Utilizing Cross-Modal Contrastive Learning to Improve Item Categorization BERT Model

Lei Chen[*], **Hou Wei Chou**[*]
Rakuten Institute of Technology
Boston, MA, USA
{lei.a.chen,houwei.chou}@rakuten.com

Abstract

Item categorization (IC) is a core natural language processing (NLP) task in e-commerce. As a special text classification task, fine-tuning pre-trained models, e.g., BERT, has become a main stream solution. To improve IC performance further, other product metadata, e.g., product images, have been used. Although multimodal IC (MIC) systems show higher performance, expanding from processing text to more resource-demanding images brings large engineering impacts and hinders the deployment of such dual-input MIC systems. In this paper, we proposed a new way of using product images to improve text-only IC model: leveraging cross-modal signals between products' titles and associated images to adapt BERT models in a self-supervised learning (SSL) way. Our experiments on the three genres in the public Amazon product dataset show that the proposed method generates improved prediction accuracy and macro-F1 values than simply using the original BERT. Moreover, the proposed method is able to keep using existing text-only IC inference implementation and shows a resource advantage than the deployment of a dual-input MIC system.

1 Introduction

Item categorization (IC) is a core natural language processing (NLP) technology in e-commerce. Since millions types of products are provided in e-commerce markets, it is important to map these products to their locations in a product category tree efficiently and accurately so that buyers can easily find their interested products. Therefore, IC models with a high accuracy are needed for the success of e-commerce business. In spite that IC shares the same setup as a text classification task, it possesses its unique aspects, including (a) handling a large number of prediction labels, (b) a severe long-tailed distribution of labels, and (c) noisy raw inputs due to the fact that these inputs are generally provided by merchants in a heterogeneous way. These unique aspects make IC be a challenging task in practice.

Fine-tuning pre-trained models, e.g., BERT (Devlin et al., 2019), has become a main stream approach on building high-performance NLP applications. When using this paradigm to build IC models, is there any way to achieve an even higher performance? This is the first research question we tackled in this paper. One approach to improving IC models that generally use products' text titles alone is utilizing products' images. Previously, multimodal IC models using both text and image inputs have been actively investigated and applied in practice. However, such dual-input multimodal IC models bring more burden to the operation. Comparing to handling text data, the resource needed for storing/transferring/processing image data is much higher. For industry-scaled IC systems, adding image processing in its inference stage is costly. Is there any other way to get benefits by utilizing products' associated images but not paying such a high cost? This is the second research question we focused in this paper.

To tackle these two questions, inspired by (Zhang et al., 2020; Fang et al., 2020), we propose a solution of running a cross-modal contrastive learning, a special self-supervised learning (SSL), between products' images and text titles, to adapt pre-trained models to fit the IC task domain better. Then, the adapted pre-trained models will be used to build IC models using the fine-tuning paradigm. Moreover, by using the cross-modal SSL training, images can be used to improve text-based pre-trained models in the model training stage and a series of costly changes/operations in the inference stage can be avoided.

[*] Equal contributor

Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5), pages 217 - 223

2 Related works

Fine-tuning pre-trained models has been becoming a main stream method for building high quality IC systems. For example, in a recent data challenge for building multimodal IC systems, which was organized in the SIGIR'20 e-commerce workshop[1], fine-tuning BERT models (Devlin et al., 2019) has been used by most of the participants (Bi et al., 2020; Chordia and Vijay Kumar, 2020; Chou et al., 2020).

In order to improve IC performance, one direction has been exploring utilizing more metadata associated with products. (Zahavy et al., 2016) is a seminal work where multi-label classification using both titles and images was conducted on products listed on the Walmart.com website. They used a convolutional neural network (CNN) to extract representations from both titles and images, and designed several policies to fuse the outputs of the two models. This led to improved performance over individual models separately. In the SIGIR'20 multimodal IC data challenge, different text-image fusion methods have been explored. Roughly in order of increasing complexity, the methods included simple decision-level late fusion (Bi et al., 2020), highway network (Chou et al., 2020), and co-attention (Chordia and Vijay Kumar, 2020). By reviewing the experiment results from several teams, we can observe that (a) dual-input IC models show higher performance than any uni-modal IC model, (b) performance gains brought by using images are limited but not neglectable. When considering IC performance's profound impacts on e-commerce business values, including products' visual information is necessary.

