Align before Search: Aligning Ads Image to Text for Accurate Cross-Modal Sponsored Search

Yuanmin Tang^{1,2}, Jing Yu^{1,2*}, Keke Gai³, Yujing Wang⁴, Yue Hu¹, Gang Xiong¹, Qi Wu⁵

¹Institute of Information Engineering, Chinese Academy of Sciences
²School of Cyber Security, University of Chinese Academy of Sciences
³Beijing Institute of Technology

⁴Microsoft Research Asia

⁵University of Adelaide

{tangyuanmin, yujing02, huyue, xionggang}@iie.ac.cn, gaikeke@bit.edu.cn, yujwang@microsoft.com, qi.wu01@adelaide.edu.au

Abstract

Cross-Modal sponsored search displays multi-modal advertisements (ads) when consumers look for desired products by natural language queries in search engines. Since multimodal ads bring complementary details for query-ads matching, the ability to align ads-specific information in both images and texts is crucial for accurate and flexible sponsored search. Conventional research mainly studies from the view of modeling the implicit correlations between images and texts for query-ads matching, ignoring the alignment of detailed product information and resulting in suboptimal search performance. In this work, we propose a simple alignment network for explicitly mapping fine-grained visual parts in ads images to the corresponding text, which leverages the co-occurrence structure consistency between vision and language spaces without requiring expensive labeled training data. Moreover, we propose a novel model for cross-modal sponsored search that effectively conducts the cross-modal alignment and query-ads matching in two separate processes. In this way, the model matches the multi-modal input in the same language space, resulting in a superior performance with merely half of the training data. Our model outperforms the state-of-the-art models by 2.57% on a large commercial dataset. Besides sponsored search, our alignment method is applicable for general cross-modal search. We study a typical cross-modal retrieval task on the MSCOCO dataset, which achieves consistent performance improvement and proves the generalization ability of our method. Our code is available at https://github.com/Pter61/AlignCMSS/.

1. Introduction

Sponsored search (Jansen and Mullen 2008; Ling et al. 2017; Zhu et al. 2022) is a widely used business model on search engine platforms, where sponsored ads are presented to users with other search results. Displaying the relevant ads for a query will significantly increase the exposure of proper products while effectively satisfying the buyers' demand. Thus, the sponsored search system needs to model the relevance between queries and ads accurately. Besides ads with pure text, multi-modal ads with images have become a new trend and extend the current single-modal sponsored

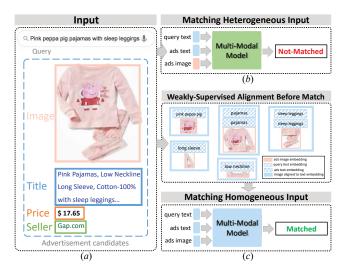


Figure 1: Illustration of our motivation. (b) Existing model with heterogeneous input (a). (c) Our model with homogeneous input is aligned by our weakly-supervised method.

search to Cross-Modal Sponsored Search (CMSS). Images in ads bring complementary product details to texts. Considering the query in Figure 1(a), the agent not only needs to semantically match the ads text of "pink pajamas" and "sleep leggings", but also has to align the query to the visual content of "pink peppa pig" in the ads image. Therefore, accurately capturing the query-relevant content from both visual and textual ads information and jointly matching it with the query is essential to achieve flexible sponsored search.

Recent works (Zhu et al. 2022) regard the cross-modal sponsored search as a traditional image-text matching task and leverage image-text correlation learning approaches to model the relevance between the text query and multi-modal ads. The typical solution, i.e., AdsCVLR (Zhu et al. 2022), is based on the vision and language pre-training (VLP) framework, which treats the query and ads text together as text input and regards the corresponding ads image as visual input. All the inputs are fed into a single-stream VLP model to measure their global similarity as illustrated in Figure 1(b).

^{*}Corresponding author

However, this kind of solution is hard to satisfy fine-grained sponsored search because of the great difference between query-ads matching and image-text matching. In sponsored search, the multi-modal ads contain more detailed product information, e.g., color, brand, style, price, and seller. The queries are short (Zhu et al. 2022) and generally about partial features of the products (e.g., the query in Figure 1(a) focuses on attribute "pink peppa pig"). Thus, aligning fine-grained visual regions in ads images to the corresponding text becomes a core issue for high-quality CMSS. Recent works, which implicitly correlate visual regions and words by attention mechanism in VLP, are challenging to accurately achieve fine-grained alignment without expensive region-word labels for training (empirical evidence is provided in Figure 5 and supplementary material). Moreover, existing works jointly learn cross-modal alignment and query-ads matching by unified transformers, which is dataexhaustive to achieve optimal performance.

To tackle these issues, we propose a new cross-modal sponsored search model that successively conducts crossmodal alignment and query-ads matching by two independent modules. As shown in Figure 1(c), in the cross-modal alignment module, the visual part embeddings of ads images are aligned to the corresponding semantic word embeddings by a weakly-supervised approach without extra labels. We name this module VALSE for short. Specifically, VALSE learns a simple and effective linear mapping for vision-tolanguage alignment. As illustrated in Figure 2, the alignment approach is based on our observation that the topological structure constructed by the co-occurrence relationships between product partial characteristics is consistent in both vision and language spaces (evidence is provided in the supplementary). Based on this hypothesis, we propose a three-stage alignment approach to progressively map the visual embeddings to the language space by unsupervised adversarial training and weakly-supervised refinement. Experimental results demonstrate the superior alignment quality and generalization ability of our alignment approach.

