ECLIP: Efficient Contrastive Language-Image Pretraining via Ensemble Confidence Learning and Masked Language Modeling

Jue Wang [1 2 *] **Haofan Wang** [2 *] **Weijia Wu** [2 3] **Jincan Deng** [2] **Yu Lu** [2 4] **Xiaofeng Guo** [2] **Debing Zhang** [2]

Abstract

While large scale pre-training has achieved great achievements in bridging the gap between vision and language, it still faces three challenges. First, the cost for pre-training is expensive. Second, there is no efficient way to handle the data noise which degrades model performance. Third, previous methods only leverage limited image-text paired data, while ignoring richer single-modal data, which may result in poor generalization to single-modal downstream tasks. In this work, we propose **E**fficient **C**ontrastive **L**anguage-**I**mage **P**retraining (ECLIP) via Ensemble Confidence Learning and Masked Language Modeling. Specifically, We adaptively filter out noisy samples in the training process by means of Ensemble Confidence Learning strategy, and add a Masked Language Modeling objective to utilize extra non-paired text data. ECLIP achieves the state-of-the-art performance on Chinese cross-modal retrieval tasks with only 1/10 training resources compared with CLIP and WenLan, while showing excellent generalization to single-modal tasks including text retrieval and text classification.

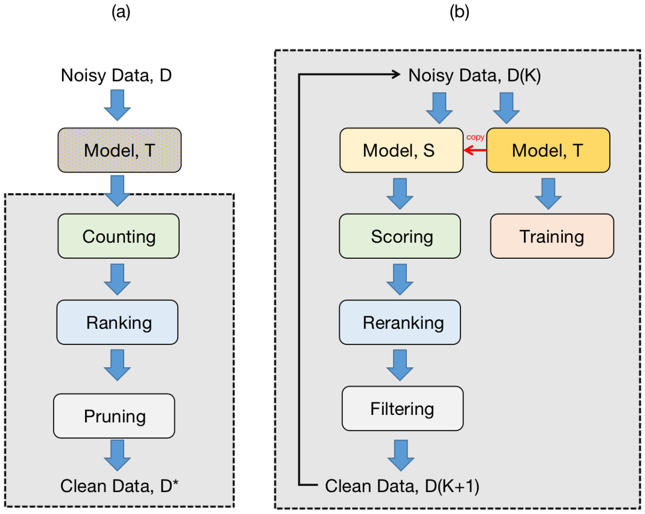

Figure 1. Procedure of Confidence Learning (a) and Ensemble Confidence Learning (b). In CL (a), the scoring model T is a constant model followed by counting, ranking and pruning, the filtering process is decoupled from training. In ECL (b), we maintain a normally training model T and a delay updated shadow model S. The T is one epoch behind by S and copied from T at the beginning of each epoch. We use the delayed S to conduct scoring, ranking, and drop noisy samples of low scores. The training dataset updates at each epoch and is adopted to train T.

1. Introduction

Pre-training has achieved great progress in natural language processing (Devlin et al., 2018; Liu et al., 2019; Radford et al., 2019) and computer vision (Chen et al., 2020; Dosovitskiy et al., 2020) tasks. Recently, cross-modal pre-training (Huo et al., 2021; Radford et al., 2021; Jia et al., 2021) attracts widespread attention, where large scale paired data from the internet is utilized to learn universal cross-modal representation via contrastive learning (Hadsell et al., 2006). However, existing cross-modal pre-training approaches face several common challenges. First, as the scale of data increases, pre-training requires expensive training resources, WenLan (Huo et al., 2021) and CLIP (Radford et al., 2021) cost 896 and 3584 GPU-days respectively. Second, large-scale web-crawled data is noisy and has negative impacts on model performance (Shen et al., 2020; Northcutt et al., 2021). Third, previous cross-modal pre-training methods only use limited image-text pairs, while ignoring single-modal text data that is more accessible, leading to poor generalization to downstream tasks (Li et al., 2020).

Previous works (Radford et al., 2021; Huo et al., 2021) usually perform an elaborate cleaning process (e.g., duplicate and sensitive information detection) to filter out low-quality pairs before pre-training, few studies have focused

[*] Equal contribution [1] Zhongnan University of Economics and Law [2] Kuaishou Technology [3] Zhejiang University [4] University of Technology Sydney. Correspondence to: Jue Wang <201821090281@stu.zuel.edu.cn>.

First Workshop of Pre-training: Perspectives, Pitfalls, and Paths Forward at ICML 2022, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

