MSRA: A Multi-Aspect Semantic Relevance Approach for E-Commerce via Multimodal Pre-Training

Hanqi Jin[*]
Alibaba Group
Hangzhou, China
jinhanqi.jhq@alibaba-inc.com

Jiwei Tan[*]
Alibaba Group
Hangzhou, China
jiwei.tjw@alibaba-inc.com

Lixin Liu
Alibaba Group
Hangzhou, China
llx271805@alibaba-inc.com

Lisong Qiu
Shaowei Yao
Alibaba Group
Hangzhou, China
qiulisong.qls@alibaba-inc.com
yaoshaowei@alibaba-inc.com

Xi Chen
Xiaoyi Zeng
Alibaba Group
Hangzhou, China
gongda.cx@taobao.com
yuanhan@alibaba-inc.com

ABSTRACT

To enhance the effectiveness of matching user requests with millions of online products, practitioners invest significant efforts in developing semantic relevance models on large-scale e-commerce platforms. Generally, such semantic relevance models are formulated as text-matching approaches, which measure the relevance between users' search queries and the titles of candidate items (i.e., products). However, we argue that conventional relevance methods may lead to sub-optimal performance due to the limited information provided by the titles of candidate items. To alleviate this issue, we suggest incorporating additional information about candidate items from multiple aspects, including their attributes and images. This could supplement the information that may not be fully provided by titles alone. To this end, we propose a multi-aspect semantic relevance model that takes into account the match between search queries and the title, attribute and image information of items simultaneously. The model is further enhanced through pre-training using several well-designed self-supervised and weakly-supervised tasks. Furthermore, the proposed model is fine-tuned using annotated data and distilled into a representation-based architecture for efficient online deployment. Experimental results show the proposed approach significantly improves relevance and leads to considerable enhancements in business metrics.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

E-commerce, semantic matching, pre-trained multimodal model

[*] Both authors are the corresponding authors.

CIKM '23, October 21–25, 2023, Birmingham, United Kingdom
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615224

ACM Reference Format:
Hanqi Jin, Jiwei Tan, Lixin Liu, Lisong Qiu, Shaowei Yao, Xi Chen, and Xiaoyi Zeng. 2023. MSRA: A Multi-Aspect Semantic Relevance Approach for E-Commerce via Multimodal Pre-Training. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615224

1 INTRODUCTION

E-commerce platforms like Amazon, Taobao, and JD.com invest considerable resources in optimizing their search engines to facilitate an efficient matching of user requests with millions of online products (items) [6]. Commercial e-commerce search engines are usually optimized to boost users' engagement and conversion, possibly at the cost of relevance in some cases [1]. Nevertheless, displaying irrelevant items can adversely affect the user's experience. Precisely measuring the relevance between queries and items is crucial for e-commerce search engines.

In e-commerce, relevance is usually defined as a semantic matching task between the query and item title [6, 17, 19], which is an important research topic in both industry and academia. However, according to statistical data from our scenario, over 60% of bad cases are due to missing crucial information in item titles caused by length limitation or seller negligence. For example, in Figure 1, the search query is for an "XXL animal T-shirt", but the item title doesn't include the size and pattern. In traditional query-title matching-based relevance models, items without size and pattern information in their titles won't be displayed to customers, negatively impacting both their experience and the merchants' interests. In fact, e-commerce platforms have defined multi-aspect descriptions, such as attributes and images, which are provided by sellers to detail the products from various perspectives.

To address the aforementioned issues, we propose a Multi-aspect Semantic Relevance Approach (MSRA) that considers the match between search queries and the title, attribute and image information of items simultaneously. Search queries, item titles, attributes (structured information in the form of key-value pairs), and images can be viewed as three different modalities: text, attribute, and image. Firstly, we design a novel unified architecture to encode the multi-modal information. As human-annotated relevance data is