In this way, VALSE unifies the multi-modal ads input in the same language space. Then we propose a novel model AlignCMSS for cross-modal sponsored search by combining VALSE with a typical VLP model. AlignCMSS receives the aligned homogeneous embeddings of query, ads text, and ads image from VALSE and matches the query and aligned multi-modal ads in the same language space via the VLP model. By successively conducting cross-modal alignment and query-ads matching, AlignCMSS significantly outperforms existing CMSS models. Moreover, this align-beforematch strategy is training-efficient, which merely uses 50% training data that results in superior performance compared to the SoTA approaches with the total training data.

The main contributions are summarized as follows. (1) We propose VALSE, a novel method for aligning fine-grained visual parts in ads images to the corresponding text without requiring expensive labeled training data of region-word pairs. VALSE leverages the prior knowledge of co-occurrence structure consistency between vision and language spaces and achieves superior alignment quality for flexible sponsored search. (2) We propose AlignCMSS, a

novel model for CMSS that effectively conducts the cross-modal alignment and query-ads matching in two separate processes. AlignCMSS demonstrates consistent improvement over various CMSS models and significantly outperforms the state-of-the-art (SoTA) model by 2.57% on the existing largest commercial dataset. Moreover, AlignCMSS is training-efficient, which sheds new lights towards the efficient pre-training on large-scale vision-language data. (3) Besides sponsored search, the proposed alignment approach is generally applicable to other cross-modal retrieval tasks and has impacts a broader range of applications. We case-study another cross-modal retrieval task on the MSCOCO dataset and achieve a consistent performance improvement of 1.3% on average precision.

2. Related Work

Relevance Modeling in Sponsored Search. Relevance modeling is a critical component of information retrieval. Previous works (Grbovic and Cheng 2018; Li et al. 2021a; Zhu et al. 2021) have focused on language models to extract information from the text while ignoring the value of visual information. FashionBERT (Gao et al. 2020) considers both image and text information to address imagetext matching in the fashion domain. AdsCVLR (Zhu et al. 2022) first proposes a VLP approach for cross-modal sponsored search to model the relevance of the query, ads text, and ads images, which obtains obvious performance boost compared with existing VLP models. AdsCLVR also trains a two-stream student model using knowledge distillation for online computation and latency constraints. Since the multimodal inputs are unaligned, it is hard for AdsCLVR to accurately correlate the regions in ads images with words in the query, resulting in incorrect prediction when the ads texts are insufficient or inaccurate. We propose a cross-modal alignment approach that enables object-level fine-grained visionlanguage alignment for accurate sponsored search.

Vision and Language Alignment. Existing methods (Zhen et al. 2019; Gao et al. 2020; Yu et al. 2018; Liu et al. 2019; Anderson et al. 2018) typically focused on specific tasks through joint space learning or cross-modal attention. Recent approaches (Su et al. 2020; Li et al. 2020b; Zhang et al. 2021; Radford et al. 2021; Yao et al. 2022) aim to develop unified models for cross-modal alignment using vision-language pre-training (VLP). However, VLP models with implicit attention mechanisms struggle to achieve accurate alignment between visual objects and words without fine-grained annotations, which are essential for sponsored search (empirical evidence in supplementary material). Align-before-fuse VLP methods (Li et al. 2021b, 2022) attempt to overcome this by employing a tunable patchbased visual encoder and an image-text contrastive loss. Nonetheless, the alignment still relies on global imagetext pair supervision, lacking the fine-grained annotations of patch-word pairs necessary for accurate alignment. To address these limitations, we propose a novel alignment method that directly aligns fine-grained image object embeddings with language in a weakly-supervised way. Our alignment stage is independent of relevance modeling, enhancing the cross-modal model (e.g., VLP) better to capture query-ads relevance with the same amount of training data. Furthermore, our method accounts for structure consistency and easily integrates with VLP models, unlike earlier studies (Gupta, Schwing, and Hoiem 2019; Wang et al. 2020) reliant on concept co-occurrence relationships.

Unsupervised Word Translation. The study of unsupervised word translation inspires our weakly-supervised solution for vision-language alignment. Existing works (Lample et al. 2018; Artetxe, Labaka, and Agirre 2019; Pourdamghani et al. 2019) use adversarial training to learn a linear mapping between two languages without parallel corpora. In (Lample et al. 2018), a bilingual dictionary is built between two languages without parallel corpora. To extend this schema to cross-modal translation, (Chung et al. 2018) proposes a speech-word-to-languageword alignment method based on adversarial training for speech-to-text translation. However, these previous works only align sequence-structured modalities, which presents difficulties in the vision-language scenario due to the semantic gap. To overcome these limitations, we propose a novel three-stage alignment approach that enables fine-grained alignment between vision and language modalities, effectively addressing the semantic gap between them.

3. VALSE

The data in the cross-modal sponsored search task is defined as a set of triples: $\mathcal{L} = \langle \mathbf{q}_i, \mathbf{a}_i, \mathbf{y}_i \rangle$, where \mathbf{q}_i denotes the user query, \mathbf{a}_i represents a multi-modal ads containing the product image and text, and $\mathbf{y}_i \in \{0,1\}$ is the relevance label between \mathbf{q}_i and \mathbf{a}_i , where 0 denotes irrelevant while 1 denotes relevant. This task aims to learn a binary classifier $f: f(\mathbf{q}_i, \mathbf{a}_i) \in \{0,1\}$ to accurately predict the relevance between user queries and multi-modal ads.