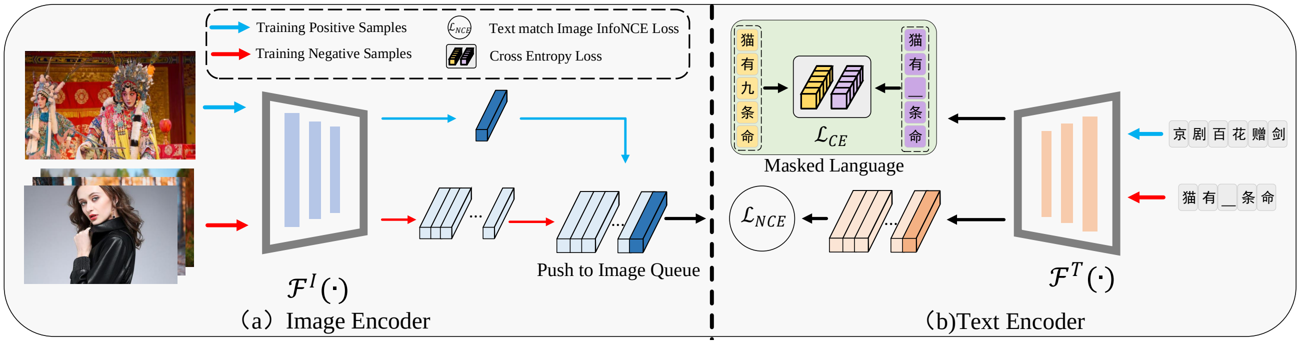

Figure 2. **Illustration of pipeline.** (a) Image Encoder is from the vision branch of CLIP ViT-B/32 and keeps frozen all the time. We construct a memory queue for image features. (b) Text Encoder builds on the top of CLIP ViT-B/32 with additional transformer layers. All layers are learnable. To further improve the representation of text, an auxiliary masked language modeling (MLM) objective is added on text branch utilizing extra text data. We train the pipeline in multitask style, and the training dataset keeps updating (noisy pairs are filtered out by ECL module). Notation: $F^I(.)$ and $F^T(.)$ denote image encoder and text encoder, respectively.

on handling noise without manual intervention. In contrast, there have been numerous works (Shen et al., 2020; Northcutt et al., 2021) that contribute to developing noise-free strategies for vision tasks. Inspired by confident learning (CL) (Northcutt et al., 2021), we introduce an Ensemble Confidence Learning (ECL) strategy, and draw the attention of the large-scale pre-training community to the adverse effects of noise on model performance, rather than simply scaling up the data scale and increasing model parameters. Instead of adopting a constant model (CL) to prune noise, we adaptively ensemble predictions from models in previous training epochs via exponential smoothing for better estimation of noise distribution, and retrain the model with a less noisy subset in the next epoch, which saves training resources (data size decreases) and speeds up convergence (data quality increases). To further boost the generalization performance on downstream tasks in single-modality, we utilize extra non-paired text data which is available in much richer scenes than paired cross-modal data. The text branch is enhanced with a masked language modeling (MLM) task (Devlin et al., 2018). We adopt the image encoder from CLIP (Radford et al., 2021) whose robustness has been validated (Shen et al., 2021). To make up for the scarcity of large-scale pre-trained cross-modal model in Chinese scenarios, we collect language-image paired data in Chinese for pre-training.

ECLIP is evaluated on both cross-modal and single-modal tasks and achieves state-of-the-art results on corresponding benchmarks. Specifically, ECLIP outperforms WenLan (Huo et al., 2021) on R@1 and R@5 by 25.13% and 19.60% respectively, on the cross-modal retrieval tasks of AIC-ICC[1] (Wu et al., 2019). It also exceeds CLIP (Radford et al., 2021) on R@1 and R@5 by 14.21% and 9.03% on text-to-image retrieval task of COCO (Veit et al., 2016). Thanks to the enhancement of the extra non-paired text data, our ECLIP shows strong generalization on single-modal downstream tasks. ECLIP outperforms benchmarks on text-text retrieval and text classification tasks by a large margin. All experiments are conducted under zero-shot settings, in addition to the text classification task. Our key contributions are summarized as below:

1. We propose ECL to dynamically filter noisy data for trade-off between the noise and dataset size, which is the first data-oriented filtering strategy in cross-modal pre-training.

2. We efficiently utilize single modal data (e.g., text data) in the cross-modal pre-training via an MLM task, which is validated to enhance the performance not only on cross-modal tasks but also on the single-modal tasks

3. Our proposed ECLIP achieves a new state-of-the-art result on cross-modal retrieval tasks in Chinese scenario and competitive results in English scenario with 1/10 resources.

2. Approach

2.1. Cross-lingual Distillation

Our framework is targeting for building up a cross-modal pre-trained model in Chinese scenarios. However, most of existed works are trained with image-text pairs in English, which leads to a lingual domain gap between English and Chinese. For convenience, instead of training from scratch, we perform cross-language knowledge distillation and distill a Chinese encoder from existed pre-trained models. Same as regular knowledge distillation frameworks, we train the student model with Chinese texts to mimic the behavior of the teacher model with English texts via MSE loss, where the parameters of the teacher model are frozen, and we only

[1] It is the largest Chinese caption dataset by now.

update the student model.

2.2. Ensemble Confidence Learning

Large scale image-text datasets crawled from the internet have been widely used in pre-training. However, as indicated by (Shen et al., 2020; Northcutt et al., 2021), excessive noisy data negatively impacts the model's performance and training efficiency. ALIGN (Jia et al., 2021) and WenLan (Huo et al., 2021) demonstrate that the large-scale pre-training with expensive resources can suppress the influence of noise to some extent, but these training resources are usually not available for general researchers. An alternative way is to establish clean datasets like COCO (Veit et al., 2016), or Conceptual Captions (Sharma et al., 2018). However, the size of these data sets is often limited due to the need for high-quality manual annotation or complex processing, which constrains the transfer ability to downstream tasks. Driven by these obstacles, a compromised way based on the idea of Confident Learning (Northcutt et al., 2021) named as Ensemble Confident Learning (ECL) is designed. Instead of removing all noisy pairs at once, ECL strategy adopts a similar way as Confident Learning and adaptively removes noisy data from the training set, as it is hard to estimate the distribution of large scale dataset. We find that such dynamic filtering retains a good balance between training dataset size and noise.

As pre-trained model generally performs more discriminative (predict with high confident score) on strong-correlated pairs, we propose to adaptively and iteratively remove the noisy pairs (weak-correlated or non-correlated) by means of the self-discriminative ability of model to data distribution. First, we use the distillation model as initialization which is already equipped with the basic discriminative power and additionally establish a scoring shadow model[2] that only updates parameters at the beginning steps of each epoch.