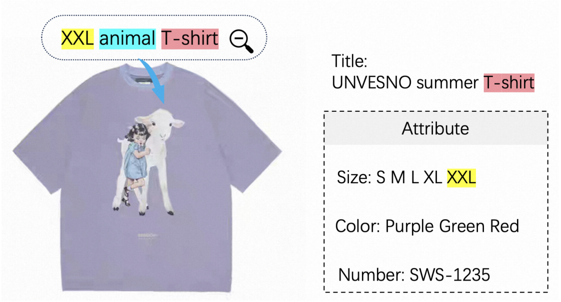

Figure 1: A search case in e-commerce where the item title is inadequate to comprehensively match the search query, but the attribute and image provide additional complementary information.

typically scarce and expensive, we use popular pre-training techniques [3, 10, 11, 15, 18] to learn better multi-modal representations with several well-designed self-supervised and weakly-supervised tasks. Subsequently, we fine-tune the pre-trained model on a small annotated relevance dataset and distill it into a representation-based model for online deployment.

We evaluate the proposed multimodal relevance model on an annotated relevance test set and find that it significantly outperforms several strong baselines. Furthermore, we conduct online experiments with the representation-based model on the real-world e-commerce platform. The results demonstrate that the proposed approach substantially improves relevance and Gross Merchandise Value (GMV).

In summary, we make the following contributions in this paper:

- We suggest incorporating attribute and image information into the relevance model as a solution to address the inadequate match between search queries and item titles.
- We propose a novel multi-aspect semantic relevance approach that takes into account the match between queries and the title, attribute, and image information of products simultaneously. Furthermore, we enhance this approach through pre-training that includes several well-designed self-supervised and weakly-supervised tasks.
- We deploy the proposed model for efficient online serving. Both offline and online experiments demonstrate significant improvement in relevance and substantial enhancement in GMV, compared to the state-of-the-art methods.

2 METHODOLOGY

2.1 Model Architecture

In this section, we introduce the architecture of Multi-aspect Semantic Relevance Approach (MSRA). As illustrated in Figure 2, the query and item are represented by multiple modalities, such as texts (queries and item titles), attributes, and images. To incorporate these diverse modalities into a unified model, we first pre-process and integrate them into a multi-modal embedding sequence. This sequence is used as input for a unified Transformer encoder [16].

2.1.1 Textual Representation. We first tokenize the raw texts of the query and title into sequences. Specifically, the query is tokenized as $Q = [w_1, ..., w_{M_1}]$, and the title is tokenized as $T = [w_1, ..., w_{M_2}]$. Each token w_i is converted into a vector representation e_{w_i} using the word embedding lookup. We assign a position embedding to indicate the position of each word in the query and title. We use the type embedding to differentiate between the query and title sequences, marking the query with "QUERY" and the title with "TITLE". The input representations of the texts are the sum of the token embeddings, position embeddings, and type embeddings. Additionally, we also add learnable special embeddings $\boldsymbol{q}_{[CLS]}$ and $\boldsymbol{t}_{[CLS]}$ to the beginning of the query and title sequences, respectively.

2.1.2 Attribute Representation. Attributes describe the item from fine-grained perspectives, such as size and color, as illustrated in Figure 1. Typically, attributes are represented as a series of property-value pairs. To encode the structural knowledge of the attributes, we propose a novel encoding method that treats each property as a special type, similar to "QUERY" or "TITLE". We first establish the explicit embedding $\boldsymbol{e}_{p_i}$ for each property p_i by maintaining a property embedding matrix. The value of the attribute is then treated as raw text and tokenized into a token sequence $[\boldsymbol{w}_{i,1}, \ldots, \boldsymbol{w}_{i,L_i}]$. We assign the position embedding to indicate the position of the token. We compute the input representations by summing the value embedding, position embedding, and property embedding. Finally, we concatenate all attribute representations with a special embedding $\boldsymbol{a}_{[CLS]}$ added at the beginning.