Contrastive learning (CL) has been found to be an effective self-supervised learning (SSL) approach for training high-quality representations. For example, in computer vision, SimCLR (Chen et al., 2020) uses the consistence between an anchor image and its transformed version and the in-consistence between the anchor and other instances in a batch (negative instances) to guide learning visual representations. Without using labels, it achieves visual representations with a quality on par or even higher than the ones trained based on traditional supervised learning. Inspired by the success of SimCLR in computer vision, CL-based text representation learning has been a hot research topic in NLP. Both SimCSE (Gao et al., 2021) and (Liu et al., 2021) used dropout operations existing in Transformer architecture (Vaswani et al., 2017) to be an effective text augmentation way and obtained effective text representations.

The SSL idea has been tried in the NLP domain for further improving BERT models. For example, (Fang et al., 2020) proposed using SSL to improve the pre-trained BERT model prior to running down-stream NLP tasks, e.g., various tasks in the GLUE benchmark test. When generating augmented instances for providing positive pairs, a back-translation method is used. The CL setup was based on MoCo architecture by using a momentum mechanism (He et al., 2020). (Su et al., 2021) also used the SSL to improve pre-trained BERT model for more accurate relation extraction (RE) task. When doing text augmentation, it considered the RE task's unique property and proposed a task-specific augmentation method. In these two works, SSL training was found to be useful for improving follow-up fine-tuning tasks' performance.

Regarding the SSL methods being used, an interesting trend is considering cross-modal signals. (Zhang et al., 2020) proposed Contrastive VIsual Representation Learning from Text (ConVIRT) model to use text descriptions associated with medical images to help training more accurate medical image representations. Note that in the medical image domain, the annotated image data size is much limited, in a contrast, medical text notes are more adequate. (Radford et al., 2021) is a seminal work from OpenAI. By using a massive set of image-text pairs, about 350 million, CLIP used a simple cross-modal contrastive learning to pre-train a quite powerful vision-language joint model. The trained model shows many impressive applications, like superior performance on many zero-shot image classification tasks.

3 Model

Figure 1 depicts our proposed model in a concise way. From a pre-trained BERT model, denoted as BERT_{origin}, a self-supervised learning (SSL) is applied to further adapt the BERT_{origin} fitting to the fine-tuning task better. When selecting the data set used in the SSL step, note that due to the self-supervised nature, we don't need human-annotated labels. Then, the adapted BERT is used to initialize the BERT model (serving as a textual feature encoder) in the fine-tuning stage. The final IC model,

[1] https://sigir-ecom.github.io/ecom2020/data-task.html

consisting of BERT and a linear classifier on top of it, is learned jointly by using a cross-entropy loss on the fine-tuning data set.

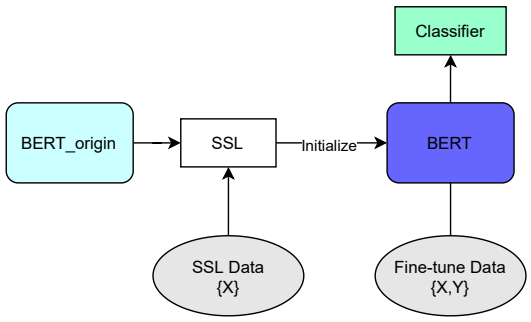

Figure 1: Self-Supervised Learning (SSL) is applied to improve pre-trained BERT, BERT_{origin}, for fitting the fine-tuning task better. Then, a conventional fine-tuning is conducted to jointly update both BERT (for representation learning) and classifier training.

3.1 SSL SimCSE

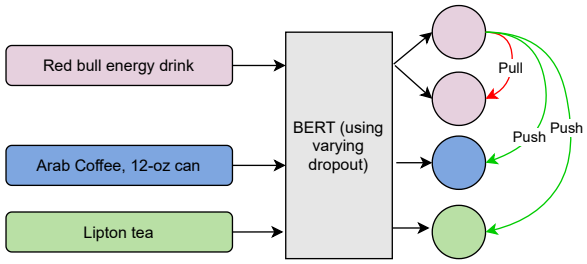

Figure 2: An illustration showing how SimCSE works. Note that we only show contrastive pairs, i.e., both positive and negative, to the top product "red bull energy drink".