We define the dataset at epoch K as $D_K = \{d_1, d_2...d_i...d_n\}$ where d_i is an image-text pair and n is the size of the current dataset. The scoring shadow model at epoch K is denoted as S_K. The pre-trained model that keeps updating in the process is denoted as T. We also maintain a moving average score C_K^i that represents the estimated correlation of d_i at epoch K. In the first epoch, T and S_K are the same. The complete iteratively ECL strategy can be decomposed to 3 steps.

(1) Scoring & Training. Given image-text pair d_i, we first obtain its correlation score $S_K(d_i)$ through the scoring shadow model, where the correlation score is the cosine similarity in our case. Then, we update the total score C_K^i based on modified exponential moving average.

$$C_K^i = \alpha * C_{K-1}^i + S_K(d_i), \tag{1}$$

where α is a constant decay factor, and is set to 0.9. In the first epoch ($K = 1$), C_0^i is set to 0. The correlation score of d_i at epoch K is the weighted average of the moving average score and current correlation score given by scoring shadow model S_K. We only update the model T with D_K and keep S_K frozen in this step.

(2) Re-ranking & Filtering. After obtaining the updated moving average score C_K^i of each pair at epoch K, we re-rank all pairs in the dataset D_K by descending order. To filter out noisy pairs, as predicted score range varies among models, it is not practical to set a hard score threshold where pairs of low correlation scores are removed. In contrast, we set a filter rank λ and drop out those pairs whose rank index is larger than $\lambda \times n$. Then, the dataset is updated to D_{K+1}.

$$D_{K+1} = \{d_1^*, d_2^*, \cdots, d_{\lambda \times n}^*\}, \tag{2}$$

where $*$ means the pairs are ordered by descending order.

(3) Shadow Updating. Before diving into the next epoch $K+1$, we update the parameters of the scoring shadow model S_K with the parameters of the pre-trained model, T which has been updated in step (1).

$$S_{K+1} = T, \tag{3}$$

Then, we return to step (1) for the training of next epoch as shown in Figure 1. We admit that non-noisy samples indeed may be filtered out in step (2), as we drop a fixed proportion of samples in each epoch. With the progress of training, the range of output score varies, which makes it unpractical to set a fixed threshold. Thus, we take this way to filter noisy data as a trade-off. We stop the filtering steps when the R@1 on validation set does not increase. The effectiveness of Ensemble Confidence Learning (ECL) is provided in Appendix B.4.1.

2.3. Single-Modality Enhancement

Existing cross-modal pre-trained models utilize large scale image-text pairs from the internet, while not realizing that those paired datasets are usually scene-limited. We observe that image-text pairs are common in some specific domains, such as Wikipedia, News and several human-annotated public datasets. However, in other more professional domains such as Technology and Medical, paired data is scarce, single-modal non-paired data is abundant instead. Driven by the idea of multi-modal few-shot learning (Tsimpoukelli et al., 2021), we additionally leverage extra single-modal text data from various scenes and aim to enhance the generalization of model to downstream tasks.

[2] A shadow model is defined as a delay updated model derived from another existing model that updates normally.

We adopt the commonly adopted masked language model (MLM) task proposed by BERT (Devlin et al., 2018) as auxiliary objective for self-supervised training. MLM takes advantage of bidirectional semantic information, and only requires a simple masking operation on the original text. We find that cross-modal pre-training benefits from additional prevalent single-modal non-paired data, and effectively relief scene-limited problem of cross-modal image-text pairs. On the one hand, the model gains more knowledge about rich scenes from additional text data, which helps avoid the problem of limited distribution of image-text pairs. On the other hand, the MLM task facilitates the model to pay more attention to the relationship between words and helps the model avoid catastrophic forgetting of token-level knowledge, which results in the improvement of transferring tasks. To validate our claim, we show the single-modality enhancement in Appendix B.4.2. The pipeline is in Figure 2.

3. Experiment

3.1. Dataset Preparation

Enabled by the large amounts of publicly available data on the internet, we construct 3 types of training datasets. The first is a Chinese-English text paired dataset collected from available public translation datasets totaling 67 million translation pairs, which is used for cross-lingual distillation to obtain a distilled text encoder in Chinese. The second is a large Chinese image-text dataset of 300 million pairs, which is collected from the internet without elaborate cleaning process and is used for pre-training. The last is a Chinese text dataset of 20 millions samples after simple cleaning as extra single-modal text for enhancement. Details are listed below.

3.1.1. Training Dataset

We attribute the training datasets for 3 purposes. The Chinese-English Paired Texts are for cross-lingual distillation. The Chinese Texts are for single-modality enhancement. The Image-Text Pairs are for cross-modal pre-training.

Chinese-English Paired Texts are collected from AI Challenge Machine Translation[3], WMT20[4], translation2019zh[5], UN Corpus[6] and other publically available sources[7], totaling 67 million translation pairs.

Chinese Texts are collected from the single-modal text dataset of CLUE[8], which is the largest Chinese language understanding evaluation benchmark. We remove text of Chinese character ratio less than 50% and meaningless symbols (“&”, “-”, etc.), has 20,329,263 texts in total.

Image-Text Pairs are crawled from the Internet and of 300 millions uncleaned pairs. Specifically, we establish a Chinese word dictionary including 4 millions Chinese words. Each word from the dictionary is used as a query to crawl image-text pairs from the largest Chinese Search Engine (Baidu Pictures and Baidu Baike).

