PREMISE: Matching-based Prediction for Accurate Review Recommendation

Abstract

We present PREMISE (PREdict with MatchIng ScorEs), a new architecture for the *matching-based* learning in the multimodal fields for the Multimodal Review Helpfulness Prediction (MRHP) task. Distinct to previous fusion-based methods which obtains multimodal representations via cross-modal attention for downstream tasks, PREMISE computes the multi-scale and multi-field representations, filters duplicated semantics, and then obtained a set of matching scores as feature vectors for the downstream recommendation task. This new architecture significantly boosts the performance for such multimodal tasks whose context matching content are highly correlated to the targets of that task, compared to the state-of-the-art fusion-based methods. Experimental results on two publicly available datasets show that PREMISE achieves promising performance with less computational cost.

1 Introduction

The e-commerce industry has experienced an unprecedented boom in the past decade. Powered by an instant trading system, online shopping platforms successfully endow buyers who seek their favorite goods and sellers who advertise their products with convenience for transaction (Boysen et al., 2019; Vulkan, 2020; Alfonso et al., 2021). However, when wandering through these shops, customers easily fall into the dilemma of deciding whether to buy a product displayed on the screen. At that time, the comments left by past customers are often considered as the most valuable reference. Therefore, how to automatically evaluate review's quality and accurately recommend these reviews becomes a challenge yet an opportunity for online shopping platforms to attract and hold customers. Formally, researchers formulate this problem as the Review Helpfulness Prediction (RHP) task (Tang et al., 2013; Ngo-Ye and Sinha, 2014), which aims to quantify the value of each review to potential customers. By sorting these reviews according to the predicted helpfulness scores in descending order, the platform can post the most valuable reviews at the conspicuous location in the shop page.

In recent years, RHP task has been extended to the multimodal scenario by incorporating the text-attached images as an auxiliary source to help the model make more accurate predictions (Liu et al., 2021), termed Multimodal RHP (MRHP).

Product Name: Gourmia GK250 (1.8 Qt/1.7 L) Cordless Stainless Steel Kettle Supreme - Speed Boil - Auto Shutoff Boil Detect - Concealed Element - 360 Swivel Base - 1500 Watts

Product description: Forget fumbling with all of those microwaves...Heat up to 1.7 liters of piping hot water in a flash! ... and auto shut off to ensure completely safe kettle usage...

Review 1 (Helpfulness Score: 4): I've had my eye on an eighty dollar stainless steel electric kettle for a while, but I didn't care to make that investment just yet. ... Kettle automatically turned off within seconds of reaching a full boil. **When filled halfway (to 1 liter), it takes about 4 and a half minutes to boil. This kettle doesn't take up much counter space either, it's easy to tuck in to the corner by my stove when I need it out of the way.** Overall I'm pretty happy with this, and **thankful for a kettle that turns itself off so I don't have to worry about forgetting it while it's boiling.**

Review 2 (Helpfulness Score: 1): The handle for the kettle which love or I thought I did is falling off. What do I do? Was this product meant to only last for a few months? I need help here! So disappointing. I did not want to report it here, but I see nowhere else to do so.

Table 1: A pair of reviews with high and low helpfulness scores from Amazon-MRHP dataset. We highlight the text that provides customers helpful information. Due to space limitation we only preserve key sentences in the product description.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 2776–2794

Previous achievements to address this problem usually employ fusion modules to learn expressive multimodal representations for prediction (Arevalo et al., 2017; Chen et al., 2018; Liu et al., 2021; Han et al., 2022; Nguyen et al., 2022). Despite the gained satisfying results, there are still several drawbacks in those models that limit the system's performance. First, explicit multi-scale modeling is missing. Extant works usually take into account single or combinations of segments in a *fixed scale*, such as tokens, phrases and image patches. Nevertheless, multi-scale modeling is necessary especially when confronted with long textual inputs like reviews, since it has been pointed out that task-related information is commonly distributed unevenly among all the sentences (Chen et al., 2019). Take a randomly picked product and attached review in Table 1 as an example, even a review of high helpfulness score (Review 1), there are many dispensable sentences (unbold text) that stray from the product it comments on (off-the-topic). Second, though fusion-based models have been demonstrated effective in a family of multimodal tasks, they often result in bulk structures and hence are time-consuming in the training process (Nagrani et al., 2021). Previous research reveals that semantic matching, i.e., the similarity between semantic elements (image regions, text tokens and their n-grams) can be regarded as a crucial factor that guide models to make the final decision (Ma et al., 2015; Huang et al., 2017; Liu et al., 2017). Based on the discovery, we postulate that quantified matching scores could be fully exploited in regression. Specifically, in MRHP task, the matching extent between the review and product description, and inside a review itself (i.e., whether the text and image of a review express a similar meaning) impact how customers rate that review—one could probably not contribute kudos unless he finds product-related contents in that review—such contents subsume the confirmation to the seller's claims, complementary illustration of the product's characteristics, and precautions for the usage, etc.

Based on these two observations and inspired by the idea from relation-based learning (Snell et al., 2017; Sung et al., 2018), we devise a simple yet effective model, PREMISE (PREdict with MatchIng ScorEs) for MRHP tasks. PREMISE gets rid of classic fusion-based architecture and use the matching scores between different modalities and fields in various scales as the feature vectors for regression. Meanwhile, we harness the theory of contrastive learning (Oord et al., 2018; He et al., 2020; Chen et al., 2020a) to further boost the model's performance as it has similar mathematical interpretations of relation-based learning. To our best knowledge, this is the first work that dedicates to utilizing semantic matching scores as logits for classification. The contributions of our work are summarized as follows:

- We propose PREMISE, a model purely based on semantic matching scores of multi-scale features for the multimodal review helpfulness prediction task. PREMISE can produce multimodal multi-field and multi-scale matching scores as expressive features for MRHP tasks.
- We design a new functional architecture named aggregation layer, which receives features from a smaller scale and outputs combinations of features in the same scale, plus the counterparts in a larger scale.
- We conduct comprehensive experiments on several benchmarks. The results compared with several strong baselines show the great advantage and efficiency of exploiting semantic matching scores for the MRHP task.

2 Related Work

Multimodal Representation Learning The fundamental solutions of current multimodal tasks focus on multimodal representation learning, which dedicates to extracting and integrating task-related information from the input signals of many modalities (Atrey et al., 2010; Ngiam et al., 2011). Recently, multimodal fusion technique becomes the predominant method for expressive representation learning, which coalesces a set of multimodal inputs by mathematical operations (e.g., attention and Cartesian product) (Liu et al., 2018; Tsai et al., 2019; Mohla et al., 2020; Hazarika et al., 2020; Han et al., 2021b). Though showing exceptional performance on those tasks, stacked attention architecture also consumes huge computational power and slows down the training and inference speed. To alleviate this issue, we devise a fusion-free model for the MRHP task, which escapes from the conventional fusion-based routine.

Relation-based Learning The idea of relation-based learning was firstly applied in the few-shot image classification task (Vanschoren, 2018;

Hospedales et al., 2020). Vinyals et al. (2016) employs quantified similarity values between the unseen test images and seen trained images to perform classification. Prototypical networks (Snell et al., 2017) and Relation Network Sung et al. (2018) further treats the correlation matrices between images and pre-computed prototypical feature vectors as logits and optimize them to improve the model's performance. Lifchitz et al. (2019) substitutes the comparison target with implanting weights for better generalization ability. Later achievements stemming from this theory encompass building up network structures for interactions between samples (Garcia and Bruna, 2017; Kim et al., 2019), incorporating small-scale computation units like image pixels (Chang and Chen, 2018; Si et al., 2018; Hou et al., 2019; Min et al., 2021) and adding correlation matrices as regularization terms (Wertheimer et al., 2021). In the multimodal scenario, semantic matching has also been chosen as the core task to pretrain large multimodal models (Chen et al., 2020b; Kim et al., 2021; Radford et al., 2021). We inherit this idea to develop a matching-based approach for the MRHP task. In our network, sorted matching scores between vectors of different scales and modalities are shaped into regression features. We will show this canonical formulation beats many fusion-based strong baselines.

