Multimodal Learning with Online Text Cleaning for E-commerce Product Search

Zhizhang Hu[1,†], Shasha Li[2], Ming Du[3], Arnab Dhua[4] and Douglas Gray[5]

[1]University of California, Merced
[2]Amazon Visual Shopping
[3]Amazon Visual Shopping
[4]Amazon Visual Shopping
[5]Amazon Visual Shopping

Abstract

Vision-language transformer models play a pivotal role in e-commerce product search. When using product description (e.g. product title) and product image pairs to train such models, there are often non-visual-descriptive text attributes in the product description, which makes the visual textual alignment challenging. We introduce MultiModal Learning with online Token Pruning (MML-TP). MML-TP leverages token pruning, conventionally used for computational efficiency, to perform online text cleaning during multimodal model training. Evaluation on the e-commerce dataset comprising over 710k unique Amazon products validates that refining text tokens enhances the paired image branch's training, which leads to significantly improved visual search performance.

Keywords

Token pruning, multimodal learning, product search

1. Introduction

Multimodal transformer models have been widely adopted in e-commerce product search, including but not limited to caption-to-image search, image-to-image search, and multimodal-to-image search [1, 2, 3, 4, 5]. The success of applying multimodal models in e-commerce product search can be attributed to its strength in understanding vision and language representations of product contents. One of the key factors for training an effective vision-language multimodal model relies on the alignment of image-text pairs in the dataset. In practice, the training dataset is usually collected in an automatic fashion with limited manual cleaning or annotation. As a result, the alignment between text and image is far from ideal.

This misalignment issue is bi-directional: it could be the case that not all the text content is reflected by the paired image, or the corresponding text does not fully describe the image content. In e-commerce applications, the former issue is more common [7, 8] because sellers are inclined to include as many as product attributes in the product title in order to promote their listings. In the example shown in Figure 1, most phrases in the product title are not visual-

eCom'24: ACM SIGIR Workshop on eCommerce, July 18, 2024, Washington, DC, USA
[†]This work is done during the internship at Amazon.
zhu42@ucmerced.edu (Z. Hu); shashli@amazon.com (S. Li); mingdu@amazon.com (M. Du); aduha@amazon.com (A. Dhua); douggray@amazon.com (D. Gray)


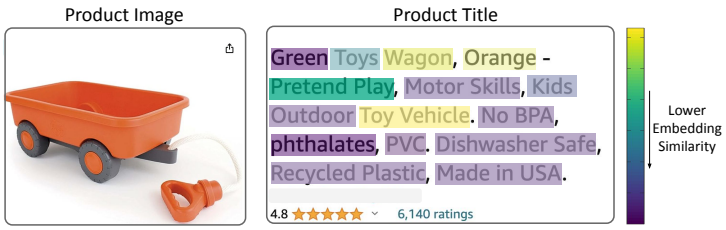

Figure 1: Example of a product's image-text pair from an e-commerce website. Phrases in the product title are color-coded by their embedding similarity to the image embedding. Both image and text embeddings are generated by the BLIP-2 [6] model.

descriptive – *how can you tell it is "No BPA" by looking at the image*? Such non-visual-descriptive phrases have significantly lower similarity to the image compared to other phrases. We assume cleaning out such phrases could enhance multimodal alignment learning and lead to better image embedding models.

In this paper, we introduce MultiModal Learning with online Token Pruning (MML-TP), a simple yet effective method for training the multimodal transformer models with noisy e-commerce image-text training data. The method leverages token pruning technique, which was conventionally used for improving models' computational efficiency by discarding unimportant tokens [9, 10], to perform online text cleaning during multimodal model training. The key idea is that given that each phrase has a different importance in describing the image, we can let the model learn to remove unimportant/unrelated tokens alongside its original multimodal training task. As a result, the model can be trained with implicitly-cleaned image-text pairs.