VALSE aims to learn a mapping to align the region embeddings in ads images to the word embeddings with corresponding semantics in ads texts and queries. In this way, multi-modal input embeddings are unified in language space for direct relevance learning in the Vision-Language Pretraning (VLP) model. The alignment strategy derives from our understanding of the origin of vision and language: both modalities originate from a shared reality. Therefore, relevant visual components appear in the same scenario while humans talk continuously about relevant things. As shown in Figure 2, "Nike", "thick sole", and "high uppers" have similar co-occurrence relationships for the product "Nike basketball shoes" in both ads image and ads text. Thus we consider that the topological structure constructed by co-occurrence relationships between partial product characteristics is consistent in both vision and language space. Our empirical study in the supplementary and a paper published in Nature Machine Intelligence (Roads and Love 2020) has a similar conclusion. Inspired by this observation, we propose to align ads image regions and corresponding text words by aligning their co-occurrence structure constructed by their embeddings instead of learning the mapping directly from the ground-truth text-image pairs. In this way, we don't require labor-extensive text-image annotations. In this section, we first introduce the vision and language embedding extraction approach to keep the co-occurrence structure in a single

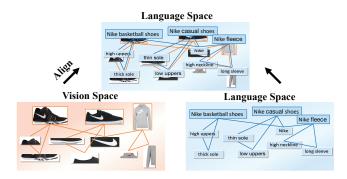


Figure 2: An illustration of structure consistency in vision and language space motivating cross-modal alignment in VALSE.

modality space. Then we propose a three-stage alignment approach based on the structure consistency across different modalities.

3.1. Vision and Language Space Construction

To keep the co-occurrence structure in each modality space, we propose representing the co-occurrence information in the object (resp., word) embeddings. Specifically, in the learned vision (resp., language) embedding space, objects (resp., words) with similar co-occurrence relationships are close to each other. Such structure-preserving constraints have been extensively studied in metric learning (Wang, Li, and Lazebnik 2016; Liu et al. 2021). To make our alignment approach more generic and applicable to existing models, we utilize the widely-used embedding approaches in existing models with implicit co-occurrence learning, e.g., object detector X152-C4 for region embeddings, and transformer for word embeddings. Such implicit co-occurrence embeddings have obtained substantial benefits for crossmodal alignment. Since our work mainly focuses on proving the effectiveness of the align-before-match framework and the alignment strategy, other regularization terms for cooccurrence structure preserving can be further studied.

Vision Embedding Space Construction. Since we select VinVL as the baseline, we use the same object detector X152-C4 (Zhang et al. 2021) in VinVL to detect ads image regions and extract their embeddings to construct the vision space. In the learning process of the object detector, objects in an image are detected and embedded jointly, which implicitly introduces object co-occurrence constraints in each detected object embedding and preserves weak co-occurrence structure. We denote the set of detected object as $\mathcal{O} = \{o_i\}_{i=1}^K (K=50)$ and represent each object o_i by a region embedding $v_i \in \mathbb{R}^{d_1}$ ($d_1 = 2054$ with 2048-dimensional visual embedding and 6-dimensional position embedding). The embeddings extracted from all the ads images in the commercial advertising dataset (Zhu et al. 2022) form the vision space.

Language Embedding Space Construction. Transformers have been widely verified the ability to model contextual information for word embeddings (Devlin et al. 2019; Floridi and Chiriatti 2020). Thus the information of co-occurring

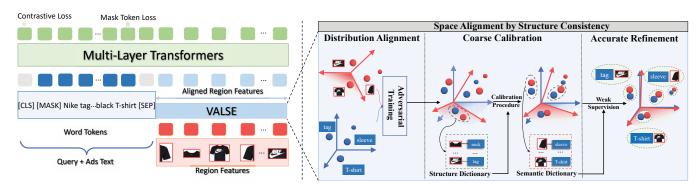


Figure 3: An overview of AlignCMSS with two parts: (left) the architecture of AlignCMSS and (right) three alignment stages in VALSE. We first train VALSE offline, then replace the original linear mapping in VinVL with VALSE and fine-tune it online.

words is implicitly correlated in the output embeddings of Transformers by the attention mechanism. Therefore, we apply the transformer-based language encoder to extract the word embeddings to construct the language space. Since our backbone VinVL is based on transformer structure, we utilize the language part in VinVL multi-modal encoder to obtain the output word embeddings for alignment. Specifically, we tokenize the query q_i and ads text in a_i and obtain a sequence of D tokens with BERT embeddings (Devlin et al. 2019). All the word embeddings with padding image region embeddings are fed into VinVL to extract word embeddings $\{t_j\}_{j=1}^D\in\mathbb{R}^{d_2}$ $(d_2$ =768) from the output layer. We only keep noun embeddings to form the language embedding space for alignment since the image regions are mostly objects, and the noun embeddings contain contextual information about the nouns.

3.2. Space Alignment by Structure Consistency

Given a set of m region embeddings $\mathcal{V} = \{v_i\}_{i=1}^m \subseteq \mathbb{R}^{d_1 \times m}$ and a set of n noun embeddings $\mathcal{T} = \{t_j\}_{j=1}^n \subseteq \mathbb{R}^{d_2 \times n}$, we aim to learn a linear mapping $\boldsymbol{W_{align}} \in \mathbb{R}^{d_2 \times d_1}$ to align the two spaces as:

$$W_{align} = \underset{\boldsymbol{W} \in \mathbb{R}^{d_2 \times d_1}}{\arg \min} (||\boldsymbol{W} \boldsymbol{\mathcal{V}} - \boldsymbol{\mathcal{T}}||^2)$$
(1)

We use linear mapping instead of complex non-linear mapping for alignment since we align the co-occurrence structures of regions and nouns instead of the one-to-one features of region-noun pairs. The co-occurrence structures of vision and language spaces have natural consistency, which has been proven effective for linear mapping to align in our study and previous bilingual word alignment study (Lample et al. 2018). In this work, we do not have a golden dictionary that specifies which v_i corresponds to t_i . We propose a weakly-supervised cross-modal alignment approach with three stages: Distribution Alignment first coarsely maps the distribution of region embeddings with the distribution of noun embeddings via adversarial training; Coarse Calibration then selects the best mutually aligned region-noun embeddings as a structural dictionary to calibrate W_{alian} ; Accurate Refinement automatically constructs region-noun semantic labels for weakly-supervised refinement of W_{align} .