For a text title $\mathbf{x}_t$, we obtain a text representation $\mathbf{t}$ with a BERT encoder $f_t(, d)$ where d is a dropout mask, and a projection function g_t, which uses a simple multiple layer perception (MLP) structure.

$$\mathbf{t} = g_t(f_t(\mathbf{x}_t, d)) \tag{1}$$

To obtain a positive pair, SimCSE just runs the same text title throughout the Transformer encoder pipeline with a different dropout mask d^+.

$$\mathbf{t}^+ = g_t(f_t(\mathbf{x}_t, d^+)) \tag{2}$$

For ith text title, the training objective of SimCSE is like:

$$\mathcal{L}_i^t = -log\frac{exp(sim(\mathbf{t}_i, \mathbf{t}_i^+)/\tau)}{\sum_{j=1}^{N,j\neq i} exp(sim(\mathbf{t}_i, \mathbf{t}_j)/\tau)} \tag{3}$$

for a mini-batch of N text titles, where $sim()$ represents a similarity computation and τ is a temperature parameter. The total loss computed by SimCSE is an average among all text titles

$$\mathcal{L}_{simCSE} = \sum_{i=1}^{N} \mathcal{L}_i^t / N \tag{4}$$

in the mini-batch.

As shown in Figure 2, the top title serves as an anchor and is sent to a BERT model twice to obtain two similar but varying text representations because of using varying dropout masks. The augmented version serves as a positive pair to the anchor while other two titles in the batch serve as negative pairs. Through using an infoNCE loss (Oord et al., 2018), BERT encoder is changed to pull positive pairs closer while pushing negative pairs away. Clearly, without any supervision, BERT encoder can be further adapted to provide a representation better fitting to the SSL data set.

3.2 SSL ConVIRT

In the SSL SimCSE method, both positive and negative pairs are from the text domain. Inspired by the cross-modal contrastive learning in (Zhang et al., 2020), we use product images that co-exist with text titles to provide self-supervision signals.

Regarding the visual encoder used to process product images to visual representation vectors, we choose a newly emerging encoder based on Transformer like BERT model. In recent years, Transformer based visual models have become popular (Han et al., 2020). Among the many visual Transformer models, we selected the ViT model (Dosovitskiy et al., 2020), which is a pure Transformer that is applied directly on an image's $P \times P$ patch sequence. In the implementation, it follows the original Transformer's design as much as possible. ViT utilizes the standard Transformer's encoder part as an image classification feature extractor and adds a MLP head to determine the image labels. The ViT model is pre-trained using a supervised learning task on a massive image data set. The size of the supervised training data set impacts ViT performance significantly. When using Google's in-house JFT 300M image set, ViT can reach a performance superior to other competitive ResNet (He et al., 2016) models. After converting a product image to $P \times P$ patches, ViT converts these patches to visual tokens. After adding a special [CLS] visual token to represent the entire image, the $M = P \times P + 1$ long sequence

is fed into a ViT model to output an encoding as $\mathbf{v} = (v_0, v_1, v_2, ...v_M)$, where $M = P \times P$.

For a product i with a text title x_t and an image x_v, we obtain its visual representation by running through a visual processing pipeline including a ViT image encoder f_v and a projection layer g_v, which is also an MLP.

$$\mathbf{v} = g_v(f_v(x_v)) \tag{5}$$

Based on text and image representations, we compute a contrastive loss from text to image direction (denoted as $t \rightarrow v$).

$$\mathcal{L}_i^{t \rightarrow v} = -log \frac{exp(sim(\mathbf{t}_i, \mathbf{v}_i)/\tau)}{\sum_{j=1}^{N, j \neq i} exp(sim(\mathbf{t}_i, \mathbf{v}_j)/\tau)} \tag{6}$$