Given the scarcity of publicly available e-commerce datasets, we establish a benchmark multimodal e-commerce dataset based on the uni-modal Amazon ESCI dataset [11] with over 710k unique products sold on Amazon.com. Extensive experiements on ALBEF [12] and CLIP [13] frameworks validates the effectiveness of MML-TP. MML-TP boosts the image retrieval performance by over **5** percentage point measured by Recall@1.

2. Related Work

The success of large-scale transformer-based pre-training in the field of Natural Language Processing [14] has boosted research works in vision-language pre-training.

Vision-language transformer models are trained on large-scale image-text pairs and learn a joint vision-language embedding space for various downstream tasks. CLIP model [13] leverages a broader source of supervision from text to train a predictive model that aligns text with image, resulting in a task-agnostic model comparable to task-specific supervised models. ALIGN [15] scales up the CLIP model with a noisy dataset without expensive filtering or post-processing steps that cover more than one billion image alt-text pairs. CLIP and ALIGN show promising results in vision-based downstream tasks, however, they ignore the interaction between two modalities and vision-language downstream tasks.

Later studies propose to learn joint embeddings of image contents and natural language during pre-training, like OSCAR [16], UNIMO [17] and UNITER [18]. These works use an

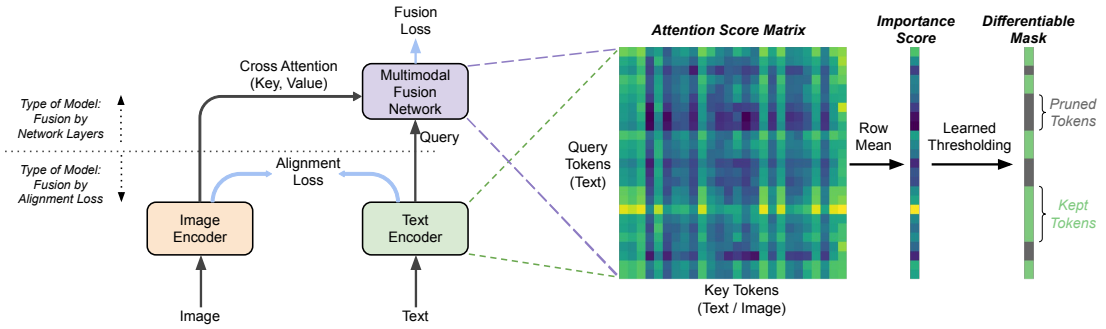

Figure 2: Overview of the MML-TP method. It is flexible to work with the self-attention matrix in the text encoder or cross-attention matrix if the model has fusion network layers. It takes the attention score matrix to calculate the importance score for each query text token. The unimportant tokens are masked following a learnable thresholding mechanism.

object detector backbone to capture vision features first, then a transformer-based model is applied to the concatenated vision and text features to learn joint embeddings. ViLT [19] further breaks through the regional feature from convolutional networks and adopts vision transformer [20] to fuse the whole global image feature with natural languages. ALBEF [12] and TCL [21] further exploit contrastive loss functions to align image and text features before modeling their joint embeddings, increasing the interaction between two modalities and achieving a state-of-the-art performance (SOTA).

3. Methodology

In a nutshell, our method masks text tokens based on token importance derived from the attention score matrix. We present how we define token importance in Section 3.1 and how to mask text tokens based on their importance scores in Section 3.2. The overview of the two components are illustrated in Figure 2.

3.1. Token Importance

For vision-language model learning frameworks like ALBEF, cross attention is used to directly measure the relevance between image and text tokens. For frameworks like CLIP with no relevance measurement between image and text tokens, self attention in the text branch measures the importance of different text tokens. Given that the learning objective aligns text embedding with image embedding, the text token attentions are learnt guided by visual features. we hypothesize that analyzing the self-attention patterns within the text encoder reveals fine-grained textual dependencies and also tokens' importance in grounding visual content. Therefore, we propose to use the attention score matrix from both self attention and cross attention to quantify the importance of text tokens.