3 Method

In this section, we first illustrate the problem definition of Multimodal Review Helpfulness Prediction (MRHP). Then we elaborate on the model architecture and training process.

3.1 Problem Definition

Given N product descriptions $\mathcal{P} = \{P_1, P_2, ..., P_N\}$ and their associated review sets $\mathcal{R} = \{R_1, R_2, ..., R_N\}$, where the review set R_i contains m_i pieces of review $R_i = \{r_{i,1}, r_{i,2}, ..., r_{i,m_i}\}$. Both the product descriptions and review pieces are presented in the modality of text $T_{p_i/r_{i,k}}$ and image $I_{p_i/r_{i,k}}$. MRHP aims to predict the helpfulness scores of reviews $\{y_{i,k}\}_{k=1}^{m_i}$ and rank these reviews according to the scores in descending order so that favorable reviews can be promoted to the top. For the simplicity of the statement, we call the product description and review *field*, denoted by the superscripts $f \in \{p, r\}$. Similarly the superscripts $m \in \{t, v\}$ refer to the *modality* of text and image (vision).

3.2 Overview

We depict the overall architecture of our model in Figure 1. At the bottom layer, the modality-specific encoders are pretrained models or word vectors that map the raw inputs into continuous embeddings. The initially embedded representations are viewed as the minimal scale to be aggregated by PREMISE. For example, they are word vectors if the encoder is Glove (Pennington et al., 2014), contextualized word representations if applying BERT (Devlin et al., 2018) or other pretrained language models, and detected hot regions in an image when adopting FastRCNN (Girshick, 2015). These representations are then passed through N stacked aggregation layers where representations from a smaller scale are collated into larger-scale counterparts. Finally, PREMISE computes the matching scores between these multi-scale feature vectors and performs regression with the sorted top-K scores.

3.3 Input Feature

Textual Representation We initialize the token representations of text in the product and review fields with word vectors or pretrained models as $\mathbf{E}'_t = \{e_1'^t, e_2'^t, ..., e_l'^t\}$, where l is the length of a review sentence. For word vector embeddings, we additionally exploit a Gated Recurrent Unit (GRU) (Cho et al., 2014) layer on each sentence to obtain the context-aware token-level representations $\mathbf{E}_t = \{e_1^t, e_2^t, ..., e_l^t\}$, where θ_t denotes the parameters of the GRU.

Visual Representation We embed images with pretrained Faster R-CNN (Ren et al., 2015), which utilizes ResNet-101 as its backbone, yieding the visual feature input $\mathbf{E}_v = \{e_1^v, e_2^v, ..., e_{n_h}^v\}$. where θ_v denotes the parameters in FastRCNN and n_h is the number of hot regions detected in the given image.

3.4 Multi-Scale Matching Network (MSMN)

The inspiration beneath MSMN is from the pyramid and network-in-network architectures (Lazebnik et al., 2006; Han et al., 2021a) and relation-based learning (Snell et al., 2017; Sung et al., 2018).

Multi-scale Feature Generation MSMN consists of several structurally-identical aggregation

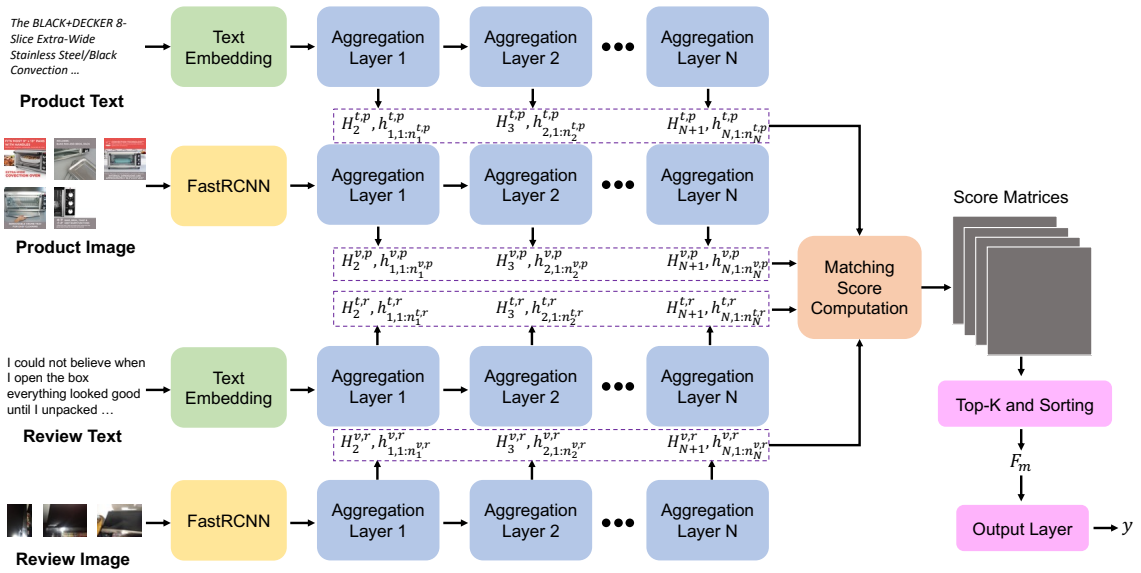


Figure 1: The overall architecture of PREMISE. We hide the data frame reorganization process betwixt two aggregation layers that merge the produced larger-scale representations into a new sequence and only show the outputs from a single block of data with the subscript i omitted for simplicity.

layers that upscale the input sample hierarchically. An aggregation layer can be further divided into many aggregation blocks, as pictured in Figure 2, each of which receives the outputs produced from the last layer to generate both the combined representations at the k-th scale $h_{1:n_k} = \{h_1, h_2, \ldots, h_{n_k}\}$ of length n_k (for better readablitiy we omit the superscript flags of modality and field) and an aggregated representation at the $k+1$-th (next larger) scale H_{K+1}:

$$\mathbf{V}_{k,i} = \{H_{k,1}, \ldots, H_{k,n_k}\} \quad (1)$$

$$H_{k+1,i}, h_{1:n_k,i} = \mathbf{Aggr}_i(\mathbf{V}_{k,i}; \theta_i) \quad (2)$$

where $\mathbf{V}_{k,i}$ is the collection of the aggregated representations from the k-th layer and the subscript i indexes the aggregation block that processes the sequence i of an input instance in the k-th layer. The output representations are all collected to calculate matching scores later, and the upscaled representations are meanwhile gathered as the input sequences for the next layer. Following Han et al. (2021a), we enforce these internal blocks to share parameters, i.e., $\theta_1 = \theta_2 = \ldots = \theta_{n_k}$. In our formulation, we set $N = 2$ to endow these scales with realistic meanings (from $k = 0$ to $k = 2$)—"word → sentence → the entire review/description" for text and "hot region→ image → the entire review/description" for images.

We adopt Transformer (Vaswani et al., 2017) as the basic architecture for aggregation layers. For each layer, We feed the sequential representations from the last layer (after adding the `[CLS]` token to their heads) into the current layer and extract the heads of the output as the next-scale representations that serve as the input to the next layer:

$$[H_{k+1,i}, h_{1:n_k,i}] = \mathbf{Transformer}(\mathbf{V}_{k,i}; \mathbf{\Theta}) \quad (3)$$

where Θ denotes the transformer parameters.

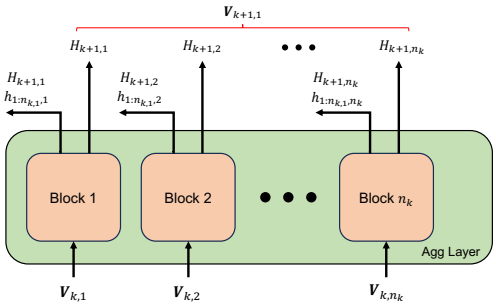


Figure 2: The inner structure of an aggregation layer.