Similarly, a contrastive loss from the other direction, image to text (denoted as $v \rightarrow t$), can be computed as

$$\mathcal{L}_i^{v \rightarrow t} = -log \frac{exp(sim(\mathbf{v}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{N, j \neq i} exp(sim(\mathbf{v}_i, \mathbf{t}_j)/\tau)} \tag{7}$$

$$\mathcal{L}_{ConVIRT} = \sum_{i=1}^{N} (\alpha \mathcal{L}_i^{t \rightarrow v} + (1 - \alpha) \mathcal{L}_i^{v \rightarrow t})/N \tag{8}$$

where α is a hyper-paramter to control two contrastive losses in the range [0,1].

As shown in Figure 3, for the anchor with a text title as “red bull drink", we use its corresponding product image to be a positive pair. Images from other in-batch products serve as negative pairs. By using the ConVIRT objective, we can adapt both BERT and ViT encoders to better fit the SSL data set. Note that in a contrast to (Zhang et al., 2020), our goal is on using the adapted BERT on text domain in the follow-up fine-tuning task.

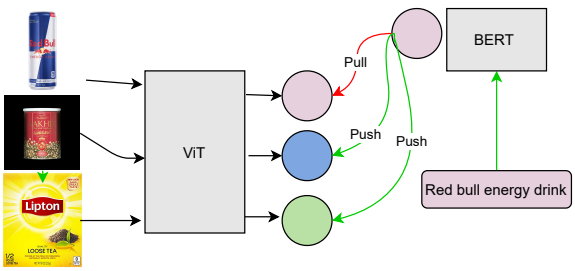

Figure 3: An illustration of ConVIRT. Note that we only showed text→image direction.

Figure 4: Example titles and images from Amazon Review Dataset

4 Experiment

4.1 Setup

Data set: The experimental data consisted of products from Amazon Review Dataset (McAuley et al., 2015; He and McAuley, 2016), focusing on three major product categories, i.e., Automotive, Beauty, and Electronics. Each product contains a text title and a product image that is downloaded from Amazon website from a set of associated images. Figure 4 provides some concrete examples. Our task, a multi-class classification problem, is to predict product categories from their titles. More details of the experimental data are shown in Table 1.

Root genre	# Class	# Data	Len. (ave.)
Automotive	953	200,907	9.91
Beauty	229	199,757	10.26
Electronics	500	107,947	14.88

Table 1: Statistics of the data obtained from Amazon Review Dataset, including the number of labels, the number of instances, and average lengths of text titles

Models: We built IC models by following the paradigm of fine-tuning pre-trained models. Three different pre-trained models were compared.

- origin: using the origin English BERT base model[2].
- SSL SimCSE: the origin BERT model is adapted by using the SimCSE SSL method described in Section 3.1.
- SSL ConVIRT: the origin BERT model is adapted by using the ConVIRT SSL method described in Section 3.2. Note that we used ViT-L-16[3] 16 means that we used 16×16 patches when feeding images.

[2] https://huggingface.co/bert-base-uncased

[3] https://github.com/asyml/vision-transformer-pytorch

Implementation details: For each root genre, from entire data set, we allocate 50% instances for SSL training while keep the remaining 50% instances for fine-tuning, i.e., 30% for Train, 10% for Dev and 10% for Testing. To test model performance on different sizes of fine-tuning data, we then incremental increased fine-tuning set from 5%, 10%, 25%, 50%, 75%, and 100% on the Train set (containing 30% of the entire data size). The fine-tuned IC classifiers were tested on the entire Testing set. Our primary evaluation metric is accuracy. In addition, to make sure all labels can be properly detected, we also evaluate macro-F1 metric. Our models are implemented in PyTorch using 4 GPUs for training and evaluation. At SSL stage, for each dataset, we use the AdamW optimizer (Loshchilov and Hutter, 2017) with an initial learning rate of $1e^{-5}$ and weight decay of $1e^{-8}$. Different from (Zhang et al., 2020), We use Vit as vision encoder and give more weights (higher α) on text-to-image contrastive loss. We set τ=0.1, α=0.75 and train 50 epochs for ConVIRT. The Transformer encoders are followed by a mean-pooling layer and a projection layer with an output dimension $d = 768$. At the fine-tuning stage, we use Adam optimizer with an initial learning rate of $5e^{-5}$, and a weight decay of $1e^{-8}$ for 30 epochs. A fixed batch size of 32 is used on both stages.