Given an input query sequence $\mathrm{x} \in \mathbb{R}^{m \times n}$ with m tokens, and input key sequence $\mathrm{z} \in \mathbb{R}^{k \times l}$

with k tokens, the attention score matrix is calculated as:

$$\text{Attn}(\mathrm{x}, \mathrm{z}) = \frac{\mathrm{x}\mathrm{W}_q\mathrm{W}_k^T\mathrm{z}^T}{\sqrt{d}}, \tag{1}$$

where $\mathrm{W}_q \in \mathbb{R}^{n\times d}$ and $\mathrm{W}_k \in \mathbb{R}^{l\times d}$ are trainable weight matrices. For self-attention, we have $m = k$ and $n = l$. This attention score matrix measures each input query token's pairwise importance on every key token.

Note that the text tokens are used as query tokens in cross-attention. Therefore we could aggregate the attention scores along key tokens to define text token importance score following [22, 23, 10]. However, we find empirically that using key [CLS] token only gives better performance. In cross-attention, the image [CLS] token encodes aggregated visual concepts. In self-attention, the text [CLS] represents the overall linguistic context. Attending to these consolidated representations provides a less noisy measure compared to all the key tokens. Therefore, we calculate importance score of the i-th query token as the average of its attention to the key [CLS] token from all heads as shown blow.

$$s(\mathrm{x}_i) = \frac{1}{H}\sum_{h=1}^{H}\text{Attn}_h(\mathrm{x}_i, \mathrm{z}_0), \tag{2}$$

where Attn_h is the attention matrix for the $h-$th head and we assume the [CLS] token is in the first (0-th) position of the key token sequence.

3.2. Pruning with Learned Threshold

Given each query token's importance score, MML-TP prunes unimportant tokens by comparing the score with a threshold τ. This process allows the model to discard noisy tokens that contribute negatively or less to multimodal alignment and fusion. However, setting the value of τ' s is a nontrivial task. The appropriate threshold may differ between tasks and datasets. The threshold may also vary across transformer layers, as deeper layers capture higher-level concepts where fewer tokens may be relevant. Therefore, we model τ as a learnable parameter, allowing it to adapt to the specific requirements of each task, data, and layer.

Inspired by Tempered Sigmoid Activations [24], the differentiable pruning mask defined for the i-th query token (x_i) in the l-th attention layer is defined as:

$$M_l(\mathrm{x}_i) = \sigma(\frac{s_l(\mathrm{x}_i) - \tau_l}{T}), \tag{3}$$

where T is the temperature parameter and τ_l is the threshold learnt for the l-th layer. To mask the text tokens, we update query token embedding features by multiplying them with their associated mask score which is between 0 and 1. For tokens whose importance scores are smaller than the threshold, their mask score is close to zero and hence they will not become major information sources in succeeding layer.

To encourage token pruning, we adopt pruning loss[10] as an additional training objective.

$$\mathcal{L}_{Prune} = \frac{1}{N}\sum_{l=1}^{L}\frac{\|M_l(\mathrm{x})\|_1}{d_l^Q}, \tag{4}$$

where d_l^Q is the sequence length of the Query at layer l. The scaling factor d_l^Q is designed for models with dynamic Query length, which is helpful for normalizing the mask's L1 norm to a unified scale. Intuitively, when more tokens are situated close to the threshold, the gradient $\frac{\mathcal{L}_{Prune}}{d\tau_l}$ becomes larger. Consequently, this causes an increase in the threshold value, resulting in the pruning of a greater number of tokens that are proximate to the threshold boundary. Generally, for models with original training objectives $\mathcal{L}_{Model}$, the updated training objective is:

$$\mathcal{L} = \mathcal{L}_{Model} + \lambda \cdot \mathcal{L}_{Prune}, \tag{5}$$

where λ is the regularization parameter to control the aggressiveness of pruning.