3.1.2. Validation Datasets

Image-Text Pairs are used for tuning hyperparameters, we extra collect image-text pairs from various scenarios on the internet including Baidu Pictures, Baidu Baike, Toutiao, hashtag, and other sources. To obtain a validation set in high quality (image-text pairs are strong correlated), we first filter out pairs whose text does not have words existed in a 40 thousand common entity vocabulary. Then, we calculate the cosine similarity for each pair applying the distillation model and take out the top10K pairs as our validation set that has no overlap with pairs in the training dataset.

3.1.3. Pre-Processing

There exists enormous meaningless symbols in the text, thus we conduct simple pre-processing on the text before pre-training. First, we remove the redundant HTML, space symbols (“. . . ”, “—”) and emojis from text, and then replace interval symbols (“&”, “-”, etc.) by “,”. Meanwhile, we filter out short text of length less than 4 and text of a Chinese character ratio less than 50%.

3.2. Implementation Details

We also provide the details of our implementation, including cross-lingual distillation, cross-model pre-training, hyperparameters setting and the implementation of SimCSE, which can be found in Appendix B.1.

3.3. Evaluation Results

To measure the capability of task-agnostic models, zero-shot evaluations have been widely adopted and proved being more representative of a model’s ability (Radford et al., 2021). We evaluate the effectiveness of ECLIP on both cross-modality and single-modality tasks under zero-shot settings. The CLIP model used is the CLIP ViT-B/32[9]. All benchmarks without special mention are pre-trained models.

3.3.1. Cross-Modality Evaluation

AIC-ICC (Wu et al., 2017) is the only publicly-available Chinese multi-modal dataset. The training set contains 210,000

[3] https://challenger.ai

[4] http://www.statmt.org/wmt20/

[5] https://www.kaggle.com/terrychanorg/translation2019zh

[6] https://conferences.unite.un.org/uncorpus

[7] https://github.com/quincyliang/nlp-public-dataset

[8] https://www.cluebenchmarks.com/

[9] https://github.com/openai/CLIP

Method	Text2Image(%)			Image2Text(%)		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP†	7.80	18.50	25.00	13.40	27.30	35.10
UNITER†	9.80	23.30	31.40	14.80	29.80	37.90
CLIP‡	11.06	24.89	33.28	18.33	34.45	43.23
WenLan	14.40	30.40	39.10	20.30	37.00	45.60
ECLIP	**18.02**	**36.36**	**46.05**	**26.67**	**44.78**	**53.17**

Table 1. Evaluation results for cross-modal retrieval tasks on the AIC-ICC test subset. † and ‡ denote translation and distillation. The results of CLIP†, UNITER† and WenLan are from WenLan. We distill CLIP‡ from CLIP† via cross-lingual distillation.

images, and the validation set contains 30,000 images. Each image has 5 descriptions. We evaluate the zero-shot transfer ability on the test subset (10,000 data) as WenLan (Huo et al., 2021). To compare with those benchmarks trained with English data, we translate text from Chinese to English via Google translation API or conduct cross-lingual distillation as mentioned in Section 2.1 to obtain a distilled Chinese encoder. Table 1 presents the cross-modal retrieval results. ECLIP outperforms all other benchmarks significantly on both text-to-image and image-to-text retrieval tasks.

To mitigate the potential adverse effect of translation on benchmarks and show our effectiveness comprehensively, we distill an English encoder from ECLIP following the same protocol as decribed in Section 2.1. We evaluate on COCO2014 test set (Veit et al., 2016) which is a commonly adopted image caption datasets of 5,000 images and 24,716 captions in English. The results are in Table 2.

Method	Text2Image(%)			Image2Text(%)		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	30.75	56.60	67.73	**50.78**	75.10	83.70
ECLIP‡	**35.12**	**61.71**	**72.77**	50.70	**76.56**	**84.98**

Table 2. Evaluation results for cross-modal retrieval tasks on the COCO2014 test set. ‡ means distillation.

As shown, ECLIP outperforms the original CLIP by a large margin on text to image retrieval subtask. On image to text retrieval subtask, ECLIP also performs better than CLIP on R@5 and R@10 while slightly worse on R@1.

3.3.2. Single-Modality Evaluation

Previous cross-modal pre-trained models usually cannot effectively adapt to single-modal (NLP) scenarios (Li et al., 2020). We bridge this gap between cross-modal retrieval and its transfer ability to single-modality tasks by introducing extra masked language modeling auxiliary task, and evaluate on text classification and text to text retrieval tasks.

Method	Text Match(%)		
	R@1	R@5	R@10
CLIP†	22.94	36.25	42.65
WenLan	31.44	45.32	52.98
CLIP‡	37.34	53.26	60.64
SimCSE	39.64	56.49	63.24
ECLIP	**43.48**	**60.36**	**67.74**

Table 3. Results for short text retrieval on AIC-ICC test subset. † and ‡ means translation and distillation, respectively.

Text Classification. To validate the NLU (Natural Language Understanding) capability of language model on short texts, we evaluate on TNEWS dataset from CLUE[10]. Each title is labeled with one of 15 news categories (finance, technology, sports, etc.) and the task is to predict which category the title belongs to. The training set contains 53,360 samples, and the test set and validation set both contain 10,000 samples. We compare with 3 pre-trained benchmarks from CLUE. The results are in Table 4.

Model	XLNET	RoBERTa	ALBERTA	ECLIP
Acc(%)	56.24	58.61	59.46	**67.20**

Table 4. Text classification results on TNEWS. Benchmarks are XLNet-mid, RoBERTa-wwm-large and ALBERT-xxlarge.