Semantics Refinement In the lower layer where the feature scale is small and dense, there are closed semantic units, which result in many duplicated matching scores and impair the ability of prediction network (as we show in the next section) To address this problem, we aim to filter the extremely long sequences of output features, only to maintain the dominant components. The algorithm is based on a faster k-means algorithm, which can produce a set of representations by clustering adjacent points so that redundant semantics are eliminated. To reduce extra overhead, we implement an approximate but faster algorithm (Hamerly, 2010) only when

the feature number exceeds a threshold. To prevent each clustered set from being too small (i.e., only 1 or 2 points) and ensure efficiency, we random sample C points as centers. The algorithm is formally depicted in Algorithm 1, where we omit the specific steps of fast k-means and readers can refer to the paper (Hamerly, 2010) for the details. It should be emphasized that k-means is a non-parametric clustering algorithm and only incurs negligible overhead in the forward pass.

Algorithm 1: Semantics Refinement

Input: Semantic elements set $S = \{s_1, s_2, ..., s_N\}$, expected cluster size r, number of centers C,
Output: Refined set $S' = \{s'_1, s'_2, ..., s'_C\}$

```
if N ≤ C × r then
    return RandomSample (S, C)
else
    return k-Means(RandomInit(C), S)
end
```

In the algorithm, C bounds the lowest number of centers required for the later computation and a typical value could be $C = \lceil\sqrt{K}\rceil$ where K is the hyperparameter in the last layer's feature selector (as stated below) implicitly and roughly. The heuristics in using squareroot comes from the fact that similarity scores are computed in pair. Here r is another important hyperparameter that controls the expected (or average) cluster size to avoid overmuch small clusters.

Prediction After obtaining the representations from both fields and modalities in all scales $(H_k, h_{1:n_k})$, we concatenate them into four matrices $\mathbf{R}^{t,p}, \mathbf{R}^{t,r}, \mathbf{R}^{v,p}, \mathbf{R}^{v,r}$ whose rows are these representation vectors. Concretely, we extract the n-gram token, sentence, and n-gram sentence representations for text, and n-gram RoI and image representations for the image. We hypothesize that the review quality depends on the semantic coherence existing within 1) The image-text pair in the review to guarantee the coherence of the review itself. Low-quality reviews are usually not self-contained ($\mathbf{R}^{t,r}, \mathbf{R}^{v,r}$). We exclude the scores of image-text pairs from product introduction because they do not have any impact to the helpfulness of a review. 2) The same modality from different fields ($\mathbf{R}^{t,p}, \mathbf{R}^{t,r}$ and $\mathbf{R}^{v,p}, \mathbf{R}^{v,r}$). This is important because user-preferred comments should directly response to the selling points in the introduction.

The matching scores are calculated as the cosine similarities between row vectors of two matrices:

$$\mathbf{S}(\mathbf{A}, \mathbf{B}) = \mathbf{cosine}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A}\mathbf{B}^T}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|^T} \quad (4)$$

where $\|\cdot\|$ is the row-wise L2 normalization. Suppose there are n_1, n_2, n_3, n_4 rows (the number of feature vectors) in the four matrices, then we have $n_1n_2 + n_2n_4 + n_3n_4$ matching scores. We then picked the highest K scores to form the last features. It should be emphasized that the top-K operation reorganizes the tensors during its computation.

$$\mathbf{h} = \mathbf{TopK}(\mathbf{FlattenAll}(\mathbf{S}')) \quad (5)$$

Therefore the gradient back-propagation path is not constant given different input samples. The predictions for training and inference are calculated from the feature:

$$\mathbf{f}_{i,j} = \sigma(\mathrm{Linear}(\mathbf{h}_{i,j})) \quad (6)$$

where σ is the Sigmoid function.

3.5 Training

We follow Nguyen et al. (2023) to apply the listwise loss for training.

$$\mathcal{L} = -\sum_{i=1}^{|P|}\sum_{j=1}^{|R_i|} y'_{i,j} \log(f'_{i,j}) \quad (7)$$

where $|\mathcal{P}|$ is the number of productions in the batch and $|R_i|$ is the number of reviews corresponding to P_i. The normalized labels y' and predictions f' are given by

$$f'_{i,j} = \mathrm{softmax}(\mathbf{f}_i)_j, y'_{i,j} = \mathrm{softmax}(\mathbf{y}_i)_j \quad (8)$$

Note that the final predictions are ranged within $(0, 1)$, which diverges from the true label distribution $y \in [0, 4]$. However, the ultimate target (same as the evaluation metrics) of the task concentrates on *ranking* (relative value) rather than absolute value. This kind of learning can still benefit the performance.

4 Experiments

4.1 Datasets and Metrics

We evaluate our model on two MRHP datasets (Liu et al., 2021), each of which subsumes same three categories: *Clothing, Shoes & Jewelry, Home & Kitchen* and *Electronics*. We train and test both datasets on a single NVIDIA RTX A6000 GPU.

Model	#Params	Statistical Learning	Fusion	Matching Score as Logits
SSE-Cross	—	✗	✓	✗
D&R Net	—	✓	✓	✗
MCR	2.33M	✗	✓	✗
SANCL	2.38M	✓	✓	✗
CMCR	2.41M	✗	✓	✗
GBDT	21.8M	✗	✓	✗
PREMISE	2.28M	✗	✗	✓

Table 2: Comparison between baseline models and ours.

The gradients are calculated and backpropagated for each batch in a single forward pass, without batch division and gradient accumulation. We compare our model with several baseline models on three common metrics for ranking tasks: the mean average precision (MAP), the N-term ($N = 3, 5$ in our experiment in accord with previous works) Normalized Discounted Cumulative Gain (NDCG@N) (Järvelin and Kekäläinen, 2017; Diaz and Ng, 2018). The helpfulness scores are labeled as the logarithm of approval votes to the corresponding reviews and are clipped into integers within $[0, 4]$. The statistics of datasets and more training details are provided in appendix.

4.2 Baselines

We compare our model with the following baselines: Stochastic Shared Embeddings enhanced cross-modal network (SSE-Cross) (Abavisani et al., 2020), Decomposition and Relation Network (D&R Net) (Xu et al., 2020). The Multimodal Coherence Reasoning network (MCR) (Liu et al., 2021) designs several reasoning modules based on fused representations for prediction. SANCL (Han et al., 2022) and contrastive-MCR (CMCR) (Nguyen et al., 2022) minimize auxiliary contrastive loss to refine the multimodal representations. Gradient-boosted decision tree (GBDT) (Nguyen et al., 2023) design a random walk policy and aggregate the helpfulness scores through from tree leaves—the endpoints of the random walk.

To provide a holistic view on the distinctions between the learning paradigms of these models, we list and compare three key characteristics between PREMISE and baselines in Table 2. From the table we observe that all baseline models contain fusion modules inside the entire structures. Moreover, D&R Net and SANCL also incorporate extra statistical correlations (Adjective-Noun Pairs and selective-attention mask creation) that inject external knowledge to bridge the semantic gap between textual and visual modality or cast more focus on contents that perceived important by human beings. Our model escapes from both complicated manually crafted features and conventional model architecture by directly computing the matching scores and automatically picking the K highest ones as features for regression.

4.3 Results

We run our models three times and report the average performance in Table 3 and 4. It can be clearly seen that our model outperforms all these baselines on two datasets. Particularly, compared with the strongest baseline—GBDT (Nguyen et al., 2023), PREMISE gains over 5 points improvement on MAP and NDCG@5 and 10 points improvement on NDCG@3 on Amazon-MRHP dataset, and 6.8∼17.5 improvement on all metrics on Lazada-MRHP datasets. When using BERT to initialize embeddings, we note a slight performance degradation compared to the implementations that use GloVe as embeddings in both PREMISEand other baselines. Such outcome demonstrates the superiority of our fusion-free model and, at least in the MRHP task, multimodal fusion is not a necessity and may hinder the model from better performance.