Distribution Alignment by Adversarial Training. This stage coarsely aligns the distribution of vision and language embeddings by making the mapped region features $W_{align}\mathcal{V}$ and noun embeddings \mathcal{T} indistinguishable. Since adversarial training is a strong unsupervised solution to achieve this goal, we apply an adversarial approach for learning W_{align} without cross-modal supervision. We define the alignment mapping W_{align} as the generator and a binary classifier as the discriminator, which is a multilayer perceptron with two hidden layers parameterized by θ_D . The generator aligns region embeddings to the language space by $W_{align}v_i$, while the discriminator distinguishes between the transformed region embeddings and noun embeddings t_j by minimizing the following objective:

$$\mathcal{L}_{D}(\theta_{D}|\boldsymbol{W_{align}}) = -\frac{1}{m} \sum_{i=1}^{m} \log P_{\theta_{D}}(\text{vis} = 1|\boldsymbol{W_{align}}\boldsymbol{v_{i}})$$
$$-\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_{D}}(\text{vis} = 0|\boldsymbol{t_{j}})$$
 (2)

where $P_{\theta_D}(\text{vis} = 1 | \boldsymbol{W_{align}}\boldsymbol{v_i})$ denotes the probability that $\boldsymbol{v_i}$ is from the vision space while $P_{\theta_D}(\text{vis} = 0 | \boldsymbol{t_j})$ is the probability that $\boldsymbol{t_j}$ is from the language space. On the contrary, the generator fools the discriminator from making correct predictions by minimizing the following objective:

$$\mathcal{L}_{W}(\boldsymbol{W_{align}}|\theta_{D}) = -\frac{1}{m} \sum_{i=1}^{m} \log P_{\theta_{D}}(\text{vis} = 0|\boldsymbol{W_{align}}\boldsymbol{v_{i}})$$
$$-\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_{D}}(\text{vis} = 1|\boldsymbol{t_{j}})$$
(3)

Coarse Calibration by Pseudo Structure Dictionary. Since the adversarial training aligns regions and nouns regardless of their frequency, many regions with a small amount of data result in worse alignment. To calibrate the mapping, we build a structure dictionary from the roughly aligned region-noun embeddings after adversarial training. To ensure a high-quality dictionary, we consider the top frequent nouns to align, which are expected to have better alignment quality with more training samples. We regard the mutual nearest neighbors (MNN) (Lample et al. 2018) between frequent noun embeddings and aligned region em-

beddings as pseudo region-noun pairs to construct the dictionary, *i.e.*, pseudo structure dictionary. We use MNN instead of *K*-nearest neighbors to avoid the hubness problem (Dinu, Lazaridou, and Baroni 2014) (embeddings tending to be nearest neighbors of many embeddings).

To this end, we select the top N (N=30) frequent nouns from \mathcal{T} , denoted as $\hat{\mathcal{T}} = \{\hat{t}_i\}$, and then compute MNN of \hat{t}_i from the aligned region embeddings $W_{align}\mathcal{V}$, denoted as $\hat{\mathcal{V}} = \{\hat{\boldsymbol{v}}_i\}$, to form the dictionary. We utilize the Cross-Domain Similarity Local Scaling (CSLS) metric (Lample et al. 2018) for MNN computation. We first select L region embeddings $\{\hat{v}_l\} \subset \hat{\mathcal{V}}$ with top M cosine similarities for each $\hat{m{t}}_j$ in a training batch. For each $(\hat{m{v}}_l,\hat{m{t}}_j)$ pair, we compute MNN as $CSLS(\hat{\boldsymbol{v}}_l, \hat{\boldsymbol{t}}_j) = 2\cos(\hat{\boldsymbol{v}}_l, \hat{\boldsymbol{t}}_j) - r_{\hat{\mathcal{T}}}(\hat{\boldsymbol{v}}_l) - r_{\hat{\mathcal{V}}}(\hat{\boldsymbol{t}}_j)$, where $r_{\hat{\mathcal{T}}}(\hat{\boldsymbol{v}}_l) = \frac{1}{K} \sum_{\hat{\boldsymbol{t}}_y \in \mathcal{N}_{\hat{\mathcal{T}}}(\hat{\boldsymbol{v}}_l)} \cos\left(\hat{\boldsymbol{v}}_l, \hat{\boldsymbol{t}}_y\right)$ computes the mean similarity between $\hat{\boldsymbol{v}}_l$ and its K nearest neighbors in $\hat{\mathcal{T}}$, *i.e.*, $\mathcal{N}_{\hat{\mathcal{T}}}(\hat{\boldsymbol{v}}_l)$. $r_{\hat{\mathcal{V}}}(\hat{\boldsymbol{t}}_j)$ shares similar operations but differs in similarity computation direction. We choose the nearest \hat{v}_l for each \hat{t}_j and form the dictionary with all the $\{\hat{\boldsymbol{v}}_l, \hat{\boldsymbol{t}}_j\}$ pairs. In practice, (Xing et al. 2015) proved that the results are optimized by enforcing an orthogonality constraint on W_{align} . Following (Xing et al. 2015), we regard Eq. 1 as the Procrustes problem and provides a closed solution by the singular value decomposition (SVD) on the structure dictionary as W_{align} = UV^T , with $U\Sigma V^T = SVD(\hat{T}\hat{V}^T)$.