4. Experiments

We in this section first describe the evaluation dataset, implementation details and evaluation metrics. We consider the evaluation of MML-TP in two application scenarios. We can directly use MML-TP to finetune a public available vision-language model. Or if we already have a model finetuned on e-commerce dataset, which is often the case, we can further finetune the model with few epochs using MML-TP on the same finetune dataset and achieve better product search performance. We evaluate both scenarios with two models, that is, CLIP and ALBEF. We also present ablation study on the two different token importance score definitions.

4.1. Dataset

Public multimodal e-commerce datasets are not suitable for our evaluation. Fashion-Gen [25], Fashion 200k [26], Shopping100 [27], and FashionIQ [28] focus on the fashion domain, instead of general purpose product search. M5 Product Data [7] and Product 1M Data [29] are in the form of Chinese product titles, as the unique characteristics of the Chinese language and its tokenizing effect on the proposed MML-TP is out of the scope of this work. We therefore establish a new benchmark multimodal e-commerce dataset based on Amazon ECSI dataset [11], which is a uni-modal dataset for product shopping queries. We add product catalog images to Amazon ESCI dataset. After removing products that are no longer available or have less than two images, the dataset covers over 710k products sold on Amazon.com. For each product, we have a product title, a main image, and multiple (1 to 10) auxiliary images. We reserve 80k products for test where 186k image-image pairs are generated for visual search (image to image retrieval) evaluation. The other 630k products are used for training where 858k image-text pairs are generated for multimodal learning. This dataset covers most common product categories, including but not limited to *Hardlines* (e.g., electronics, furniture, ...), *Softlines* (apparel, shoes,..), *Consumables* (personal care, pantry, ...), etc.

4.2. Implementation Details and Metric

All experiments were conducted using 8 NVIDIA A100 GPUs, utilizing the PyTorch deep learning framework [30] and the Ray distributed computing framework [31]. Both the CLIP and ALBEF models employ a standard ViT-B/16 [20] vision encoder with 12 layers and 86M parameters. CLIP's text encoder is a 12-layer transformer with 63M parameters, while ALBEF's text and

fusion encoders are built on a 6-layer transformer, totaling 124M parameters. In token pruning, layer-wise thresholds are initialized with linearly rising values, ending with a fixed threshold of 0.01 at the final layer. The temperature parameter T is set at $1e^{-4}$. From empirical exploration, a pruning loss's regularization parameter λ of 0.1 is found suitable for all experiments.

We adopt the standard evaluation metric in image to image retrieval, i.e., Recall@K (denoted as R@K), which is defined as the proportion of test queries for which the correct targets are successfully identified within the top-K retrieved samples [32]. Unless specified, the unit in tables of retrieval performance is the percentage (%.)

4.3. MML-TP for Public Model Finetune

	R@1	R@5	R@10
Pre-train	42.56	51.29	56.62
Standard finetune	53.59	63.24	69.03
MML-TP finetune	**55.21**	**65.13**	**70.97**
↑	1.62	1.89	1.94

	R@1	R@5	R@10
Pre-train	38.60	47.40	53.03
Standard finetune	51.68	62.27	68.68
MML-TP, CA only	54.59	65.09	71.36
MML-TP finetune	**57.06**	**67.54**	**73.74**
↑	5.38	5.27	5.06

Table 1
Left table for CLIP results and right table for ALBEFL results. CLIP finetune with MML-TP leads to 1-2pps recall increase compared to standard finetune. ALBEF fintune with MML-TP leads to 5+pps recall increase. CA only in the right table means token pruning applied only for cross-attention layers.

We finetune CLIP with 100 epochs with a batch size of 1360, using the AdamW optimizer [33] with a weight decay of 0.02. The learning rate was initialized at $5e^{-6}$, warmed up to $2e^{-5}$ after 10 epochs, and then decreased to $5e^{-6}$ using the cosine decay strategy. Evaluation results in Table 1 left part shows that MML-TP finetune improves CLIP image to image retrieval performance by 1-2 percentage points (pps) compared to standard finetune. This implies that in vision-language frameworks where image and text tokens are not attended to each other directly, token pruning in text self-attention layers still improves multimodal learning.