2.1.3 Visual Representation. Inspired by previous work on visual pre-training[2, 8, 12, 14], we reshape the input image of the item into a sequence of flattened 2D patches V, where H, W, and C represent the image height, width, and channel, respectively. P and N indicate the size and number of patches, respectively. We use ResNet [4] as the backbone network to extract 2048-D features $\boldsymbol{e}_{v_i}$ from each patch v_i. Position embedding is assigned to indicate the position of the patch in the original image. To differentiate the image from other inputs, we use the type embedding to mark the image with "IMAGE". We compute the image representations by summing the patch features, position embeddings, and type embeddings. Finally, we prepend a learnable special embedding $\boldsymbol{v}_{[CLS]}$ to the sequence.

2.1.4 Unified Multimodal Encoder. We utilize a unified Transformer encoder to encode the query, title, attribute, and visual representations. As shown in Figure 2, we concatenate them into a multimodal embedding sequence with a special embedding token $\boldsymbol{m}_{[CLS]}$. This concatenated multi-modal embedding sequence is fed into the Transformer encoder, allowing for cross-modal attention between the query, title, attribute, and image representations for the modality fusion. The final output from the multimodal Transformer encoder is a list of hidden states. For convenience, we denote the corresponding output of $\boldsymbol{m}_{[CLS]}$, query, title, attribute, and image as $h_{\boldsymbol{m}_{[CLS]}}$, h_q, h_t, h_a, and h_v respectively.

2.2 Model Pre-Training

In e-commerce, semantic relevance is different from click-through rate prediction in that no direct training signal is available [19]. Most previous works [6, 17, 19] attempt to learn relevance models from user click-through data that are cheap and abundant. Unfortunately, click behavior is noisy and misleading, which is affected by not only relevance but also factors including price, image and

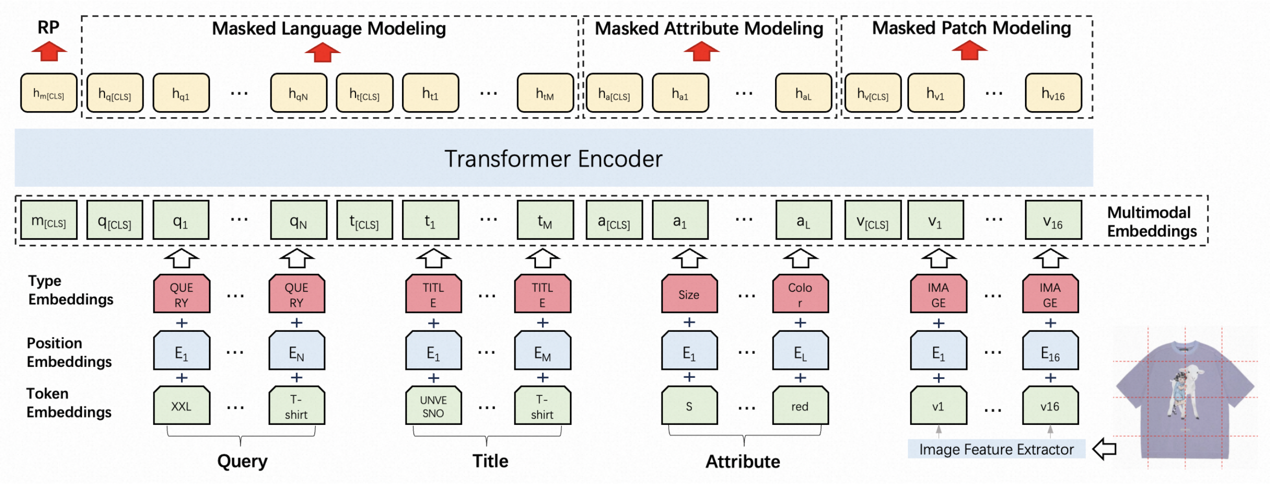

Figure 2: The overview of the proposed MSRA model.

attractive titles. It is challenging but valuable to learn relevance models from click-through data. In this paper, we elaborately design several self-supervised and weakly-supervised tasks to pre-train the model from the click-through data.