Accurate Refinement by Pseudo Semantic Dictionary. Though the above two stages coarsely align vision and language spaces, they are difficult to align specific product characteristics accurately, e.g., mapping the ads image region to the distinctive description 'black Nike T-shirt', since there lack of semantic annotations for fine-grained alignment. Therefore, we propose to automatically construct a semantic dictionary as supervision to refine further W_{align} . Specifically, we first detect all the object regions $\{o_i\}$ with corresponding object labels $\{l_i\}$ in the ads images. Since the query generally describes product features by nouns, we select nouns $\{w_i\}$ in each query that match object labels in the relevant ads image satisfying $w_j \in \{l_i\}$ to form pseudo region-noun pairs $\{o_i, w_i\}$. To represent w_i with comprehensive product features, we locate all w_i in both query and ads text and average their embeddings to represent w_i . The region embeddings are the same as in Sec. 3.1. In this way, we build a dictionary with about 50K region-noun pairs and sample 20% of them according to the distribution of nouns to form the semantic dictionary for refinement according to Eq. 1 as Coarse Calibration.

3.3. AlignCMSS: Cross-Modal Sponsored Search Model with VALSE

Since VALSE is learned independently of the sponsored search model, we introduce the training strategy of incorporating W_{align} with a recently proposed VLP model, e.g., VinVL (Zhang et al. 2021). We name our model as AlignCMSS. Following the design in VinVL, we also feed the object tags detected in the ads image together with the query, ads text, and ads image to the multi-modal encoder.

As shown in Figure 3 (left), we directly replace the original linear mapping of the input layer in VinVL by VALSE to align the region embeddings to the language ones and fine-tune it with VinVL together. We initialize VinVL with its pre-trained weights and further pre-train AlignCMSS on the ads search dataset with two tasks: Masked Token Model (MTM) and Query-Ads Contrastive Learning.

Masked Token Model. Similar to Mask Language Model (Devlin et al. 2019), we randomly mask each input token h_i with the probability of 15% from m input tokens and replace the masked token with a special token [MASK]. The goal of MTM is to predict the masked tokens based on other tokens, denoted as \hat{h}_i , by minimizing the negative log-likelihood:

$$\mathcal{L}_{MTM} = -\mathbb{E}_{\boldsymbol{h}_{j} \sim \mathcal{D}} \log p(\boldsymbol{h}_{i}|\hat{\boldsymbol{h}}_{i})$$
 (4)

Query-Ads Contrastive Learning. Following VinVL (Zhang et al. 2021), we construct negative samples for each query-ads pair by replacing the query with a different query from the training set with a probability of 25%. The training data with both positive and negative samples is denoted as \mathcal{D} . We feed the output embedding of [CLS] to a fully-connected layer as a classifier $f(\cdot)$ to predict whether the query matches the ads (c=0) or not (c=1). The contrastive loss is defined as:

$$\mathcal{L}_{CL} = -\mathbb{E}_{(\boldsymbol{q}_i, \boldsymbol{a}_i; c) \sim \mathcal{D}} \log p(c|f(\boldsymbol{q}_i, \boldsymbol{a}_i))$$
 (5)

The final pre-training loss for AlignCMSS is defined as: $\mathcal{L} = \mathcal{L}_{MTM} + \mathcal{L}_{CL}.$

Fine-tuning. We formulate the cross-modal sponsored search task as a binary classification problem as defined in (Qi et al. 2020). Given a query-ads pair (q_i, a_i) with ground-truth label $y_i \in \{0, 1\}$, the output <code>[CLS]</code> token $t_{(q_i, a_i)}$ of AlignCMSS is fed to a binary classifier to predict the relevance, i.e. $\hat{y}_i = p(t_{(q_i, a_i)})$. We construct negative samples by replacing queries q_i from positive samples, thereby making the model aware of the importance of user queries. The cross-entropy loss is defined as:

$$\mathcal{L}_{BCE} = -\mathbb{E}_{(\boldsymbol{q_i}, \boldsymbol{a_i}) \sim \mathcal{L}}[\boldsymbol{y_i} \log \hat{\boldsymbol{y}_i} + (1 - \boldsymbol{y_i}) \log (1 - \hat{\boldsymbol{y}_i})]$$
(6)

In the testing stage, the probability of \hat{y}_i is used to predict whether a given query-ads is relevant.

4. Experiments

Dataset and Evaluation Metric. We test AlignCMSS on a large commercial advertising dataset proposed in (Zhu et al. 2022), containing 480K query-ad pairs where 400K for training and 80K for testing. Queries are labeled with relevant and irrelevant ads. We evaluate the model by the Area under Reciever Operating Characteristic Curve (AUC) as in (Cortes and Mohri 2003).