We finetune ALBEF with its pre-training configuration and adjust the batch size to 320 due to memory limitation. We use MML-TP in two different setups, token pruning only on cross-attention layers and token pruning on all the attention layers. As shown in 1 right part, MML-TP with cross attention layers improves standard finetune method by about 3pps and MML-TP with all attention layers further improves the performance by about 2pps. The results not only validate the effectiveness of MML-TP but also imply the importance of token pruning in self-attention layers.

4.4. MML-TP for Second-Stage Fintune

When there exists a production model which is finetuned from public model with e-commerce dataset. It takes time and resources to repeat the finetune process with MML-TP method. We therefore propose to do a second-stage finetune, where only few-epoch MML-TP finuetune is conducted based on the finetuned model. Note that it's possible that standard second-stage

	R@1	R@5	R@10
FT1	53.59	63.24	69.03
FT2, standard	54.55	64.29	70.08
FT2, MML-TP	**55.95**	**65.81**	**71.57**
↑	2.36	2.57	2.54

	R@1	R@5	R@10
FT1	51.68	62.27	68.68
FT2, standard	53.83	64.44	70.83
FT2, MML-TP	**56.75**	**67.39**	**73.65**
↑	5.07	5.12	4.97

Table 2
Left table for CLIP results and right table for ALBEFL results. Second-stage MML-TP fintune on CLIP improves image retrieval performance by about 1.5pps. Second-stage MML-TP fintune on ALBEF improves image retrieval performance by about 5pps.

finetune improves the first-stage finetune performance because the latter is under-fitting. We therefore provide standard second-stage finetune results as baseline.

Results in Table 2 shows that second-stage finetune with MML-TP improves CLIP first-stage finetune retrieval performance by about 2.5pps. The improvement is even more significant, about 5pps, for ALBEF model, probably because we have token pruning for both text self attention layers and cross attention layers in ALBEF while it's only text self attention layers for CLIP.

4.5. Ablation Study

	R@1	R@5	R@10
Average over tokens	54.92	64.75	70.65
[CLS] token only, ours	**55.21**	**65.13**	**70.97**

	R@1	R@5	R@10
Average over tokens 3	56.82	67.46	73.70
[CLS] token only, ours	**57.06**	**67.54**	**73.74**

Table 3
Left table for CLIP results and right table for ALBEFL results. Token importance based on key [CLS] token scores performs better than averaging the attention scores among all the key tokens.

We do ablation study for the two ways of calculating token importance score. One is to follow [22, 23, 10] and calculate the query token importance by using the average attention scores over all the key tokens. The other one is proposed by us, that is, to use only the key [CLS] token instead of all the key tokens. We experiment with the two methods to finetune the public CLIP and ALBEF models. Results in Table 3 shows consistent better performance achieved by using only the [CLS] token. Our explanation is that attending to consolidated [CLS] token representation provides a less noisy importance measure compared to attending to all the key tokens.

4.6. Grad-CAM Visualization

To get more insights on how the token pruning mask works for vision-language transformer models. We visualize the attention map on the product image associated with each word in

