2022), and in this work, we also find that the scaling of model size for Chinese CLIP can achieve steady performance improvement in different downstream tasks, including retrieval and classification. Recent studies have scaled ViT and also CLIP-like models to a much larger scale than our largest CN-CLIPViT-H/14, e.g., the 3B Swinv2 (Liu et al., 2022), the 4B ViT-e (Chen et al., 2022a), etc. In the future, we will continue exploring scaling up models in line with scaling up data in order to build a more effective Chinese CLIP. Another issue of model scaling connected with the real-world application is how to build effective small models. Experimental results show that our smallest Chinese CLIP CN-CLIPRN50 performs much worse than the ViT variants. However, in real-world applications, effective small models that are available for deployment are usually more welcomed. Thus it is necessary to explore distillation for CLIP so that the capability of large models can be transferred to small models for application. Ethics Statement The proposed model is a contrastive-learning-based vision-language foundation model, which generates features for images and texts. Those features can be representation of visual and linguistic information, and they can support applications such as search engine, recommender system, etc. Besides, this model can play as a foundation model support recent image generation models, e.g., diffusion models (Ramesh et al., 2022). This may create risks as the AI generated contents may reflect harmful information, such as hate, bias, pornography, etc. In most cases, these cases should be attributed to the training of the image generation models. Still, we cannot avoid the negative effects from the CLIP representations to the generation. In the future, we will study how to filter pretraining data to avoid the potential risks.

Appendix

A.1 Model Architecture Details We develop 5 Chinese CLIP models of different sizes, spanning from around 77 to 958 million parameters. We include 1 ResNet50 model CN-CLIPRN50 and 4 ViT models, i.e., CN-CLIPViT-B/16, CN-CLIPViT-L/14, CN-CLIPViT-L/14@336px and CN-CLIPViT-H/14, where models are pretrained on images of the resolution of 224 × 224 without specification. Table 5 presents the details of the model architecture. The smallest model CN-CLIPRN50 consists of a ResNet-50 for the image encoder and a RBT3 for the text encoder. The base-size model CN-CLIPViT-B/16 consists of a ViT-B/16@224px for the image encoder and a RoBERTa-wwm-Base for the text encoder. The large-size model CN-CLIPViT-L/14 consists of a ViT-L/14@224px for the image encoder and a RoBERTa-wwm-Base for the text encoder, while CN-CLIPViT-L/14@336px consists of a ViT-L/14@336px and a RoBERTa-wwm-Base. Specifically, we pretrain CN-CLIPViT-L/14@336px by continuing pretraining on the pretrained CN-CLIPViT-L/14. For the adaptation to a larger resolution, we initialize the image positional embedding by applying interpolation to the positional embedding of CN-CLIPViT-L/14, following Ilharco et al. (2021). The huge-size model CN-CLIPViT-H/14 consists of a ViT-H/14 for the image encoder and RoBERTa-wwm-Large for the text encoder. More implementation details are presented in Appendix A. We provide more details of their model architectures in Table 6 and Table 7. We keep the architecture of ResNet-50, ViT-B/16, and ViT-L/14 backbones conformed with OpenAI CLIP and the architecture of ViT-H/14 same with LAION CLIP8 . This enables us to initialize the Chinese