Implementation Details. The vision space contains 1.5 million region embeddings, and the language space has 1.6 million word embeddings extracted from the dataset. VALSE is pre-trained by an SGD optimizer with 100k iterations, where the mini-batch size is 1024, and the learning rate is 0.1. The input length of AlignCMSS is 152. AlignCMSS is pre-trained by Adam optimizer with 180k iterations, where the mini-batch size is 64, and the dropout ratio is 0.1. We set

Ads Input	Model	AUC
Text	BERT-base (Devlin et al. 2019)	82.15
	CLIP (Radford et al. 2021)	
Image	ViLT (Kim, Son, and Kim 2021)	
	Unicoder-VL (Li et al. 2020a)	83.16
	CLIP@(Qt-OtOi) (Zhu et al. 2022)	81.98
	CLIP-MH@(Qt-OtOi) (Zhu et al. 2022)	82.18
	ALBEF (Li et al. 2021b)	82.74
	BLIP (Li et al. 2022)	83.51
	VL-BERT (Su et al. 2020)	86.27
	OSCAR (Li et al. 2020b)	87.45
Multi-Modal	Unicoder-VL@(Qt-OtOi) (Zhu et al. 2022)	87.90
Muiti-Modai	VinVL (Zhang et al. 2021)	88.56
	AdsCVLR (Zhu et al. 2022)	89.16
	VL-BERT+VALSE	90.13
	OSCAR+VALSE	90.62
	AlignCMSS (50%)	90.46
	AlignCMSS (70%)	91.02
	AlignCMSS (100%)	91.73

Table 1: Comparison with state-of-the-art models.

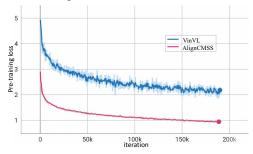


Figure 4: Comparison of the pre-training loss curves.

the learning rate to 1×10^{-4} . The model is fine-turned by 40 epochs, and the learning rate is 2×10^{-5} . VALSE is trained for 16 hours, while AlignCMSS is pre-trained for 216 hours and fine-tuned for 24 hours with 3 NVIDIA V100 (32G).

4.1. Comparison with State-of-the-Art Methods

Table 1 shows the results of three kinds of state-of-the-art (SoTA) models: text-only model (Devlin et al. 2019), image-only models (Radford et al. 2021; Kim, Son, and Kim 2021; Li et al. 2020a), and multi-modal models (Su et al. 2020; Li et al. 2020b; Zhu et al. 2022; Zhang et al. 2021; Li et al. 2021b, 2022). Besides VinVL, we also combined VALSE with another two widely compared VLP models, *i.e* VL-BERT and OSCAR, in the last block to prove the generalization ability of our align-and-match strategy. We also pre-trained and fine-tuned AlignCMSS with 50%, 70%, and 100% training data to prove its effectiveness.

The results show that AlignCMSS consistently outperforms all existing approaches, achieving a new SoTA performance that significantly outperforms the SoTA model AdsCVLR by 2.57%. VALSE also significantly improves VL-BERT and OSCAR by 3.86% and 3.17%, respectively.

	Method	AUC		
1.	AlignCMSS (full model)	91.73		
Ablation of Three Stage Alignment				
2.	w/o Accurate Refinement	90.34		
3.	w/o Coarse Calibration	91.16		
4.	w/o Distribution Alignment & Coarse Calibration	90.65		
5.	w/o Coarse Calibration & Accurate Refinement	89.97		
6.	w/o VALSE Alignment	89.51		
Abla	Ablation of Training Strategies			
7.	w/o \mathcal{L}_{MTM}	91.37		
8.	w/o \mathcal{L}_{CL}	91.46		
9.	w/o \mathcal{L}_{MTM} & \mathcal{L}_{CL}	91.13		
10.	w/o Fine-tuning	91.04		
Abla	Ablation of Pre-trained Knowledge in VinVL			
11.	w/o VinVL pre-training	90.14		
Abla	Ablation of Object Tags in VinVL			
12.	w/o Object Tags	91.55		

Table 2: Ablation of key components in AlignCMSS.

Moreover, AlignCMSS outperforms AdsCVLR by 1.3% with only 50% of the training data, proving that unifying heterogeneous inputs in language space through VALSE significantly enhances the relevance of measuring the VLP model's ability. Besides the advantages of using less training data, AlignCMSS achieves a faster and more stable learning process than VinVL, as shown by the pre-training loss curves in Figure 4, because VALSE effectively bridges the semantic gap between ads images and text and alleviates the burden of cross-modal correlation learning.

4.2. Ablation Study

In Table 2, we evaluate the effectiveness of three alignment stages, training losses, and pre-trained knowledge in AlignCMSS. (1) In models '2-6', we observe that each alignment stage is indispensable for the performance boost. When removing accurate refinement (model '2'), the AUC score decreases by 1.39%, which is more significant than removing coarse calibration. It indicates that weak supervision is essential for bridging the semantic gap across different modalities. Only weakly-supervised alignment (model '4') or only unsupervised adversarial training (model '5') leads to further performance decrease. Without VALSE (model '6'), AUC decreases obviously by 2.22%. It proves that the three stages of VALSE work jointly to achieve the maximum boost. (2) In models '7-10', we conclude that the masked token model and contrastive learning are effective pre-training strategies, while fine-tuning is also necessary to achieve the best performance. Combining with other pre-training tasks, such as image classification in AdsCVLR (Zhu et al. 2022), can be studied in future work. (3) Without the pre-trained weights in VinVL (model '11'), the AUC score drops by 1.59%, indicating that implicit vision-language correlations learnt by pre-training benefit relevance learning. It is worth noticing that the AUC score of model '11' is still 0.63% higher than model '6', proving that the explicit vision and language alignment by VALSE is more beneficial for relevance modeling than implicit alignment by pre-training. (4)

	Model	AUC	
Semantic Dictionary Size			
1.	Semantic dictionary size-100%	91.82	
2.	Semantic dictionary size-20%	91.73	
3.	Semantic dictionary size-2%	91.20	
4.	Semantic dictionary size-0.2%	90.96	
5.	Semantic dictionary size-0%	90.34	
Nou	Noun Selection in the Semantic Dictionary		
6.	Noun-head-20%	91.16	
7.	Noun-tail-20%	90.86	
8.	Noun-random-20%	91.61	
Unsu	Unsupervised Alignment Method		
9.	Wasserstein Procrustes	90.15	
10.	Accurate refinement & Dictionary size-100%	91.18	
Lang	Language Embedding		
11.	BERT noun token Embedding	90.79	
12.	Single noun Embedding	91.22	

Table 3: Analysis of key components in VALSE.