Besides, we highlight the size of the feature vectors in the last layer. Fusion-based baselines usually concatenate representations from both fields and modalities to perform final regression, which requires at least 512 (128×4) dimensions of feature vectors. The vector is even longer in MCR and SANCL (over 1000) since there are many extra features taken into account. Nevertheless, the feature vector lengths for the best performance in PREMISE are apparently smaller. As shown in Figure 3, the optimal choices of K range from 64 to 128. This fact manifests that there could be many redundant elements in the vectors generated by fusion-based models, and PREMISE successfully enhance the efficiency of unit-length features through a simple representation learning policy.

4.4 Ablation Study

We run our models several ablative settings under feature selection. These settings are all corresponding to excluding some features from different scales when computing the multi-scale matching scores; During implementation, we substitute these representations with zero vectors in equal sizes, which amounts to masking them out.

Model	Cloth. & Jew.			Electronics			Home & Kitchen		
	MAP	**N@3**	**N@5**	**MAP**	**N@3**	**N@5**	**MAP**	**N@3**	**N@5**
SSE-Cross	65.0	56.0	59.1	53.7	43.8	47.2	60.8	51.0	54.0
D&R Net	65.2	56.1	59.2	53.9	44.2	47.5	61.2	51.8	54.6
MCR	66.4	57.3	60.2	54.4	45.0	48.1	62.6	53.5	56.6
SANCL	67.3	58.6	61.5	56.2	47.0	49.9	63.4	54.3	57.4
CMCR	67.4	58.6	61.6	56.5	47.6	50.8	63.5	54.6	57.8
GBDT	82.6	80.3	79.3	74.2	68.0	69.8	81.7	76.5	78.8
PREMISE (Ours)	**92.3**	**90.4**	**91.5**	**81.4**	**78.6**	**75.6**	**88.6**	**88.3**	**88.4**
MCR+BERT	65.8	55.9	58.8	55.9	46.8	49.4	62.4	52.9	56.1
GBDT+BERT	80.3	78.7	77.1	73.8	68.3	69.5	81.4	76.9	79.4
PREMISE +BERT (Ours)	**91.5**	**89.7**	**90.1**	**79.3**	**77.7**	**78.6**	**87.7**	**85.9**	**86.2**

Table 3: Results on the Amazon-MRHP (English) dataset. All reported metrics are the average of five runs. Baseline results are from Nguyen et al. (2023). PREMISE outperform the strongest baseline with p-value<0.05 based on the paired t-test.

Model	Cloth. & Jew			Electronics			Home & Kitchen		
	MAP	**N@3**	**N@5**	**MAP**	**N@3**	**N@5**	**MAP**	**N@3**	**N@5**
SSE-Cross	66.1	59.7	64.8	76.0	68.9	73.8	72.2	66.0	71.0
D&R Net	66.6	61.3	65.8	76.5	69.5	74.3	72.7	66.7	71.8
MCR	68.8	62.3	67.0	76.8	70.7	75.0	73.8	67.0	72.2
SANCL	70.2	64.6	68.8	77.8	71.5	76.1	75.1	68.4	73.3
CMCR	70.3	64.7	69.0	78.2	72.4	79.6	75.2	68.8	73.7
GBDT	78.5	77.1	79.0	87.9	86.7	88.1	85.6	78.8	83.1
PREMISE (Ours)	**95.3**	**94.6**	**95.0**	**96.9**	**96.1**	**96.8**	**93.9**	**91.5**	**92.8**

Table 4: Results on the Lazada-MRHP (Indonesian) dataset.

The results are summarized in Table 5, from which we have the following discoveries. First, discarding representations of any scale causes degradation in the model's performance, indicating that all these chosen features contribute to the accurate prediction. Besides, the performance plummets more severely when removing features of smaller scales (or in bottom layers), including single scale (e.g., "n-gram token repr" v.s. "n-gram sent repr", "n-gram RoI repr" v.s. "n-gram image repr") or the combinations (e.g., "n-gram token & n-gram RoI repr" v.s. "n-gram sent repr & image repr"). This outcome reveals that lower-level feature is more fundamental to the model's performance than higher ones, since a large portion of matching scores are computed from them.

5 Analysis

5.1 The Impact of Selected Feature Numbers

A unique hyperparameter in PREMISE could be K for the selection of last features which determines how many highest scores to be included to form the final feature vectors. To explore how K affects the model's performance, we run our model with various values of K and plot how the performance changes in Figure 3. It can be found that to achieve ideal performance in both datasets, an appropriate choice of K (from 64 to 128) is necessary. When setting K too high or too low, i.e., dispersing the model's concentration on too many scores or forcing the model to focus on only a few highest scores, the model fails to reach the optimum. This phenomenon reveals that a promising filter should provide comprehensive coverage of the matching scores for the prediction layer.

5.2 The Impact of Lower Layer Filter

Apart from the last-layer feature selector, we also insert many filters into the lower layers. To verify the efficacy of this design, we performed additional experiments by varying r_{min} in algorithm 1 while fixing $K = 96$ and $C = \lceil \sqrt{K} \rceil = 10$. The results on the two datasets are shown in Figure 4, from which we notice that in all categories, a proper

Description	MAP	N@3	N@5
PREMISE (Amazon)	**87.4**	**85.8**	**86.5**
-w/o n-gram token repr	84.8	82.1	83.3
-w/o sent repr	86.2	84.3	85.0
-w/o n-gram sent repr	85.7	83.9	84.6
-w/o n-gram RoI repr	83.9	81.8	82.6
-w/o image repr	86.5	84.1	85.3
-w/o n-gram token & n-gram RoI repr	75.3	69.8	72.2
-w/o n-gram sent repr & image repr	84.1	82.5	83.0
PREMISE (Lazada)	**95.4**	**94.0**	**94.9**
-w/o n-gram token repr	91.0	88.9	89.5
-w/o sent repr	93.5	91.6	92.2
-w/o n-gram sent repr	94.3	92.9	93.8
-w/o n-gram RoI repr	92.7	90.1	91.7
-w/o image repr	94.8	94.2	94.6
-w/o n-gram token & n-gram RoI repr	80.1	78.6	79.5
-w/o n-gram sent repr & image repr	92.3	89.6	90.1

Table 5: Ablation experiments of PREMISE on two datasets. The values are averaged over all three categories.

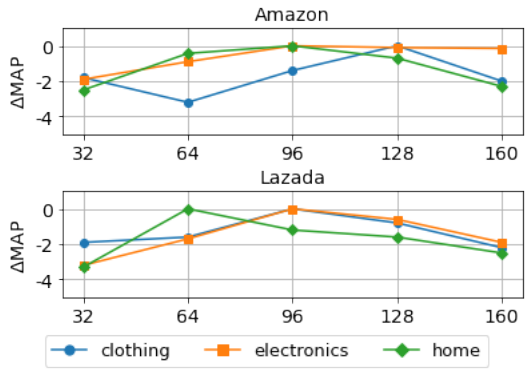


Figure 3: The relative MAP drop (the absolute value of ΔMAP) from the optim to different K. Performance when $K > 160$ or $K < 32$ is far lower than the optimum so we do not include in the figure.

choice of r value ($k = 4$ in our experiments) can further enhance the performance by removing duplicated semantics in lower aggregation layers. This suggests that the semantics redundancy removal procedure the combination of k-means and random sampling can serve as a primary filter for the feature selection.