2.2.1 Masked Language Modeling (MLM) & Masked Patch Modeling (MPM). Inspired by BERT [3], we employ Masked Language Modeling (MLM) to self-supervise the learning of query and title representations. To achieve this, we randomly mask some tokens in the query and title, and the model is trained to predict these masked tokens t_m from all the unmasked tokens $t_{\backslash m}$, attributes A the image patches V.

$$\mathcal{L}_{\mathrm{T}} = -\mathbb{E}_{\{Q,T,A,V\}\in\mathcal{D}} \log P_\theta\left(t_m \mid t_{\backslash m}, Q, A, V\right) \tag{1}$$

To enhance the visual representation of image patches, we adopt Masked Patch Modeling (MPM) to supervise the learning of patch representations, as detailed in [9]. This involves sampling image patches and masking the patch embeddings with a probability of 20%. The masked patch embeddings are replaced by zeros, and the model is trained to reconstruct the masked patches v_m based on the query Q, title T, attribute A, and the remaining visual regions $v_{\backslash m}$.

$$\mathcal{L}_V = -\mathbb{E}_{\{Q,T,A,V\}\in\mathcal{D}} f_\theta\left(v_m \mid v_{\backslash m}, Q, T, A\right) \tag{2}$$

As the patch features are high-dimensional and continuous, we utilize the feature regression objective to regress the contextualized patch representations h_{v_i} to its patch features v_i, which can be formulated as:

$$f_\theta\left(v_m \mid v_{\backslash m}, Q, T, A\right) = \left\| r\left(h_{v_m}\right) - v_m \right\|^2 \tag{3}$$

where r indicates a fully connected layers to convert h_{v_i} into a vector of the same dimension as v_i.

2.2.2 Masked Attribute Modeling (MAM). In this section, we propose a novel Masked Attribute Modeling (MAM) task to supervise the learning of attribute representations. We randomly mask some properties and values, and predict these masked properties and values from other attributes together with the query, title, and image. Masked property prediction promotes the model to infer the property from the corresponding value, thereby learning better representations for the attribute. Masked value prediction promotes the model to learn the alignment between attributes, queries, titles, and images. We use 15% masking probability, and replace masked properties or tokens of values with one of the following: special [MASK] tokens, random tokens, or the original tokens with probability 80%, 10%, and 10%, respectively. It is guaranteed that the property and value of the same attribute will not be masked at the same time. We optimize the following negative logarithm likelihood by predicting these masked values a_m based on the remaining attributes $a_{\backslash m}$, query Q, title T, and the image patches V.

$$\mathcal{L}_{\mathrm{A}} = -\mathbb{E}_{\{Q,T,A,V\}\in\mathcal{D}} \log P_\theta\left(a_m \mid a_{\backslash m}, Q, T, V\right) \tag{4}$$

2.2.3 Relevance Prediction (RP) Pre-Training. While human-annotated relevance data is typically scarce and expensive, it is beneficial for the model to acquire relevance knowledge during the pre-training stage. Therefore, we leverage click signals as weakly-supervised relevance labels for the Relevance Prediction (RP) task. The model output $\boldsymbol{h}_{m_{[CLS]}}$ is fed into the linear and sigmoid layers to predict the relevance score. The objective is to predict the relevance label y based on the query Q, title T, attribute A, and image patches V, by minimizing the cross-entropy loss with weakly-supervised label:

$$\mathcal{L}_{\mathrm{R}} = -\mathbb{E}_{\{Q,T,A,V,y\}\in\mathcal{D}} \log P_\theta\left(y \mid Q, T, A, V\right) \tag{5}$$

The final pre-training objective is the sum of the losses of the four pre-training tasks.

$$\mathcal{L} = \mathcal{L}_{\mathrm{T}} + \mathcal{L}_{\mathrm{V}} + \mathcal{L}_{\mathrm{A}} + \mathcal{L}_{\mathrm{R}} \tag{6}$$

2.3 Model Fine-Tuning and Online Deployment

So far, the proposed method does not involve manually labeled data at all. In practice, there are usually more or less labeled data available, and it would be desirable to further improve the model using this data. Fine-tuning is an effective approach to improving model performance with high-quality training data. Therefore, we fine-tune the MSRA model on human-annotated relevance data using the Relevance Prediction (RP) task to improve performance.