Text Retrieval. To measure the discriminative ability of text embedding and the zero-shot transfer ability, we evaluate on AIC-ICC (Wu et al., 2017) test subset (only use the texts) where each image has 5 corresponding descriptions. We randomly select one of the 5 texts as the query, and the rest of 4 texts as the key. Results are in Table 3.

ECLIP outperforms all benchmarks by a large margin on both text classification and text retrieval tasks, which shows better transfer ability to downstream single-modality tasks.

4. Conclusion

In this work, we introduce Efficient Contrastive Language-Image Pretraining (ECLIP) via Ensemble Confidence Learning and Masked Language Modeling. We investigate the first data-oriented filtering strategy in cross-modal pretraining and adaptively filter the noisy data, and show the value of single-modality enhancement with non-paired text data as auxiliary objective to improve transferability. ECLIP achieves the leading performance on cross-modal retrieval tasks in Chinese scenario and competitive results in English scenario, and also outperforms benchmarks on single-modality downstream NLP tasks.

[10] https://github.com/CLUEbenchmark/CLUE

References

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gao, T., Yao, X., and Chen, D. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

Hadsell, R., Chopra, S., and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pp. 1735–1742. IEEE, 2006.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Huo, Y., Zhang, M., Liu, G., Lu, H., Gao, Y., Yang, G., Wen, J., Zhang, H., Xu, B., Zheng, W., et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

Joulin, A., Van Der Maaten, L., Jabri, A., and Vasilache, N. Learning visual features from large weakly supervised data. In *European Conference on Computer Vision*, pp. 67–84. Springer, 2016.

Li, W., Gao, C., Niu, G., Xiao, X., Liu, H., Liu, J., Wu, H., and Wang, H. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

Liu, X., Chen, Q., Deng, C., Zeng, H., Chen, J., Li, D., and Tang, B. Lcqmc: A large-scale chinese question matching corpus. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1952–1962, 2018.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Northcutt, C. G., Jiang, L., and Chuang, I. L. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 2021.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Shen, S., Li, L. H., Tan, H., Bansal, M., Rohrbach, A., Chang, Kai-Wei andf Yao, Z., and Keutzer, K. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

Shen, Y., Ji, R., Chen, Z., Hong, X., Zheng, F., Liu, J., Xu, M., and Tian, Q. Noise-aware fully webly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11326–11335, 2020.

Tsimpoukelli, M., Menick, J., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *arXiv preprint arXiv:2106.13884*, 2021.

Veit, A., Matera, T., Neumann, L., Matas, J., and Belongie, S. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.

Wu, J., Zheng, H., Zhao, B., Li, Y., Yan, B., Liang, R., Wang, W., Zhou, S., Lin, G., Fu, Y., et al. Large-scale datasets for going deeper in image understanding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1480–1485. IEEE, 2019.

A. Related Work

A.1. Cross-Modal Pretraining

Vision-language pre-training has attracted growing attention during the past few years. CLIP (Radford et al., 2021) trains with 400 million image-text data pairs from the internet after simple data cleaning, and achieves excellent performance on image-text retrieval tasks and zero-shot image recognition tasks. ALIGN (Jia et al., 2021) further expands the scale of image-text paired data, and uses 1 billion data to train without any cleaning, indicating that the expansion of the data scale can suppress the influence of noise for the model to some extent. WenLan (Huo et al., 2021) trains with Chinese paired data and achieves the best performance on image-text retrieval tasks in the Chinese scene. However, all of these works pay few attention on the adverse effect of noisy data while increasing the scale of training dataset, which motivates us to propose a data-oriented filtering strategy in cross-modal community for efficient pre-training.

A.2. Learning with Noisy Data

A number of approaches (Joulin et al., 2016; Northcutt et al., 2021) has been proposed to train models with noisy labeled data. Confident Learning (Northcutt et al., 2021) is the state-of-the-art for weak supervision, finding label errors in datasets, which works by 'learning from confident examples', and confident examples are identified as examples with high predicted probability for their training labels. Confident learning moderately increases model accuracy by cleaning data prior to training process through continuous steps of Count, Rank and Prune. Instead of decoupling the filtering process and model training, we maintain one normally training model and one delay updated model to score each sample, and adopt an iterative strategy to filter out noisy samples at each epoch. We show their difference in Figure 1.

A.3. Masked Language Modeling

Masked language modeling (MLM) (Devlin et al., 2018) is a self-supervised pretraining objective, which has been widely adopted in natural language processing for learning text representations. Specifically, MLM trains a model to predict a random sample of input tokens that have been replaced by a [MASK] placeholder in a multi-class setting over the entire vocabulary. When pretraining, it is common to use MLM to improve downstream performance. In our work, we introduce MLM into cross-modal pre-training utilizing extra text data and validate its value for improving the performance on both cross-modal and single-modal tasks.

B. Experiment

B.1. Implementation Details

B.1.1. Cross-Lingual Distillation

To obtain a Chinese text encoder efficiently, we conduct cross-language knowledge distillation. Specifically, we initialize our text encoder on the top of CLIP ViT-B/32 (Radford et al., 2021) with additional transformer layers, all layers are learnable. We train the Chinese text encoder on the Chinese-English text paired dataset following the scheme described in Section 2.1 with batch size of 256.