5.3 Why Does BERT Fail?

As mentioned above, it is weird that after replacing the word vectors (GloVe and FastText) with the pretrained language model in the embedding layer, both the fusion-based and fusion-free models fail to produce a significant increase as in other multimodal tasks. We surmise that this is mainly due to informal text input. Upon manual inspection, we find many pieces of low-quality reviews—especially those of low helpfulness scores. Take review 2 in Table 1 as an example, the review passage is readable by sentence except for some grammatical errors, but the logic is messy and out of the topic. The results of previous work on tasks related to spoken language (informal text) have shown that BERT may not lead to a performance improvement (Gu and Yu, 2020). To further verify our hypothesis, we carry out a group of blank control experiments on Amazon-MRHP dataset. Specifically, we run regression directly on: A) representations encoded by a single layer GRU with Glove 300d as word embeddings in both fields; B) the representations at the position of [CLS] token using BERT-base-uncased as the pretrained encoder. The results are shown in Table 6. From the table we find the performance between word vectors and BERT pretrained models are very close. This outcome looks consistent with the results in Gu and Yu (2020) and may substantiate our aforementioned hypothesis.

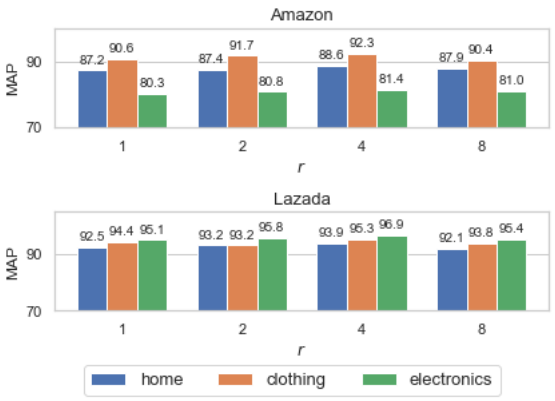


Figure 4: The performance (MAP) under different r on two datasets.

Category	Setting	MAP	N@3	N@5
Clothing	A	64.83	55.62	58.95
	B	64.75	55.51	59.03
Electronics	A	53.63	43.77	47.31
	B	53.90	43.85	47.02
Home	A	61.08	51.17	54.26
	B	61.03	51.09	54.14

Table 6: A group of blank control experiments on Amazon-MRHP dataset.

6 Conclusion

In this work, we propose a novel matching-based learning model, PREMISE, for the task of mul-

timodal review helpfulness prediction (MRHP). PREMISE calculates matching scores between refined semantics across modalities and data fields for fast and accurate regression and ranking. Experiments and analysis demonstrate that our model exceeds many strong fusion-based approaches, which provides a possible idea for such kind of tasks.

Limitations

The major limitation of PREMISE is its applicable scenarios or restricted adaptation ability to other multimodal tasks. Ideally, we expect PREMISE to behave as a generic model that can also work on many other multimodal tasks, but now we have only empirically demonstrated its efficacy in the MRHP task. Intuitively, we believe that at least in the tasks where the extent of semantic matching matters, our method should produce satisfying results, e.g., multimodal (image/text) retrieval and sarcasm detection where low correlation usually implies that sarcasm exists. But currently we only yield fair results that fall behind the current SOTA significantly on the aforementioned tasks (see appendix for details).

Another limitation is the efficiency. We actually adopt a brute-force computing strategy, which can be further improved through more careful module design. We hold this as our future potential direction to work on.

References

Mahdi Abavisani, Liwei Wu, Shengli Hu, Joel Tetreault, and Alejandro Jaimes. 2020. Multimodal categorization of crisis events in social media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14679–14689.

Viviana Alfonso, Codruta Boar, Jon Frost, Leonardo Gambacorta, and Jing Liu. 2021. E-commerce in the pandemic and beyond. *BIS Bulletin*, 36(9).

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. 2017. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*.

Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. 2010. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16(6):345–379.

Nils Boysen, René De Koster, and Felix Weidinger. 2019. Warehousing in the e-commerce era: A survey. *European Journal of Operational Research*, 277(2):396–411.

Jia-Ren Chang and Yong-Sheng Chen. 2018. Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418.

Cen Chen, Minghui Qiu, Yinfei Yang, Jun Zhou, Jun Huang, Xiaolong Li, and Forrest Sheng Bao. 2019. Multi-domain gated cnn for review helpfulness prediction. In *The World Wide Web Conference*, pages 2630–2636.

Cen Chen, Yinfei Yang, Jun Zhou, Xiaolong Li, and Forrest Bao. 2018. Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 602–607.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. 2021. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gerardo Ocampo Diaz and Vincent Ng. 2018. Modeling and prediction of online product review helpfulness: a survey. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 698–708.

Victor Garcia and Joan Bruna. 2017. Few-shot learning with graph neural networks. *arXiv preprint arXiv:1711.04043*.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Jing Gu and Zhou Yu. 2020. Data annealing for informal language understanding tasks. *arXiv preprint arXiv:2004.13833*.

Greg Hamerly. 2010. Making k-means even faster. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 130–140. SIAM.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. 2021a. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021b. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 6–15.

Wei Han, Hui Chen, Zhen Hai, Soujanya Poria, and Lidong Bing. 2022. SANCL: Multimodal review helpfulness prediction with selective attention and natural contrastive learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 5666–5677, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1122–1131.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2020. Meta-learning in neural networks: A survey. *arXiv preprint arXiv:2004.05439*.

Ruibing Hou, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. 2019. Cross attention network for few-shot classification. *Advances in Neural Information Processing Systems*, 32.

Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318.

Kalervo Järvelin and Jaana Kekäläinen. 2017. Ir evaluation methods for retrieving highly relevant documents. In *ACM SIGIR Forum*, volume 51, pages 243–250. ACM New York, NY, USA.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Jongmin Kim, Taesup Kim, Sungwoong Kim, and Chang D Yoo. 2019. Edge-labeling graph neural network for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11–20.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR.

Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 2169–2178. IEEE.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662.

Yann Lifchitz, Yannis Avrithis, Sylvaine Picard, and Andrei Bursuc. 2019. Dense classification and implanting for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9258–9267.

Junhao Liu, Zhen Hai, Min Yang, and Lidong Bing. 2021. Multi-perspective coherent reasoning for helpfulness prediction of multimodal reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5927–5936.

Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4107–4116.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE international conference on computer vision*, pages 2623–2631.

Juhong Min, Dahyun Kang, and Minsu Cho. 2021. Hypercorrelation squeeze for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6941–6952.

Satyam Mohla, Shivam Pande, Biplab Banerjee, and Subhasis Chaudhuri. 2020. Fusatnet: Dual attention based spectrospatial multimodal fusion network for

hyperspectral and lidar classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 92–93.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*.

Thomas L Ngo-Ye and Atish P Sinha. 2014. The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems*, 61:47–58.

Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Anh Tuan Luu, Cong-Duy Nguyen, Zhen Hai, and Lidong Bing. 2023. Gradient-boosted decision tree for listwise context model in multimodal review helpfulness prediction. *arXiv preprint arXiv:2305.12678*.

Thong Nguyen, Xiaobao Wu, Anh Tuan Luu, Zhen Hai, and Lidong Bing. 2022. Adaptive contrastive learning on multimodal transformer for review helpfulness prediction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10085–10096, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C Kot, and Gang Wang. 2018. Dual attention matching network for context-aware feature sequence based person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5363–5372.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

Jiliang Tang, Huiji Gao, Xia Hu, and Huan Liu. 2013. Context-aware review helpfulness rating prediction. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 1–8.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Joaquin Vanschoren. 2018. Meta-learning: A survey. *arXiv preprint arXiv:1810.03548*.

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

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Nir Vulkan. 2020. *The Economics of E-commerce*. Princeton University Press.

Zheng Wang, Zhenwei Gao, Kangshuai Guo, Yang Yang, Xiaoming Wang, and Heng Tao Shen. 2023. Multilateral semantic relations modeling for image text retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2830–2839.