E-commerce engine processes hundreds of millions of search requests every day. Since the MSRA model proposed in this paper performs interactive encoding between queries and items, it cannot be directly deployed online due to computing and resource constraints. To address this issue, we distill the interaction-based model into a representation-based two-tower model for online serving, following the approach of ReprBERT [20].

3 EXPERIMENTS

3.1 Experimental Details

3.1.1 Dataset and Implementation Details. We collect search logs from Taobao for a year to construct a dataset of 10 billion samples, which is used in the pre-training stage. Additionally, we construct an annotated dataset for fine-tuning and evaluation. Specifically, we randomly select two million query-item pairs from the search logs and employ experienced human annotators to label them as either Good (relevant) or Bad (irrelevant). The dataset consists of approximately 1.60 million pairs for training, 200k pairs for validation, and 200k pairs for testing.

We set our model parameters based on preliminary experiments on the validation set. The word vocabulary size is $66,323$, while we add $45,195$ high-frequency brand and category words to the BERT vocabulary as coarse-grained tokens. We set the number of Transformer encoder layers to 12 while each layer has 768 hidden units and 12 self-attention heads. We use Adam optimizer [7] with weight decay $\epsilon = 10^{-5}$. We pre-train the model for 5 epochs with a batch size of 1,000 on 50 NVIDIA P100 GPUs and warm up the learning rate to 1e-4 in the first 4,000 iterations. We set the batch size to 300 and the learning rate to 2e-5 for fine-tuning and distillation.

3.1.2 Metrics and Baselines. We use the ROC-AUC metric to automatically evaluate our results. To assess the impact of attributes and images on semantic relevance, we train a base model that only takes query and title as inputs. Additionally, we introduce a baseline method to test the effectiveness of our proposed encoding method in capturing attribute information. We concatenate the raw text of properties and values with the query and title, and refer to this as Base+attr(CPVT) (Concatenating both Property and Values as Text). We also compare our results to several state-of-the-art methods. All baselines are fine-tuned with the training set to achieve the best performance on the validation set. We compare the proposed model with existing state-of-the-art methods, where ReprBERT [20] is a strong baseline with the pre-training and fine-tuning technology.

3.2 Experimental Results

We present the results on the annotated test set in Table 1. Among the interaction-based methods, both *base+attr* and *base+img* show improvement over the *base model*, indicating the effectiveness of attributes and images. Our proposed encoding method *base+attr* performs better at encoding attribute information than the baseline *base+attr(CPVT)*, and integrating attributes and images together results in further gains, with *base+attr+img* achieving the best results. Surprisingly, *Base* achieves better results than *ReprBERT*, emphasizing the importance of pre-training with e-commerce data.

Table 1: ROC-AUC evaluation results on the test set.

Model	Interaction-based	Representation-based
MASM[19]	-	0.793
SBERT[13]	-	0.765
Poly-encoders[5]	-	0.808
ReprBERT[20]	-	0.894
Base	0.934	0.915
Base+attr	0.942	0.921
Base+attr(CPVT)	0.935	0.916
Base+img	0.938	0.918
Base+attr+img	**0.945**	**0.924**

Table 2: Results of interaction-based models on the validation set.

Model	ROC-AUC
MSRA	0.946
w/o MLM	0.914
w/o MPM	0.943
w/o MAM	0.939
w/o RP	0.942

We also conduct online A/B testing to validate the effectiveness of MSRA. In the baseline experiment, we use the ReprBERT [20] model. While in the test experiment, we replace ReprBERT with our proposed representation-based MSRA. Both experiments accounted for approximately 2% of the traffic and lasted for seven days. Our method outperforms the previous semantic relevance system with an improvement of 0.6% points on overall relevance. Additionally, the proposed model achieves an average improvement of 0.91% in Gross Merchandise Volume (GMV) and 0.44% in the number of transactions over the seven days. The online A/B testing confirms that our proposed model is superior to previous state-of-the-art models and can achieve significant online profits considering the extremely large traffic of our platform every day.