CLIP image encoders with their weights. The text encoders are Chinese Roberta models (Cui et al., 2020). Specifically, our most lightweight tiny-size Chinese CLIP uses the architecture of the 3-layer RBT3 model. The base-size and large-size Chinese CLIP models use the 12-layer architecture of RoBERTa-wwmBase. For the huge-size CN-CLIP, we use the 24- layer architecture of RoBERTa-wwm-Large. The vocabulary size of text tokenizer is 21, 128.A.2 Pretraining Details Initialization As mentioned in Section 2.2, we initialize the image encoders of CN-CLIPRN50, CN-CLIPViT-B/16 and CN-CLIPViT-L/14 using the OpenAI CLIP weights. The image encoder of CN-CLIPViT-H/14 is initialized with LAION CLIP. Besides the ResNet or ViT parameters, the temperature and visual output projection parameters are also initialized with the pretrained CLIP weights. For the text encoder, we initialize the parameters using the released Chinese Roberta weights of the corresponding model scale, with their pooler weights discarded. The text output projection weight is randomly initialized with normal distribution. Stage 1 The pretraining hyperparameters of Stage 1 are shown in Table 8, which are shared for CN-CLIPRN50, CN-CLIPViT-B/16, CN-CLIPViT-L/14 and CN-CLIPViT-H/14. The values of hyperparameters are generally similar to those in OpenAI CLIP (Radford et al., 2021). As to data augmentation, we use random resize cropping and AutoAugment (Cubuk et al., 2019) on input images. We leverage all-gather communications across GPU workers to compute contrastive loss on the global batch. The above 4 models are pretrained for around 20, 44, 64, and 26 epochs in this stage respectively, with the image encoder frozen. The running variance and mean of batch normalization layers are not updated in this stage for CN-CLIPRN50. The optimal epochs of pretraining are determined by measuring the mean-recall under the 3 downstream zero-shot retrieval tasks during training. Mixed-precision training is activated. In this stage, we pretrain 1.6 days using 64 NVIDIA V100 GPUs for CN-CLIPRN50, 4.5 days using 128 NVIDIA V100 GPUs for CN-CLIPViT-B/16, 11.5 days using 128 NVIDIA V100 GPUs for CN-CLIPViT-L/14 and 3.8 days using 184 NVIDIA A100 GPUs for CN-CLIPViT-H/14. Stage 2 In Stage 2, we unfreeze the image encoder and update all the model parameters. Except for the peak learning rate, batch size and training epochs, all other hyperparameters mentioned in Stage 1 are kept unchanged. We decrease the learning rate to 2e − 5 for subtler optimization. For CN-CLIPRN50, CN-CLIPViT-B/16 and CN-CLIPViT-L/14, the batch size is shrunk to 16, 384, 16, 384 and 4, 608 respectively due to the limitation in GPU memory. When scaling to CN-CLIPViT-H/14, we implement gradient checkpointing, which enables a larger batch size of 32, 768. These 4 models are pretrained for around 44, 15, 7 and 7 epochs in Stage 2, respectively. In this stage, we pretrain CN-CLIPRN50 for 5.8 days using 64 NVIDIA V100 GPUs, CN-CLIPViT-B/16 for 3.0 days using 128 NVIDIA V100 GPUs, CN-CLIPViT-L/14 for 8.0 days using 128 Nvidia V100 GPUs, and CN-CLIPViT-H/14 for 2.2 days using 184 NVIDIA A100 GPUs. To pretrain a model of a larger resolution, we implement interpolation to the image positional embedding of CN-CLIPViT-L/14 for adapting to a larger resolution and continue pretraining with images of the resolution of 336 × 336. We start from CN-CLIPViT-L/14 and continue pretraining by 2 epochs. The pretraining only costs the use of 128 NVIDIA A100 GPUs for 0.7 days.