B.1.2. Cross-modal Pre-Training

We adopt the image encoder from CLIP ViT-B/32 (Radford et al., 2021) and keep it frozen in the pre-training process. The distilled Chinese encoder is used as text encoder. Following MOCO (He et al., 2020) and CLIP (Radford et al., 2021), we set the learning rate to 2×10^{-3}, weight decay to 10^{-4}, dropout to 10^{-1}, and use cosine warm-up schedule as learning rate adjuster. As for the size of image memory queue, we set it to 5×10^4 as a trade-off between performance and training efficiency. The filter rank is set to 0.9.

In the pre-training process, we design two data loaders for the image-text pairs and text data. The batch size of image-text data and text data are set to 180 and 40 respectively. The text data is loaded similarly with BERT: the masking probability is set to 15% for each token, and each masked token has a 20% chance to be replaced by other tokens.

ECLIP is trained via a regular multitask learning scheme, and its performance is affected by the weights between the contrastive learning task and extra masked language modeling task. As the contrastive learning task is major objective and the other is auxiliary objective, we simply set the weights proportional to their batch sizes without complicated parameter search. It is worth mentioning that we remove the MLM task from the pre-training process when the filtering stops, making the model focusing on contrastive learning tasks.

B.1.3. Hyperparameters Setting

Cross-Lingual Distillation We build our text encoder on the top of CLIP ViT-B/32 (Radford et al., 2021) added with 32 transformer layers[11]. For each layer, the feature dimension is 512, depth is 20 and sequence length is 77. During training, we adopt Adam optimizer where learning rate is 2×10^{-4}, eps is 1×10^{-8}, betas is (0.9,0.98) and weight decay is 0.01. The cosine schedule warm-up is used to adjust learning rate adaptively. For the tokenize method, we adopt the method from CPM[12].

Cross-modal Pre-Training We find hyperparameters based on the results on AIC-ICC validation set. For the queue size of image encoder, we train 2×10^5 steps and find that the R@1 does not increase significantly when the queue size reaches 5×10^4. We adopt this value for the trade-off between R@1 and training efficiency.

Queue size	2×10^2	2×10^3	2×10^4
R@1	14.34	14.99	16.54
Queue size	5×10^4	1×10^5	2×10^5
R@1	17.15	17.18	17.2

Table 5. Effect of queue size.

For the choice of filter rank, we empirically test with 4 values and adopt 0.9 in all experiment setting.

Filter Rank	0.7	0.8	**0.9**	0.99
R@1	17.66	17.82	**18.02**	14.70

Table 6. Effect of filter rank.

B.1.4. Implementation of SimCSE

In single-modality evaluation on text retrieval, we train a SimCSE utilizing wiki2019zh[13] and BQ Corpus[14], following the same hyperparameter setting as the original SimCSE (Gao et al., 2021). We divide text into multiple shorter sentences with a length of less than 77 as input data, and conduct pre-processing as mentioned in 3.1.3. We initialize our Chinese SimCSE with the pre-trained SimCSE-RoBERTa-large model. We add [CLS] token to each text, and use the representation at the position of [CLS] token as its sentence embedding. We set the learning rate to 1×10^{-5}, the weight decay to 1×10^{-4}, dropout to the default 0.1. We train the model in unsupervised style with wiki2019zh dataset and then fine-tune it on BQ Corpus for supervised learning.

B.2. Single-Modality Evaluation on other datasets

B.2.1. Text Retrieval in Chinese Domain

We evaluate on AFQMC and LCQMC (Liu et al., 2018) which are also commonly adopted semantic matching datasets in Chinese in addition to AIC-ICC. The results are shown in Table 7.

B.2.2. Text Retrieval in English Domain

Besides of evaluation on Chinese domain, we also evaluate on COCO2014 test set (only use its captions) for a more convincing validation. The results can be found in Table 8.