4.2 Result

Table 2 reports on accuracy metrics on the three genres. We compared the three fine-tuning methods, including (a) routine fine-tuning using BERT_{origin}, (b) fine-tuning on the BERT self-supervised by using the SimCSE method on texts only, and (c) fine-tuning on the BERT self-supervised by using the ConVIRT method on both texts and product images. We incrementally increased fine-tuning portions from 5% to 100% of the Train set. Comparing with method (a), both methods using SSL methods show performance gains in most of cases. Between the two SSL methods, we found that using cross-modal contrastive learning is more effective, especially when fine-tuning portion is low. Success of a SSL method depends on effectiveness of augmentation operation. Compared to the dropout used in SimCSE, the cross-modal CL losses in ConVIRT may be more genuine and powerful. This is also suggested by recent success of using cross-modal contrastive signals, such as (Radford et al., 2021).

Genre	FT%	origin	SimCSE	ConVIRT
Beauty	5	0.526	0.525	**0.599**
	10	0.577	0.577	**0.623**
	25	0.618	0.616	**0.633**
	50	0.642	0.643	**0.663**
	75	0.659	0.659	**0.670**
	100	0.661	0.669	**0.676**
Auto.	5	0.473	0.499	**0.563**
	10	0.563	0.574	**0.617**
	25	0.648	0.646	**0.668**
	50	0.685	0.690	**0.697**
	75	0.702	0.709	**0.712**
	100	0.718	0.720	**0.721**
Elec.	5	0.319	0.344	**0.483**
	10	0.428	0.441	**0.533**
	25	0.539	0.533	**0.581**
	50	0.575	0.583	**0.601**
	75	0.599	0.604	**0.619**
	100	0.615	0.618	**0.626**

Table 2: Accuracy by fine-tuning BERT models from (a) origin, (b) intra-modal self-supervised by using SimCSE, and (c) cross-modal self-supervised by using ConVIRT on three genres. The underline shows that BERT after SSL cannot show further performance gains. The bold fonts suggest that BERT after SSL can bring further performance gains.

Figure 5 plots macro-F1 values on the two label groups. Based on instance sizes belonging to class labels, we divide all of the labels into two categories, *head* (labels contain 80% of instances) and *tail* (labels only contain 20% of instances). We observed ConVIRT SSL training helps on improve $F-1$ on both head and tail labels. However, SimCSE SSL training does not show noticeable F-1 increases comparing to the baseline of using origin BERT. On tail labels, the gains brought by SSL are quite consistent no matter how large portion of data was used in fine-tuning. This shows that SSL benefit to mitigate long-tailed issue in the item classification task.

5 Discussion

With the success of pre-trained models like BERT (Devlin et al., 2019) on many NLP tasks, fine-tuning BERT models has become an leading approach. To further improve fine-tuned IC model performance, inspired by related works (Fang et al., 2020), self-supervised learning (SSL) is used to adapt original BERT models to better match with the IC task. Regarding the SSL method, we com-