3.3 Ablation Study

We conduct an ablation study on the validation set to investigate the influence of different pre-training objectives in our proposed model. Our pre-training objectives include MLM, MPM, MAM, and RP, and we remove them individually to explore their impact on the results as shown in Table 2.

4 CONCLUSION AND FUTURE WORK

In this paper, we contend that conventional relevance approaches may not perform optimally when the information included in product titles is inadequate to comprehensively match search queries. As a solution, we propose a novel multi-aspect semantic relevance approach that considers the match between queries, item titles, attributes, and images simultaneously. We evaluate the proposed method on both offline data and online A/B testing, and our experimental results demonstrate that it significantly improves relevance and results in considerable improvements in GMV.

In the future, we will incorporate more information, such as item reviews, detail pages, and knowledge graphs, into the relevance model to further enhance its performance.

REFERENCES

[1] David Carmel, Elad Haramaty, Arnon Lazerson, Liane Lewin-Eytan, and Yoelle Maarek. 2020. Why Do People Buy Seemingly Irrelevant Items in Voice Product Search?: On the Relation between Product Relevance and Customer Satisfaction in eCommerce. In *WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020*, James Caverlee, Xia (Ben) Hu, Mounia Lalmas, and Wei Wang (Eds.). ACM, 79–87. https://doi.org/10.1145/3336191.3371780

[2] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-TExt Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science, Vol. 12375)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 104–120. https://doi.org/10.1007/978-3-030-58577-8_7

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, 4171–4186. https://doi.org/10.18653/v1/n19-1423

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 770–778. https://doi.org/10.1109/CVPR.2016.90

[5] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SkxgnnNFvH

[6] Yunjiang Jiang, Yue Shang, Rui Li, Wen-Yun Yang, Guoyu Tang, Chaoyi Ma, Yun Xiao, and Eric Zhao. 2021. A unified Neural Network Approach to E-CommerceRelevance Learning. *CoRR* abs/2104.12302 (2021). arXiv:2104.12302 https://arxiv.org/abs/2104.12302

[7] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[8] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*. AAAI Press, 11336–11344. https://ojs.aaai.org/index.php/AAAI/article/view/6795

[9] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2592–2607. https://doi.org/10.18653/v1/2021.acl-long.202

[10] Yiding Liu, Weixue Lu, Suqi Cheng, Daiting Shi, Shuaiqiang Wang, Zhicong Cheng, and Dawei Yin. 2021. Pre-trained Language Model for Web-scale Retrieval in Baidu Search. In *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM, 3365–3375. https://doi.org/10.1145/3447548.3467149

[11] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[12] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *CoRR* abs/2001.07966 (2020). arXiv:2001.07966 https://arxiv.org/abs/2001.07966

[13] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[14] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=SygXPaEYvH

[15] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *CoRR* abs/1904.09223 (2019). arXiv:1904.09223 http://arxiv.org/abs/1904.09223

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[17] Rong Xiao, Jianhui Ji, Baoliang Cui, Haihong Tang, Wenwu Ou, Yanghua Xiao, Jiwei Tan, and Xuan Ju. 2019. Weakly Supervised Co-Training of Query Rewriting and Semantic Matching for e-Commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019*, J. Shane Culpepper, Alistair Moffat, Paul N. Bennett, and Kristina Lerman (Eds.). ACM, 402–410. https://doi.org/10.1145/3289600.3291039

[18] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 5754–5764. https://proceedings.neurips.cc/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html

[19] Shaowei Yao, Jiwei Tan, Xi Chen, Keping Yang, Rong Xiao, Hongbo Deng, and Xiaojun Wan. 2021. Learning a Product Relevance Model from Click-Through Data in E-Commerce. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 2890–2899. https://doi.org/10.1145/3442381.3450129

[20] Shaowei Yao, Jiwei Tan, Xi Chen, Juhao Zhang, Xiaoyi Zeng, and Keping Yang. 2022. ReprBERT: Distilling BERT to an Efficient Representation-Based Relevance Model for E-Commerce. In *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, Aidong Zhang and Huzefa Rangwala (Eds.). ACM, 4363–4371. https://doi.org/10.1145/3534678.3539090