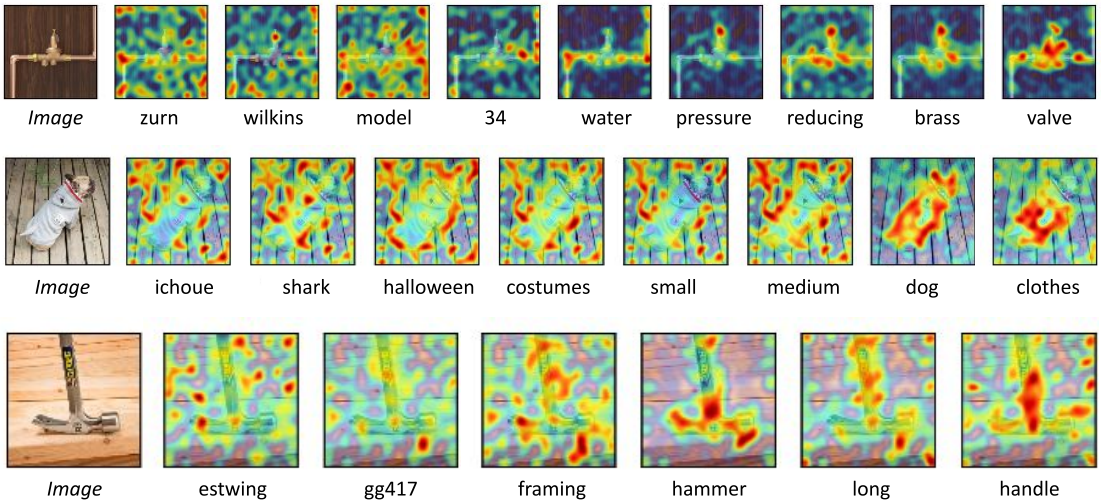

Figure 3: Grad-CAM visualizations on the cross-attention maps of the MML-TP fintuned ALBEF model, corresponding to individual words in the product title.

the product title calculated by the MML-TP finetuned ALBEF model. We do Grad-CAM [34] visualization on the fusion encoder's third layer following [12].

The attention maps reveal distinct patterns of focus. Words that are visually descriptive, such as "valve," "dog," and "handle," exhibit concentrated attention areas. This suggests that the model emphasizes regions in the image that correspond to these descriptive terms. In contrast, brand names or words that lack a direct visual counterpart in the image, like "zurn," "ichoue," and "estwing," show diffused and scattered attention patterns.

The difference in attention distribution demonstrates the model's ability to discern between text tokens. The model diminishes its attention toward text tokens that are potentially noisy or less relevant while honing in on tokens that provide meaningful visual cues. Such behavior aligns with our fundamental hypothesis and motivation: to prioritize informative text tokens and reduce the influence of extraneous ones. This selective attention mechanism not only highlights the model's capability to differentiate between visually grounded and non-grounded textual information but also provides a rationale for our token pruning approach.

5. Conclusion

In this paper, we address the challenge of noisy image-text pair alignment in e-commerce datasets and propose MML-TP. Leveraging token pruning, MML-TP facilitates multimodal transformer model learning with cleaner image-text pairings. By pruning noisy text tokens implicitly, MML-TP denoises the text branch and strengthens the vision encoder, leading to a more efficient multimodal model for e-commerce applications. Our evaluation with a large-scale e-commerce dataset has demonstrated MML-TP's effectiveness in improving visual search performance. Also, the proposed method is flexible and compatible with models like CLIP that rely on alignment loss and those like ALBEF with fusion networks.

References

[1] H. Ma, H. Zhao, Z. Lin, A. Kale, Z. Wang, T. Yu, J. Gu, S. Choudhary, X. Xie, Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18051–18061.

[2] B. Chen, L. Jin, X. Wang, D. Gao, W. Jiang, W. Ning, Unified vision-language representation modeling for e-commerce same-style products retrieval, arXiv preprint arXiv:2302.05093 (2023).

[3] X. Zheng, Z. Wang, S. Li, K. Xu, T. Zhuang, Q. Liu, X. Zeng, Make: Vision-language pre-training based product retrieval in taobao search, in: Companion Proceedings of the ACM Web Conference 2023, 2023, pp. 356–360.

[4] Y. Zhu, H. Zhao, W. Zhang, G. Ye, H. Chen, N. Zhang, H. Chen, Knowledge perceived multi-modal pretraining in e-commerce, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 2744–2752.

[5] Y. Tang, X. Xiong, S. Sun, B. Cui, Y. Zheng, H. Tang, Tmml: Text-guided mulimodal product location for alleviating retrieval inconsistency in e-commerce, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, pp. 3275–3279.