[11] https://github.com/lucidrains/DALLE-pytorch

[12] https://github.com/yangjianxin1/CPM

[13] https://www.kaggle.com/ziyunshuang/wiki2019zh

[14] http://icrc.hitsz.edu.cn/Article/show/175.html

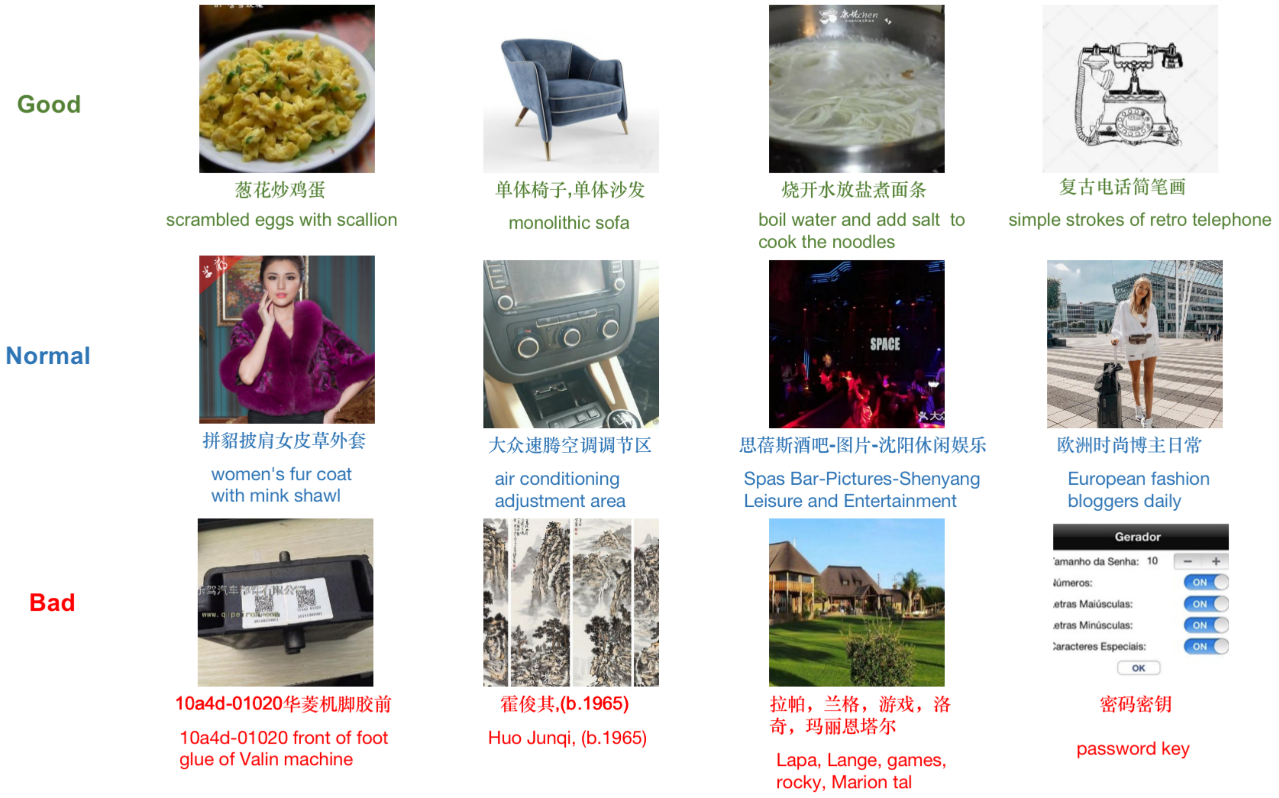
Good
葱花炒鸡蛋
scrambled eggs with scallion
单体椅子,单体沙发
monolithic sofa
烧开水放盐煮面条
boil water and add salt to cook the noodles
复古电话筒笔画
simple strokes of retro telephone
Normal
拼貂披肩女皮草外套
women's fur coat with mink shawl
大众速腾空调调节区
air conditioning adjustment area
SPACE
思蓓斯酒吧-图片-沈阳休闲娱乐
Spas Bar-Pictures-Shenyang Leisure and Entertainment
欧洲时尚博主日常
European fashion bloggers daily
Bad
10a4d-01020华菱机脚胶前
10a4d-01020 front of foot glue of Valin machine
霍俊其,(b.1965)
Huo Junqi, (b.1965)
Gerador
'amanho da Senha: 10
Júmeros: ON
.etras Maiúsculas: ON
.etras Minúsculas: ON
Caracteres Especiais: ON
OK
拉帕，兰格，游戏，洛奇，玛丽恩塔尔
Lapa, Lange, games, rocky, Marion tal
密码密钥
password key

Figure 3. Manually annotated examples of pairs in 'Good', 'Normal' and 'Bad' quality. The original text is in Chinese, we provide translated English text for those non-native Chinese speakers. The ratio of 'Bad' dramatically reduced after ECL module.

B.3. Visualization

B.3.1. Annotation Guideline

We manually check the dataset quality in a qualitative way. Each annotator is asked to label 100 image-text pairs randomly sampled from the corresponding dataset. We define the 'Quality' of pairs into 3 types. **Good case** is a strong-correlated pair, where the text exactly describes what happens in its corresponding image. **Normal case** is a weak-correlated pair, where the text just contains objectives that appear in the image. **Bad case** is a non-correlated pair, where the text is absolutely irrelevant with the image. For more intuitively understanding of each type, we provide visual results in B.3.2.

B.3.2. Visualization Results

We provide visualization results of manually annotated pairs in Figure 3 for better understanding. We show that ECL effectively filter out those bad pairs.

B.4. Ablation Studies

B.4.1. Effect of Ensemble Confidence Learning

The key contribution in our approach is the Ensemble Confidence Learning (ECL) strategy, where we adaptively filter noisy data for improving training efficiency and model performance. To show the effectiveness of ECL, we train ECLIP under 4 different settings, and evaluate them on 3 tasks. Specifically, we first train ECLIP (300M) using the original 300 millions noisy pairs without ECL and any other filtering process. Then we train ECLIP (200M*) and ECLIP (100M*) with 200 millions and 100 millions pairs that are less noisy than the original data, these data are the subset of 300 millions after ECL process, the denoise is conduct at the beginning and no filtering is adopted in the training. In ECLIP+ECL (300M), we train ECLIP with 300 millions noisy pairs and filter out noisy data adaptively with ECL.

Method	AFQMC(%)			LCQMC(%)		
	R1	R5	R10	R1	R5	R10
CLIP[†]	6.43	14.80	19.96	57.17	78.22	82.74
WenLan	9.72	18.68	23.92	66.27	87.01	90.82
CLIP[‡]	12.03	24.59	31.54	74.53	93.50	96.18
SimCSE	13.34	27.22	33.78	78.56	95.52	97.32
ECLIP	**15.77**	**30.72**	**36.54**	**81.58**	**98.10**	**98.88**

Table 7. Results for short text retrieval on AFQMC and LCQMC. † and ‡ means translation and distillation, respectively.

Method	Text Match(%)		
	R@1	R@5	R@10
CLIP	38.34	59.44	68.52
ECLIP	**40.40**	**62.20**	**71.80**

Table 8. Results on short text retrieval on COCO2014 test set.