Without the object tags in AlignCMSS (model '12'), the AUC score only slightly decreases by 0.18%, which indicates that VALSE achieves the major contribution for alignment instead of tags.

4.3. Key Component Analysis in VALSE

We conduct extensive experiments on alternative methods of key components in VALSE to prove the advantages of the selected approach in Table 3. (1) In models '1-5', we evaluate the effect of semantic dictionary size on the search performance. We range the dictionary size from 100% to 0% (i.e. without the accurate refinement stage) and conclude that more pseudo semantic supervision achieves better performance. Model '2' (full model) with only 20% pseudo region-noun pairs obtains comparable performance with the fully supervised counterpart (model '1'), which proves the effectiveness of VALSE with relatively low training resources. (2) In models '6-8', we assess the influence of noun classes in the semantic dictionary on performance. We build four dictionaries with the same data size (20%) in different ways: head 30 frequent nouns (model '6'), tail 30 frequent nouns (model '7'), and randomly selected nouns (model '8'). The AUC score of models '6' and '8' are higher, which indicates that the more consistent between the distribution of supervised data and aligned data, the better alignment is achieved. (3) In model '9', we replace VALSE with another typical unsupervised distribution alignment method Wasserstein Procrustes (Grave, Joulin, and Berthet 2019). The performance decreases by 1.58% compared with AlignCMSS. In model '10', we only retain the weakly-supervised stage with 100% training data, which achieves inferior performance compared to the full model with the unsupervised alignment stages. In summary, the unsupervised alignment strategy in VALSE is effective and complementary to the weakly-supervised stage. (4) In models '11-12', both the original BERT noun token embeddings and the single noun

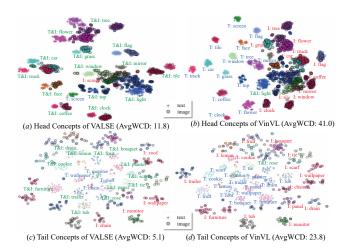


Figure 5: T-SNE visualization of the aligned region and noun embeddings. 'Cross' and 'T' indicate nouns, 'circle' and 'I' mean aligned regions, and 'T&I' means well-aligned regions and nouns. Best viewed in color with 300% zoom.

	1K Test Set		5K Test Set	
Model	Text Retrieval	Image Retrieval	Text Retrieval	Image Retrieval
	R@1/5/10	R@1/5/10	R@1/5/10	R@1/5/10
VinVL	89.8 / 98.8 / 99.7	78.2 / 95.6 / 98.0	74.6 / 92.6 / 96.3	58.1 / 83.2 / 90.1
AlignCMSS	91.2 / 99.4 / 99.8	80.1 / 97.3 / 99.2	77.4 / 94.1 / 97.0	60.4 / 84.3 / 90.7
$\Delta\uparrow$	1.4/ 0.6/ 0.1	1.9 / 1.7 / 1.2	2.8 / 1.5 / 0.7	2.3 / 1.1 / 0.6

Table 4: Results of image-text retrieval on MSCOCO.

embeddings (without averaging) are inferior to the averaged noun embeddings output by VinVL, which contain richer co-occurrence information for structure alignment.

4.5. Alignment Analysis

We conduct experiments to prove the alignment quality between regions and nouns and VALSE's generalization ability. Please refer to the supplementary for a detailed analysis.

- (1) Alignment Quality between Regions and Words. We use t-SNE (Van der Maaten and Hinton 2008) to visualize both head and tail nouns and aligned region embeddings. We compare VALSE with the outputs of VinVL in Figure 5. VALSE aligns region embeddings with related nouns much closer than VinVL over both head and tail concepts. Moreover, we obtain the same conclusion from the quantitative evaluation by Average Within-Cluster Distance (AvgWCD) (Edwards and Cavalli-Sforza 1965). The smaller AvgWCD means the better aligned quality. VALSE consistently outperforms VinVL on this metric.
- (2) Alignment Generalization Ability. VALSE is a generic alignment approach and can be applied to other cross-modal scenarios. We case-study another cross-modal retrieval task on the MSCOCO dataset (Lin et al. 2014) to prove its generalization ability. Table 4 shows the retrieval results on 1K and 5K test sets. Compared to VinVL, AlignCMSS achieves a consistent performance boost on text retrieval and image retrieval tasks over all the metrics.

5. Conclusion

In this paper, we propose a novel alignment method VALSE to align ads regions to ads texts based on their structure consistency. Combined VALSE with VinVL, we propose a new cross-modal sponsored search model AlignCMSS with state-of-the-art performance on a large commercial dataset. Extensive experiments also prove its advantages of high alignment quality, efficient pre-training, and generalization ability. The application of VALSE in other cross-modal scenarios will be explored in our future work.