Davis Wertheimer, Luming Tang, and Bharath Hariharan. 2021. Few-shot classification with feature map reconstruction networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8012–8021.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786.

A Dataset Specifications

We list the specifications of train/validation/test split of the two datasets (six categories) in Table 7 and Table 8. The numbers "X/Y" represent that the split contains X product descriptions and Y reviews. Amazon-MRHP is an pure English dataset, while Lazada-MRHP is written in

Indonesian so that we do not conduct BERT related experiments on it.

Amazon-MRHP (Products/Reviews)			
Cat.	Cloth. & Jew.	Elec.	Home & Kitch.
Train	12,074/277,308	10,564/240,505	14,570/369,518
Dev	3,019/122,148	2,641/84,402	3,616/92,707
Test	3,966/87,492	3,327/79,750	4,529/111,593

Table 7: Statistics of the Amazon-MRHP dataset.

Lazada-MRHP (Products/Reviews)			
Cat.	Cloth. & Jew.	Elec.	Home & Kitch.
Train	6,596/10,4093	3,848/41,828	2,939/36,991
Dev	1,649/26,139	963/10,565	736/9,611
Test	2,062/32,274	1,204/12,661	920/12,551

Table 8: Statistics of the Lazada-MRHP dataset.

B Training Details

B.1 Initialization of Embeddings

To stay consistent with previous works, we embed the text input of Amazon-MRHP with GloVe-300d (Pennington et al., 2014) and Lazada-MRHP with Fasttext (Joulin et al., 2016), respectively. In BERT-related experiments we employ the Huggingface toolkit for pretrained models[1].

B.2 Hyperparameter Search space

The optimal hyperparameter settings are listed in Table 9, 10 and 11. The search space of these hyperparameters are: learning rate in $\{1e^{-4}, 5e^{-4}\}$, text embedding dropout fixed at $\{0.2\}$, shared space hidden dimension in $\{128, 256\}$.

Amazon-MRHP Glove Hyperparameters			
	Cloth. & Jew.	Elec.	Home & Kitch.
learning rate	$1e^{-4}$	$5e^{-4}$	$5e^{-4}$
text embedding dim	300	300	300
text embedding dropout	0.2	0.2	0.2
image embedding dim	256	256	256
text embedding dim	128	128	128
shared space hidden	128	128	128
r	4	4	4
K	96	128	64
batch size	32	32	32

Table 9: Hyperparameters for Amazon-MRHP using glove-300d embeddings.

[1]https://huggingface.co/docs/transformers/index

Lazada-MRHP fastText Hyperparameters			
	Cloth. & Jew.	Elec.	Home & Kitch.
learning rate	$1e^{-4}$	$5e^{-4}$	$1e^{-4}$
text embedding dropout	0.2	0.2	0.2
image embedding dim	256	256	256
text embedding dim	128	128	128
shared space hidden	128	128	128
r	4	4	4
K	96	96	128
batch size	32	32	32

Table 10: Hyperparameters for Lazada-MRHP using fasttext embeddings.

Amazon-MRHP BERT Hyperparameters			
	Cloth. & Jew.	Elec.	Home & Kitch.
learning rate	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$
bert learning rate	$1e^{-5}$	$1e^{-5}$	$1e^{-5}$
text embedding dropout	0.2	0.2	0.2
image embedding dim	512	512	512
text embedding dim	256	256	256
shared space hidden	256	256	256
r	4	4	4
K	128	128	128
batch size	16	16	8

Table 11: Hyperparameters for all categories using BERT as encoder

B.3 Sampling of Production Description-Review Pairs in Training

We mentioned in §3.5 that the training pairs are sampled from the training set. Now we describe how do we sample these training pairs. First, we sample B products from the training set where B is the batch size. Next, for each product, we randomly sample one of its positive review (rating is greater than 2 and N_r^- negative reviews (rating is less than or equal to 2 from the corresponding review set. The dataset has been filtered during manufacture time so that there is always at least one positive/negative review under each product. To put it in a nutshell, a sampled batch contains B product descriptions, B positive reviews and $N_r^- B$ negative reviews.

C The Differentiability of top-K Operation in PyTorch

Given a vector $S = \{s_1, s_2, ..., s_L\} \in \mathbb{R}^L$ where L is the length of that vector, when passing through the top-K operation, most fundamentally its largest K values are selected and sorted in descending order to form a new vector $T = \{t_1, t_2, ..., t_K\} \in \mathbb{R}^K$. Suppose the indices of T's elements in S are $I = \{i_1, i_2, ..., i_K\} \subset \{1, 2, ..., L\}$, then the pro-

cess equals to that concurrently there is a mask $M \in \{0,1\}^L$ automatically created and "multiplies" on S. Each value m_j in M is

$$m = \begin{cases} 1, & if \quad j \in I \\ 0, & if \quad j \notin I \end{cases}$$

With this mask, the forward and backward propagation can proceed as in conventional routines.

D Training Speed

D.1 Theoretical Analysis

For simplicity, we consider the case of a pair of modality sequence. Let $X_1 \in \mathbb{R}^{l_1 \times d}$, $X_2 \in \mathbb{R}^{l_2 \times d}$ be a pair of input modality sequences. Here we assume they have been both projected to the same dimension as a common practice that both the fusion–prediction routines and our matching approach exercise. The multihead attention operation in fusion-based models can be written as:

$$X_{12} = Att(X_1, X_2) \tag{9}$$

Note that attention is a *directional* operation, i.e. $Att(M_1, M_2) \neq Att(M_2, M_1)$. Due to this, a fusion-based learning model $\mathcal{M}$ always adopts a pair of conjugate attention. Therefore, for a model of N layers, the total computational complexity C_f is:

$$C_f = (2l_1l_2 + l_1^2 + l_2^2)Nd \tag{10}$$

Now consider matching-based models. We only have self-attention for each modality per layer. The whole computational complexity consists of the self-attention (att) and multi-scale matching score (mm).

$$C_m = C_{att} + C_{mm} \tag{11}$$

Since the number of scales decreases as the aggregation proceeds, we denote the decreasing ratio at layer i for modality j as $k_{i,j}$. Hence, the total computational complexity is:

$$C_{att} = 2\sum_p l_p^2 d \left(1 + \frac{1}{k_{p,1}^2} + \frac{1}{k_{1,p}^2 k_{2,p}^2} + ...\right) \tag{12}$$

In our settings, $k_{p,1}$ is large (typically greater than 10), therefore $\frac{1}{k_{p,j}^2} < 0.01$ and can be ignored:

$$C_{att} = 2(l_1^2 + l_2^2)d \tag{13}$$

As for the second term, we have:

$$\begin{aligned} C_{mm} &= l_1\left(\frac{1}{k_{1,1}} + ...\right) l_2 \left(\frac{1}{k_{2,1}} + ...\right) d \\ &< \frac{l_1 l_2 d}{k_{1,1}k_{2,1}(1 - k_1^{-1})(1 - k_2^{-1})} \end{aligned} \tag{14}$$

In the MRHP dataset, $l_1 \approx l_2 = l$. For the typical value $N = 2, k_1 = k_2 = 10$, we have $C_f = 8l^2d$ and $C_m < (4 + \frac{1}{81})l^2d = 4.01l^2d$, or

$$\frac{C_m}{C_f} \approx 0.5 \tag{15}$$

which is closed to the measured acceleration in 5. In fact, let $C_m = C_f$ and $N = 2$, we have $l_1 \approx 2.42l_2$, which seldom happens in the whole dataset. We observe that the number of hot regions is greater than the text length.

D.2 Numerical Results

We measure the average training time per batch of MCR, SANCL (the state-of-the-art baseline) and PREMISE, as shown in Figure 5. The average values are calculated by counting the total time of iterations over 100 batches for 5 random intervals during the whole training process.