A.3 Finetuning Details As reported in Table 1, 2 and 3, we mainly finetune CN-CLIP on 3 cross-modal retrieval datasets: MUGE, Flickr30K-CN, and COCO-CN. Most finetuning experiments are conducted on 32 NVIDIA A100 GPUs. The finetuning strategy and loss are consistent with the pretraining process. For time efficiency and full utilization of computation resources, we set the batch size as large as possible. We implement gradient checkpointing in the finetuning process of CN-CLIPViT-L/14@336px and CN-CLIPViT-H/14 for a larger batch size. Table 9 shows the specific settings of batch size, peaking learning rate, maximum epochs, and warmup iterations in the finetuning process. We set other hyperparameters to be the same as those in pretraining by default. We save the model parameters at the end of each epoch. For MUGE, we report the best results on the validation set. For Flickr30K-CN and COCO-CN, we choose the checkpoint with the best performance on the validation set and report the results on the test set. A.4 Cross-modal Retrieval with Longer Texts The results reported in Section 3.1.2 demonstrate the excellent cross-modal retrieval capability of Chinese CLIP. Note that the average text lengths of MUGE, Flickr30K-CN, and COCO-CN are 7.4, 19.7, and 16.8, respectively. We also conduct finetuning experiments on the ICR (Xie et al., 2022) dataset with an average text length of 45.3. Experimental results are shown in Table 10. Sincthe texts in the ICR dataset are longer, we set the maximum text length to 128 for finetuning. The results show that Chinese CLIP achieves state-ofthe-art performance in cross-modal retrieval tasks with longer texts. A.5 Details About Experiments on Zero-shot Image Classification We present the data statistics and metrics of the 20 image classification datasets of the track ICinW in the ELEVATER benchmark in Table 11. For the adaptation of Chinese CLIP to the English-native benchmark, we apply a series of preprocessing strategies. Specifically, we translate the text descriptions of the labels and the templates for manual prompts to Chinese. For example, the labels in CIFAR-10 include “car, dog, ...”, and we manually translate the words into Chinese. There are also particular cases, such as the labels in FGVCAircraft (Maji et al., 2013), which are difficult to translate or transliterate. We search the names on Google and figure out the best Chinese name for each label. Be that as it may, we cannot guarantee that we have the best Chinese translation, and more importantly, it is still hard for the Chinese pretrained model to understand some of the concepts, which may lead to unsatisfactory performance in the related downstream tasks. As to the templates, for some datasets, we use our translation of the templates provided by the ELEVATER toolkit,9 and for the others, we use the translation of the templates from OpenAI CLIP. We present the experimental results of all Chinese CLIP models on zero-shot image classification in Table 12. It can be found that the scaling of model size can consistently bring improvements in model performance. The predictable improvements of scaling Chinese CLIP indicate that we can further scale up the model for better performance in the future work. However, it is still a pity that the tiny-size CN-CLIPRN50 saliently performs much worse than the ViT variants which are significantly larger. This shows that there is still much roomfor the small model to improve, and the knowledge transfer of CLIP from large models to small models should be an important research topic in multimodal representation learning. A.6 Deployment Chinese CLIP is supported to be deployed into ONNX-based10 and TensorRT-based11 models, enabling faster text and vision representation generation (especially for online inference). In this section, we provide more details on the model conversion, as well as the performance improvement. Specifically, we employ the ONNX module in PyTorch with ONNXMLTOOLS12 package to convert Chinese CLIP PyTorch models to ONNXbased models in FP16 precision. With the support of ONNXRUNTIME-GPU13 package, the ONNXbased models are able to infer on NVIDIA GPUs. The TENSORRT package enables the TensorRTbased models obtained from ONNX-based models and provides the GPU inference runtime. Our TensorRT-based models are also in FP16 precision. We benchmark the PyTorch implemented Chinese CLIP models with converted ONNX-based and TensorRT-based models using a server with a single NVIDIA T4 GPU. The server contains 16 Intel Xeon (Skylake) Platinum 8163 CPU cores with 64GB memory. For each model, we inference the vision and text representations for 100 batches and compute the average time. Simulating the scenario of online deployment, we use batch size of 1. All the models infer with FP16 precision. Table 13 shows the comparisons of inference time cost. For almost all the model scales, ONNX-based and TensorRT-based models have optimized inference speed over native PyTorch implemented Chinese CLIP models, especially on smaller model sizes. For vision representation inference, the TensorRT-based models are around 1.3 (CN-CLIPViT-H/14) to 9.5 (CN-CLIPRN50) times as fast as the Pytorch-based models. For text representation inference, the TensorRT-based models are around 6.2 (CN-CLIPViT-H/14) to 8.2 (CN-CLIPViT-L/14) times as fast as the PyTorch counterparts. We also evaluate the quality of ONNX-based and TensorRT-based model representations by measuring their zero-shot performance on MUGE retrieval dataset. Table 14 provides the experimental zero-shot results, which shows that the converted ONNX-based or TensorRT-based models keeps the quality of vision and text representations well, with no more than 0.1 MR degradation in retrieval performance.