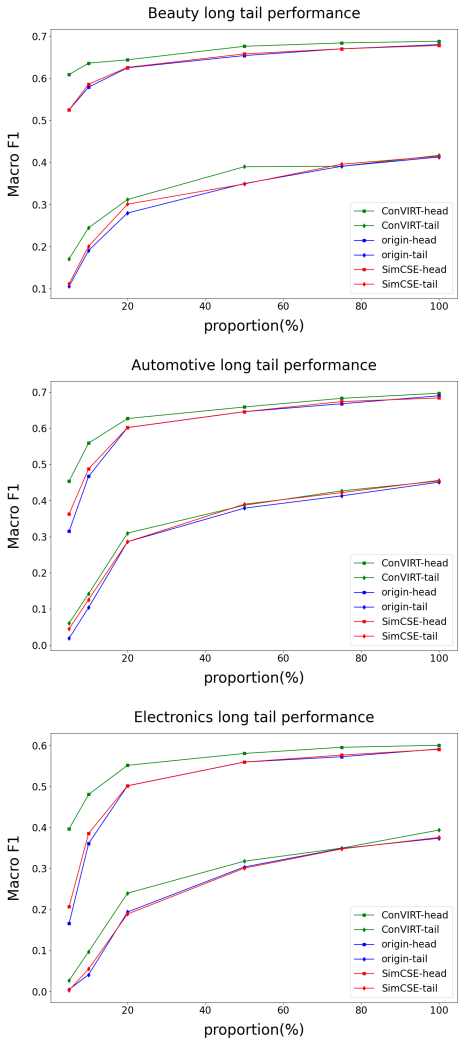

Figure 5: On head class labels (#instance $\geq$ 80%) and tail class labels (#instance $\leq$ 20%), we measured macro-F1 on the three methods, i.e., (a) fine-tuning on BERT$_{origin}$, (b) intra-modal self-supervised using SimCSE, and (c) cross-modal self-supervised using ConVIRT.

pared two approaches. The first approach is unsupervised SimCSE (Gao et al., 2021) that only uses text titles. The SimCSE provided a simple but effective way to generate semantically similar pairs (positive) by feeding the same product title into a BERT model with varying dropout masks. Using other titles in the mini-batch to be negative pairs, the BERT model can be adjusted by minimizing the infoNCE loss (Oord et al., 2018). To improve SSL further, we proposed using the cross-modal contrastive learning to utilize product images to bring additional modeling power to improve our text representations. Following (Zhang et al., 2020), for a product title, we use its associated product image to be a positive pair while other products' images to be negative pairs. The same cross-modal computation was also applied on a product image to provide contrastive signal from the other direction.

Our experiments on the three genres in the Amazon Review Dataset show that both SSL enriched BERT models have higher fine-tuning performance. Between the two SSL methods, the SSL using ConVIRT method is more effective. Moreover, this new way of utilizing images means that we only need process images during the model training stage and can only deploy text-only BERT (enriched by ConVIRT SSL training) in our inference stage. This will dramatically reduce engineering and computation costs compared to the method of deploying a dual-input MIC systems.

One limitation of our research is that we only explored the SimCSE method when using unlabeled text data. It is possible that other semi-supervised learning methods like UDA (Xie et al., 2019) may help on improving the final fine-tuning performance. We will leave the exploration on more semi-supervised learning methods in future. In addition, there are several directions to extend the current work in the future, including (1) improving contrastive learning based SSL, for example using nearest neighbor to get better positive pairs similar to (Li et al., 2021), and (2) improving our algorithms to better address training set bias such as high label noise and very imbalanced data set in real IC data.

References

Ye Bi, Shuo Wang, and Zhongrui Fan. 2020. A Multimodal Late Fusion Model for E-Commerce Product Classification. *arXiv preprint arXiv:2008.06179*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

V. Chordia and B.G. Vijay Kumar. 2020. Large Scale Multimodal Classification Using an Ensemble of Transformer Models and Co-Attention. In *Proc. SIGIR'20 e-Com workshop*.

H. Chou, Y.H. Lee, L. Chen, Y. Xia, and W.T. Chen. 2020. CBB-FE, CamemBERT and BiT Feature Extraction for Multimodal Product Classification and Retrieval. In *Proc. SIGIR'20 e-Com workshop*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language

Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

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

Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive self-supervised learning for language understanding. *arXiv preprint arXiv:2005.12766*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, and Yixing Xu. 2020. A Survey on Visual Transformer. *arXiv preprint arXiv:2012.12556*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517.

Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1459, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Peng Su, Yifan Peng, and K Vijay-Shanker. 2021. Improving bert model using contrastive learning for biomedical relation extraction. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 1–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, \Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training. *arXiv preprint arXiv:1904.12848*.

Tom Zahavy, Alessandro Magnani, Abhinandan Krishnan, and Shie Mannor. 2016. Is a picture worth a thousand words? A Deep Multi-Modal Fusion Architecture for Product Classification in e-commerce. *arXiv preprint arXiv:1611.09534*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.