References

- Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6077–6086.
- Artetxe, M.; Labaka, G.; and Agirre, E. 2019. An Effective Approach to Unsupervised Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 194–203.
- Chung, Y.-A.; Weng, W.-H.; Tong, S.; and Glass, J. 2018. Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces. *Advances in Neural Information Processing Systems*, 31: 7354–7364.
- Cortes, C.; and Mohri, M. 2003. AUC Optimization vs. Error Rate Minimization. In Thrun, S.; Saul, L.; and Schölkopf, B., eds., *Advances in Neural Information Processing Systems*.
- Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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*, 4171–4186.
- Dinu, G.; Lazaridou, A.; and Baroni, M. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv e-prints*, arXiv:1412.6568.
- Edwards, A. W.; and Cavalli-Sforza, L. L. 1965. A method for cluster analysis. *Biometrics*, 362–375.
- Floridi, L.; and Chiriatti, M. 2020. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30: 681–694.
- Gao, D.; Jin, L.; Chen, B.; Qiu, M.; Li, P.; Wei, Y.; Hu, Y.; and Wang, H. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2251–2260.
- Grave, E.; Joulin, A.; and Berthet, Q. 2019. Unsupervised alignment of embeddings with wasserstein procrustes. In *Proceedings of Machine Learning Research*, 1880–1890.
- Grbovic, M.; and Cheng, H. 2018. Real-time personalization using embeddings for search ranking at airbnb. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 311–320.
- Gupta, T.; Schwing, A.; and Hoiem, D. 2019. Vico: Word embeddings from visual co-occurrences. In *Proceedings of*

- the IEEE/CVF International Conference on Computer Vision, 7425–7434.
- Jansen, B. J.; and Mullen, T. 2008. Sponsored search: an overview of the concept, history, and technology. *International Journal of Electronic Business*, 6(2): 114–131.
- Kim, W.; Son, B.; and Kim, I. 2021. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, 5583–5594.
- Lample, G.; Conneau, A.; Ranzato, M.; Denoyer, L.; and Jégou, H. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.
- Li, C.; Pang, B.; Liu, Y.; Sun, H.; Liu, Z.; Xie, X.; Yang, T.; Cui, Y.; Zhang, L.; and Zhang, Q. 2021a. Adsgnn: Behavior-graph augmented relevance modeling in sponsored search. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, 223–232.
- Li, G.; Duan, N.; Fang, Y.; Gong, M.; and Jiang, D. 2020a. Unicoder-VL: A Universal Encoder for Vision and Language by Cross-Modal Pre-Training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11336–11344.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, 12888–12900.
- Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021b. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *Advances in Neural Information Processing Systems*, 9694–9705.
- Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *European Conference on Computer Vision*, 740–755.
- Ling, X.; Deng, W.; Gu, C.; Zhou, H.; Li, C.; and Sun, F. 2017. Model ensemble for click prediction in bing search ads. In *Proceedings of the 26th international conference on world wide web companion*, 689–698.
- Liu, H.; Feng, Y.; Zhou, M.; and Qiang, B. 2021. Semantic ranking structure preserving for cross-modal retrieval. *Applied Intelligence*, 51: 1802–1812.
- Liu, X.; Wang, Z.; Shao, J.; Wang, X.; and Li, H. 2019. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1950–1959.
- Pourdamghani, N.; Aldarrab, N.; Ghazvininejad, M.; Knight, K.; and May, J. 2019. Translating Translationese: A Two-Step Approach to Unsupervised Machine Translation.

- In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3057–3062.
- Qi, D.; Su, L.; Song, J.; Cui, E.; Bharti, T.; and Sacheti, A. 2020. ImageBERT: Cross-modal Pre-training with Large-scale Weak-supervised Image-Text Data. *arXiv*:2001.07966.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the International Conference on Machine Learning*, 8748–8763.
- Roads, B. D.; and Love, B. C. 2020. Learning as the unsupervised alignment of conceptual systems. *Nature Machine Intelligence*, 2(1): 76–82.
- Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *Proceedings of the International Conference on Learning Representations*.
- Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11): 2579–2605.
- Wang, H.; Zhang, Y.; Ji, Z.; Pang, Y.; and Ma, L. 2020. Consensus-aware visual-semantic embedding for image-text matching. In *European Conference on Computer Vision*, 18–34.
- Wang, L.; Li, Y.; and Lazebnik, S. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5005–5013.
- Xing, C.; Wang, D.; Liu, C.; and Lin, Y. 2015. Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the North American chapter of the association for computational linguistics: human language technologies*, 1006–1011.
- Yao, L.; Huang, R.; Hou, L.; Lu, G.; Niu, M.; Xu, H.; Liang, X.; Li, Z.; Jiang, X.; and Xu, C. 2022. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *International Conference on Learning Representations*.
- Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1307–1315.
- Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.
- Zhen, L.; Hu, P.; Wang, X.; and Peng, D. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10394–10403.
- Zhu, J.; Cui, Y.; Liu, Y.; Sun, H.; Li, X.; Pelger, M.; Yang, T.; Zhang, L.; Zhang, R.; and Zhao, H. 2021. Textgnn: Improving text encoder via graph neural network in sponsored search. In *Proceedings of the Web Conference 2021*, 2848–2857.

Zhu, Y.; Han, C.; Zhan, Y.; Pang, B.; Li, Z.; Sun, H.; Li, S.; Shi, B.; Duan, N.; Deng, W.; Zhang, R.; Zhang, L.; and Zhang, Q. 2022. AdsCVLR: Commercial Visual-Linguistic Representation Modeling in Sponsored Search. In *Proceedings of the ACM International Conference on Multimedia*, 444–452.