[6] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, arXiv preprint arXiv:2301.12597 (2023).

[7] X. Dong, X. Zhan, Y. Wu, Y. Wei, M. C. Kampffmeyer, X. Wei, M. Lu, Y. Wang, X. Liang, M5product: Self-harmonized contrastive learning for e-commercial multi-modal pretraining, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21252–21262.

[8] F. Liu, D. Chen, X. Du, R. Gao, F. Xu, Mep-3m: A large-scale multi-modal e-commerce product dataset, Pattern Recognition 140 (2023) 109519.

[9] Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, C.-J. Hsieh, Dynamicvit: Efficient vision transformers with dynamic token sparsification, Advances in neural information processing systems 34 (2021) 13937–13949.

[10] S. Kim, S. Shen, D. Thorsley, A. Gholami, W. Kwon, J. Hassoun, K. Keutzer, Learned token pruning for transformers, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 784–794.

[11] C. K. Reddy, L. Màrquez, F. Valero, N. Rao, H. Zaragoza, S. Bandyopadhyay, A. Biswas, A. Xing, K. Subbian, Shopping queries dataset: A large-scale ESCI benchmark for improving product search (2022). arXiv:2206.06588.

[12] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Advances in neural information processing systems 34 (2021) 9694–9705.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[14] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[15] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR, 2021, pp. 4904–4916.

[16] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al., Oscar: Object-semantics aligned pre-training for vision-language tasks, in: European Conference on Computer Vision, Springer, 2020, pp. 121–137.

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

[18] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer, 2020, pp. 104–120.

[19] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.

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

[21] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, J. Huang, Vision-language pre-training with triple contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15671–15680.

[22] S. Goyal, A. R. Choudhury, S. Raje, V. Chakaravarthy, Y. Sabharwal, A. Verma, Power-bert: Accelerating bert inference via progressive word-vector elimination, in: International Conference on Machine Learning, PMLR, 2020, pp. 3690–3699.

[23] G. Kim, K. Cho, Length-adaptive transformer: Train once with length drop, use anytime with search, arXiv preprint arXiv:2010.07003 (2020).

[24] N. Papernot, A. Thakurta, S. Song, S. Chien, Ú. Erlingsson, Tempered sigmoid activations for deep learning with differential privacy, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 9312–9321.

[25] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, L. Shao, Kaleido-bert: Vision-language pre-training on fashion domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12647–12657.

[26] X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, L. S. Davis, Automatic spatially-aware fashion concept discovery, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 1463–1471.

[27] K. E. Ak, J. H. Lim, J. Y. Tham, A. A. Kassim, Efficient multi-attribute similarity learning towards attribute-based fashion search, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2018, pp. 1671–1679.

[28] H. Wu, Y. Gao, X. Guo, Z. Al-Halah, S. Rennie, K. Grauman, R. Feris, Fashion iq: A new dataset towards retrieving images by natural language feedback, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11307–11317.

[29] X. Zhan, Y. Wu, X. Dong, Y. Wei, M. Lu, Y. Zhang, H. Xu, X. Liang, Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp.

11782–11791.

[30] A. Paszke, S. Gross, S. Chintala, Y. Wei, Z. Wang, J. Turner, A. Desmaison, L. Antiga, J. Donahu, Pytorch: An imperative style, high-performance deep learning library, Advances in Neural Information Processing Systems 32 (2019).

[31] P. Moritz, R. Nishihara, S. Wang, A. Tumanov, R. Liaw, E. Liang, M. Elibol, Z. Yang, W. Paul, M. I. Jordan, et al., Ray: A distributed framework for emerging {AI} applications, in: 13th USENIX symposium on operating systems design and implementation (OSDI 18), 2018, pp. 561–577.

[32] Y. Chen, L. Bazzani, Learning joint visual semantic matching embeddings for language-guided retrieval, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16, Springer, 2020, pp. 136–152.

[33] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[34] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.