Mathematically, denote the counted time of the i^{th} interval as t_i, the speed is calculated as follows

$$speed = \frac{\sum_{i=1}^{5} t_i}{100 \times 5} \tag{16}$$

It can be seen that the training time has been greatly shortened by 42% and 65% compared to SANCL (the fastest baseline, athough they have closed number of parameters as shown in Table 2) and GBDT (the strongest baseline), which approximately matches the conclusion given by mathematical deduction.

E Experiment on Multimodal Retrieval

We test PREMISE on multimodal retrieval (bidirectional) task, the results of both image-to-text and text-to-image retrieval on the MSCOCO test set are shown in Table 12. It can be seen that although our formulation process is completely based on "learning-from-relation" in MRHP task, the constructed model still has some generalizability to other tasks that our hypothesis stands.

Model	Image-to-Text		Text-to-Image	
	PRMP	R@1	PRMP	R@1
VCRN (Li et al., 2019)	29.70	**53.00**	29.90	**40.50**
PCME (Chun et al., 2021)	34.10	41.70	<u>34.40</u>	31.20
MSRM (Wang et al., 2023)	**35.62**	<u>44.32</u>	**35.81**	<u>33.40</u>
PREMISE	<u>34.23</u>	42.06	33.92	31.50

Table 12: Results on MSCOCO-5K test set. The highest values in each metric are in bold, while the second-highest are indicated with an underline.

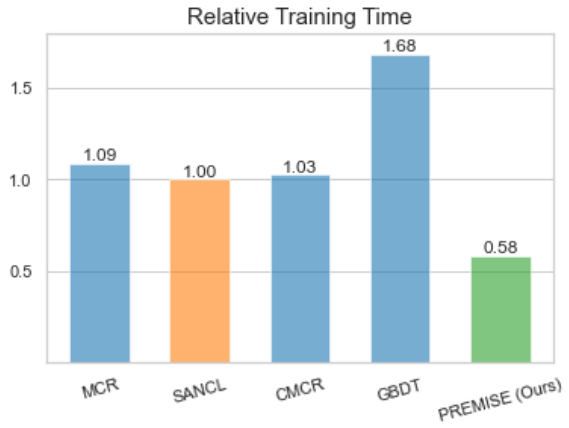


Figure 5: The relative training time of different models. The fastest baseline (SANCL) is highlighted in orange, while our model is highlighted in green. Others are in blue.

F Case Study

F.1 How matching scores affect the prediction?

To further understand the model's functional mechanism, we randomly pick an example from the Amazon–electronics and visualize some matching scores during the test time in Figure 6. There are several valuable points to underline. First, when the model achieves the best performance, its matching scores can reflect the correlation between some semantic matching feature pairs. For instance, the RoI-RoI matching score of -0.17 is produced by the two RoIs that enclose different objects in their respective images, and thus the correlation between them is very weak, and a near 0 score is obtained, while the two boxes that contain the port hub achieve a relatively high matching score. Second, text-text matching may act as word matching. It is hard to attribute the 0.89 matching score of those two sentences to the high semantic similarity between those two text snippets since their semantic meanings are different, only to share some common words. These two discoveries reveal that PREMISE attends to more than semantic matching, and just a certain number of correct matching scores could make up the last features for its accurate prediction, in accord with the conclusion about K values.

F.2 Direct comparison with GBDT

We further randomly draw two examples which PREMISE gives accurate predictions but GBDT, the state-of-the-art baseline, fails. The original review context (including text and attached image) together with the predictions from GBDT and our model towards these two examples are displayed in Table 13 to Table 16.

We find that PREMISE ranks these reviews in correct order (it has the same ranking sequence $1 \rightarrow 2 \rightarrow 3$ and $1 \rightarrow 2 \rightarrow 3 \rightarrow 4$ as the ground truth's in the two examples respectively) even trained and tested with normalized score whose values range from 0 to 1 (different from the annotated scores $s \in [0, 4]$). In the given examples, GBDT flips the order of 'B005NGQWL2-9' and 'B005NGQWL2-65' in example 1 and the order of 'B00H4O1L9Y-111' and 'B00H4O1L9Y-122' in example 2. This could imply that matching-based modeling can make more accurate predictions than fusion-based modeling.

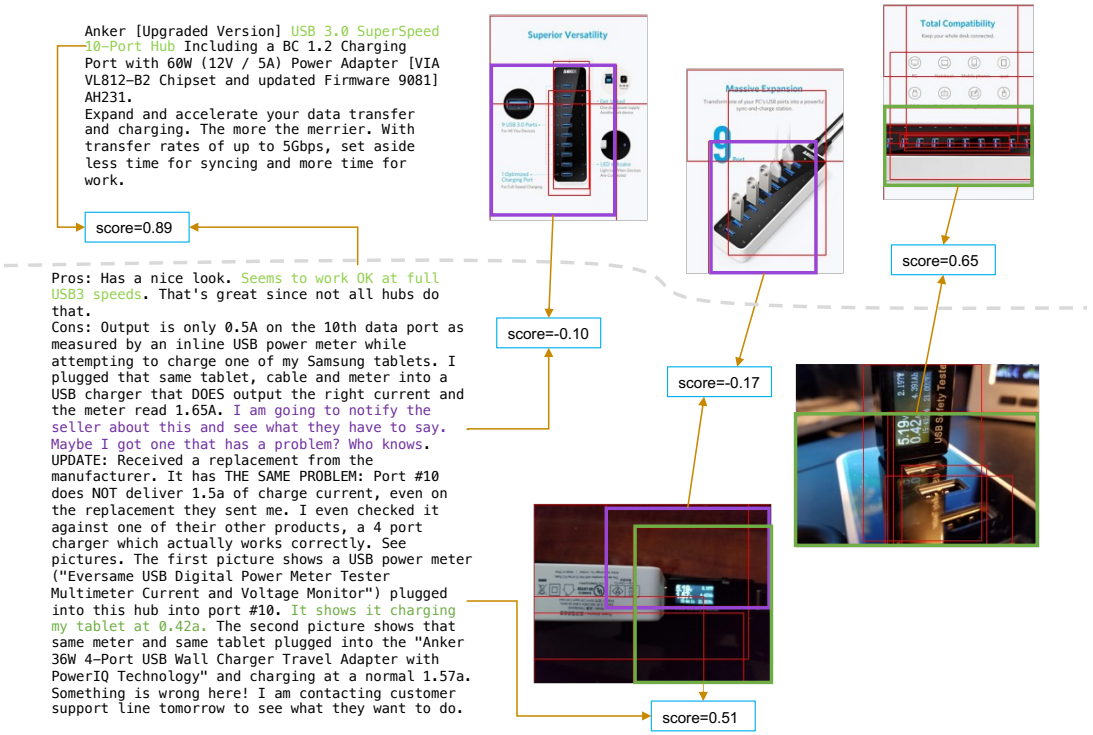


Figure 6: A case study from Amazon-MRHP dataset. The upper and lower part of the figure is the product and review post respectively. Green and purple are instance pairs that produce high and low scores that are selected/not selected into the final feature vector. For the matching of n-gram words, we display the largest matching scores between individual words in that scope and the other elements.