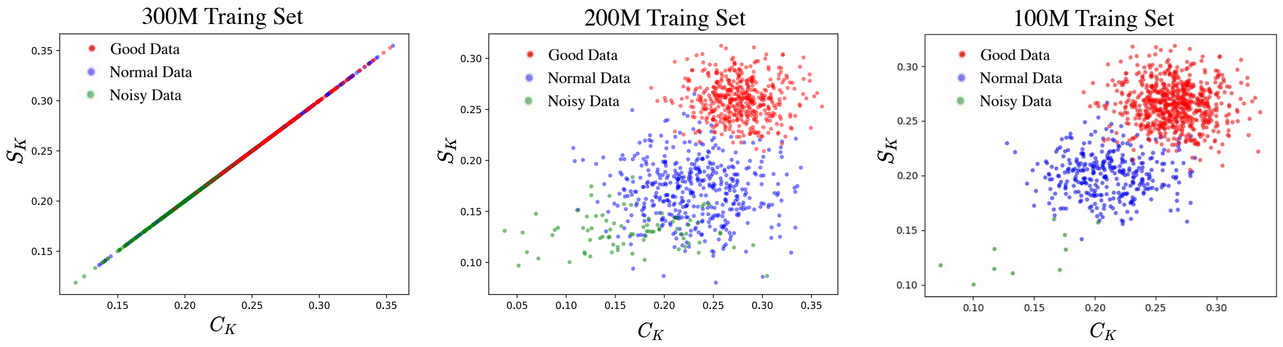

Figure 4. **Confident score distribution of training sets.** 300M, 200M and 100M datasets are of noise level by descending order. C_K and S_K are scores given by moving average and the shadow scoring model. In the first epoch (K=1), C_K is the same as S_K. With the progress of filtering, ECL successfully separates pairs in 'Good', 'Normal' and 'Noisy' based on their confident scores.

Method	T2I	TC	TR
ECLIP (300M)	10.77	65.20	32.70
ECLIP (200M*)	14.87	66.70	40.65
ECLIP (100M*)	16.72	65.92	40.38
ECLIP+ECL (300M)	**18.02**	**67.20**	**43.48**

Table 9. Effect of ECL strategy. T2I, TC and TM denote text-to-image retrieval, text classification and text retrieval respectively. T2I reports the R@1 result for cross-modal retrieval tasks on the AIC-ICC test subset, TC reports the text classification result on TNEWS dataset, TM reports the R@1 result for short text retrieval on AIC-ICC test subset. M stands for a million of uncleaned data. $*$ means the data is cleaned with ECL strategy.

We evaluate these models on text-to-image retrieval, text classification and text retrieval tasks. As the first 3 lines of table 9 shown, as the quality of data increases, although the scale of data decreases, the retrieval performance still gains significant improvement, but when the dataset size is reduced to a certain extent, the downstream performance of NLP degrades. Second, instead of using the highly correlated data (denoted with $*$) at the beginning, ECL strategy adopts an adaptive way to select a training subset from a huge noisy dataset, which is shown to be a good trade-off between dataset size and noise. ECLIP equipped with ECL achieves improvement on both cross-modality and single-modality tasks.

To further provide evidence for the choice of scoring shadow model in ECL, we also evaluate the performance when we only adopt the fixed distillation model for scoring (step 1 in ECL). Results are shown in Table 10.

Models	T2I	TC	TR
ECLIP (wo)	17.03	67.20	42.04
ECLIP (w)	**18.02**	**67.20**	**43.48**

Table 10. Effect of the scoring model. T2I, TC and TR denote text-to-image retrieval, text classification and text retrieval respectively. w and wo represent with or without updating scoring model.

To verify the effectiveness of ECL module for adaptive filtering noisy pairs, we randomly sample 1000 samples from the 300 million, 200 million, and 100 million datasets (The last 2 datasets are cleaned by ECL from the 300 million dataset) respectively, and annotate them manually given 3 candidate labels. 'Clean' represents those pairs that are strong-correlated in semantic, 'Noisy' represents those pairs that are non-correlated in semantic, the rest of pairs are attributed to 'Good' that usually has weak-correlation. We count the proportion and show the human annotated results in Table 11. We visualize the distribution of the manual data, as shown in Figure 4. It can be seen that ECL has a strong ability to remove noisy data and separate the good cases from the other cases. We also provide annotation guideline and actual 'Noisy', 'Normal' and 'Good' image-text pairs and visualization.

dataset size	300M	200M	100M
Noisy %	28.0	8.0	1.0
Normal %	51.0	45.0	33.0
Good %	21.0	47.0	66.0

Table 11. Denoise performance of ECL.

B.4.2. Effect of Masked Language Modeling

The another contribution of our work is that we introduce the auxiliary objective MLM to cross-modal pre-training and show its effectiveness to improve generalization. We also evaluate the effect of the MLM objective on text-image retrieval, text classification and text retrieval tasks.

Models	T2I	TC	TR
ECLIP (wo)	17.27	63.40	40.30
ECLIP (w)	**18.02**	**67.20**	**43.48**

Table 12. Effect of Masked Language Modeling on cross-modal pre-training. T2I, TC and TM denote text-to-image retrieval, text classification and text retrieval respectively. T2I reports the R@1 result for cross-modal retrieval tasks on the AIC-ICC test subset, TC reports the text classification result on TNEWS dataset, TM reports the R@1 result for short text retrieval on AIC-ICC test subset.

As shown in Table 12, the extra single-modal (text) pre-training branch not only enhances model's zero-shot transfer ability on text classification and text retrieval (single-modal) tasks, but also improves the performance on text-to-image retrieval (cross-modal) tasks. As expected, the performance gains from text retrieval (3.18%↑) and text classification (3.80%↑) are higher than cross-modal retrieval (0.75%↑).