Product ID	Introduction
B005NGQWL2	Expand and accelerate your data transfer and charging. The more the merrier. With transfer rates of up to 5Gbps, set aside less time for syncing and more time for work. With 10 data terminals to choose from, forget about ever having to switch or unplug again. Fast charging. 10th-port dual functionality enables fast charges of up to 1.5 amps with BC 1.2 charging-compliant devices, while simultaneously transferring data. Charge via a power adapter for higher 2 amp speeds with all USB-enabled devices when hub is disconnected from an active USB port, or your computer is off or in sleep mode. Dual functions, duly facilitated. A mainstay for the future. Featuring a high-grade chipset and a powerful 60W adapter, this hub ensures a stable power supply while you work. Get steady operation for years to come. Whether at home or in the office, add another can't-do-without fixture to your desk. BC 1.2 Charging-Compliant Devices: Apple: iPhone 5 / 5s, iPad Air, iPad mini / mini 2 Samsung: Galaxy S3 / S4, Galaxy Note 1 / 2, Galaxy Mega, Galaxy Mini, Exhilarate, Galaxy Tab 2 10.1 Google: Nexus 4 / 5 / 7 / 10 Sony: Xperia TX Nokia: Lumia 920, Lumia 1020 System Requirements Windows (32/64 bit) 10 / 8.1 / 8 / 7 / Vista / XP, Mac OS X 10.6-10.9, Linux 2.6.14 or later. Mac OS X Lion 10.7.4 users should upgrade to Mountain Lion 10.8.2 or later to avoid unstable connections. Compatibility 2.4GHz wireless devices, MIDI devices and some USB 3.0 devices may not be supported. Try using the host port or a USB 2.0 connection. Power Usage For a stable connection, don't use this hub with high power-consumption devices, such as external hard drives. The hub will sync but not charge tablets and other devices which require a higher power input. 

Table 13: (Example 1 of 2) Product introduction of an example from Amazon Electronics dataset. Some special characters have been removed for better readability.

Review ID	Content	GT	GBDT	Ours
B005NGQWL2-14	Pros: Has a nice look. Seems to work OK at full USB3 speeds. That's great since not all hubs do that. Cons: Output is only 0.5A on the 10th data port as measured by an inline USB power meter while attempting to charge one of my Samsung tablets. I plugged that same tablet, cable and meter into a USB charger that DOES output the right current and the meter read 1.65A. I am going to notify the seller about this and see what they have to say. Maybe I got one that has a problem? Who knows. UPDATE: Received a replacement from the manufacturer. It has THE SAME PROBLEM: Port #10 does NOT deliver 1.5A of charge current, even on the replacement they sent me. I even checked it against one of their other products, a 4 port charger which actually works correctly. See pictures. The first picture shows a USB power meter ("Eversame USB Digital Power Meter Tester Multimeter Current and Voltage Monitor") plugged into this hub into port #10. It shows it charging my tablet at 0.42a. The second picture shows that same meter and same tablet plugged into the "Anker 36W 4-Port USB Wall Charger Travel Adapter with PowerIQ Technology" and charging at a normal 1.57a. Something is wrong here! I am contacting customer support line tomorrow to see what they want to do. 	3.00	0.86	0.89
B005NGQWL2-9	I rarely write a negative review, in fact almost never. This AnkerDirect 10-Port USB Data hub lasted only about 2 months. Now none of the USB ports work. For the first couple of weeks all the ports seemed fine. Then one by one they stopped working. Power gets to the unit and the USB ports on my MAC work fine. I've been waiting for a replacement from the company ever since April 20th 2017, after sending their support group my address as requested, but it had not arrived. I wish to amend this review by saying that the Anker customer service folks were very helpful in rectifying this situation. After some checks on the original item at their direction, the Anker folks came to the conclusion that I had a defective product and quickly replaced it with a new model. I've had a couple of days to test it out and it appears to be working just fine. I have always felt that a product or service can go bad but it is the company's response to that problem, should it arise, that gains my respect and future business. 	2.00	0.64	0.52
B005NGQWL2-65	Works very great, powers all of my USB connections, I have a Asus Gaming Laptop which only has 4 USB ports and I needed to have a blue yeti mic, a Logitech Webcam c920, razer keyboard chroma, 2tb hard drive, and a Xbox one controller wireless adapter connected to it. So far nothing has disconnected or malfunctioned. I would definitely recommend this to my friends and familiy. 	1.00	0.75	0.31

Table 14: (Example 1 of 2) Comparison between our model and GBDT on an example from Amazon electronics dataset. The ground truth (GT) scores are annotated ones, while the scores below GBDT and ours are normalized ones.

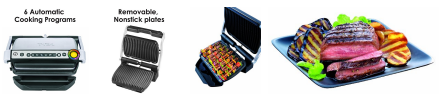
Product ID	Introduction
B00H4O1L9Y	Cook food to perfection with the T-fal OptiGrill GC702D53 electric indoor grill. This indoor grill offers versatility and convenience for any grilled meals. Choose from six pre-set programs: Burger, Poultry, Sandwich, Sausage, Red Meat, and Fish. The grills precision grilling technology with sensors measures the thickness of food for auto cooking based on the program selected. When the flashing light turns solid purple, the grill has properly preheatedplace food on the grill, lower the lid, and it takes care of the rest. A cooking-level indicator light changes from yellow to orange to red signifying the cooking progress with audible beeps that alert when food gets to each stage: rare, medium, and well-done. Take food off the grill once its reached your preferred level of doneness. Along with the six pre-set programs, the electric grill provides two additional cooking options: Frozen mode for defrosting and fully cooking frozen food and Manual mode for cooking vegetables or personal recipes. (Note: when preheating for a pre-set program, keep the lid closed or the grill will automatically switch to Manual mode.) The OptiGrill features a powerful 1800-watt heating element, user-friendly controls ergonomically located on the handle, and die-cast aluminum plates with a nonstick coating for effortless food release. The slightly angled cooking plates allow fat to run away from food and into the drip tray for healthier results, and the drip tray and cooking plates are removable and dishwasher-safe for quick cleanup. Housed in brushed stainless steel, the OptiGrill electric indoor grill makes an attractive addition to any counter. 

Table 15: (Example 2 of 2) Product introduction of an example from Amazon Home dataset.

Review ID	Content	GT	GBDT	Ours
B00H4O1L9Y-111	I want to preface this by saying that I always prefer food grilled on our big outdoor propane grill. It's just a superior method of cooking. That being said, if you don't have an outdoor grill, or even if you do and are sometimes unable to cook with it due to lack of time, running out of propane, inclement weather, laziness... then this is a FABULOUS option to still get the grilled food you loved SUPER FAST and SUPER EASY!! We got 5 (yes FIVE) George Foreman grills for our wedding. I re-gifted 4 of them and kept one and have used it off and on for a long time, but every time I have to clean it afterward I swear I'm never going to use it again because it's such a pain and it never quite gets clean, especially in the area where the hinges are. That problem is no more.... 	4.00	0.58	0.71
B00H4O1L9Y-122	My mom got this on her account. I thought she was crazy to spend so much on what looked like a glorified George Foreman grill but I was wrong, this thing is the bomb. Here is what I like about it;-Heats up super quick. I remember my old George Forman grill took a lot longer.-The presets for the type of food you are cooking must be working because nothing turns out overcooked. The nonstick removable plates. So far, I haven't had any food stick to the plates and I don't use spray or oil. Being able to take them off and wash them in the sink or dishwasher is by far the best part, I used to hate wasting a million paper towels and burning my hands on my old foreman grill and still didn't feel like it was clean. Doesn't create a lot of smoke. When I used to use my old foreman grill I was always setting off the smoke alarm, this grill doesn't do that. 	3.00	0.65	0.62
B00H4O1L9Y-148	I have used this grill 4 times now and everything I've cooked has turned out amazing! I am so impressed with this grill. It is real quick to preheat and has cooked everything perfectly so far from burgers to chicken sausages to kabobs. We haven't tried anything from the cookbook included but we definitely want to. One of the best things is that the plates are detachable and dishwasher safe! Super easy cleanup. 	2.00	0.49	0.42
B00H4O1L9Y-59	One of the best tools for preparing clean food (if that's what you choose). Cooks in minutes and cleans just as fast. Gives that grill experience within a compact structure. Definitely saves time, I use this thing at least ounce a day, on prep days 3-5 times. If you want to loose wait; it starts in YOUR kitchen by preparing your meals. 	1.00	0.35	0.27

Table 16: (Example 2 of 2) Comparison between our model and GBDT on the example from Amazon Home dataset. The ground truth (GT) scores are annotated ones, while the scores below GBDT and ours are normalized ones.