SANCL: Multimodal Review Helpfulness Prediction with Selective Attention and Natural Contrastive Learning

Anonymous ACL submission

Abstract

With the boom of e-commerce, Multimodal Review Helpfulness Prediction (MRHP) that identifies the helpfulness score of multimodal product reviews has become a research hotspot. Previous work on this task focuses on attention-based modality fusion, information integration, and relation modeling, which primarily exposes the following drawbacks: 1) the model may fail to capture the really essential information due to its indiscriminate attention formulation; 2) lack appropriate modeling methods that takes full advantage of correlation among provided data. In this paper, we propose SANCL: **S**elective **A**ttention and **N**atural **C**ontrastive **L**earning for MRHP. SANCL adopts a probe-based strategy to enforce high attention weights on the regions of greater significance. It also constructs a contrastive learning framework based on natural matching properties in the dataset. Experimental results on two benchmark datasets with three categories show that SANCL achieves state-of-the-art baseline performance with lower memory consumption.

1 Introduction

We have witnessed an acceleration towards an e-commerce boom that has transpired over the past decades (Vulkan, 2020). In the virtual bazaar, countless deals are made between mutually invisible sellers and customers from time to time. For customers, it may be their biggest headache to determine whether they should pay for a good when being overwhelmed by tempting advertisements, as they can hardly learn about the true information about a product in face of the seller's meticulous promotion without any references. In this situation, reviews in e-shops that can provide justification information, are thus of great value to customers. However, the quality of reviews under a certain product page can be disparate—many customers are willing to leave informative feedback on the product, while many others arbitrarily write a few words and even paste irrelevant messages in their comments. Therefore, from the perspective of online shopping platforms, they would be welcome and attractive to customers if they provide a service that can intelligently filter and place the most helpful reviews at the top position. The task in the machine learning field to solve this problem is Review Helpfulness Prediction (RHP) (Tang et al., 2013).

As the thriving of multimodal learning research and the handy accessibility of multimodal data in this Internet era, the latest progress incorporated image (vision modality) information into the review helpfulness prediction (RHP) (Liu et al., 2021) as Multimodal RHP (MRHP). Although previous work attained excellent results in MRHP, there are still some drawbacks. First, the attention mechanism in these works for representation learning follows the most basic setting—it directly computes out the attention scores based on the representation vectors of tokens or sentences, without any further intervention on the obtained weights (Fan et al., 2019). Generally, the amount of task-related information in each sentence in a given piece of review may vary greatly—since customers usually casually write these reviews and may insert some meaningless words, such as emotional appreciation or complaint that can not benefit the viewers. We observed that due to dataset characteristics in the MRHP task, there are cues to help locate those key sentences in the review text. Therefore, we proposed a probe-based selective attention mechanism to employ them for better attention results.

Secondly, it has been revealed that the correlation, e.g., the similarity of feature vectors, among multimodal and multi-domain data is an essential factor in modeling (Xu et al., 2020; Chen et al., 2019). Nevertheless, existing studies (Xu et al., 2020; Liu et al., 2021) simply quantified them in similarity metrics, such as cosine value, for direct classification use. Though gained appreciative results, we believe they can be better utilized through the contrastive learning scheme to enhance the quality of learned representations.

In this paper, we propose a novel framework, SANCL, to incorporate these two points. In SANCL we first generate a special "probe" mask that highlights the key sentences from the product and review text. The mask then attends the computation attention modules to help focus more on those task-related sentences. Then we construct a contrastive learning framework to learn better modality representations with internal correlations of data. Based on contrastive predictive coding (CPC) (Oord et al., 2018), the framework is composed of two feature spaces (domains). Each domain takes specific combinations of projected representations as input, according to their relation types from our analysis. Through optimization over the contrastive score, the multimodal and multi-domain representations can learn from the inherent relations. Our contribution can be summarized as follows:

- We design a selective attention approach, including the probe mask generation and mask-based attention computation, for the information aggregation in MRHP tasks.
- We analyze the characteristics and relations in multimodal reviews and formulate a contrastive learning framework to refine the learned representations.
- Extensive experiments on three publicly available datasets show our approach achieves state-of-the-art performance with lower memory consumption.

2 Related Work

In this section, we briefly recap some relevant work in the field of review helpfulness prediction and multimodal contrastive learning.

Review Helpfulness Prediction Customer reviews play an important role in helping customers investigate products before determining whether to purchase. (Zhu and Zhang, 2010; Diaz and Ng, 2018; Gamzu et al., 2021). Support vector regression (SVM) was first employed to automatically judge the review helpfulness (Kim et al., 2006; Zhang and Varadarajan, 2006; Tsur and Rappoport, 2009). Later, linear regression (Lu et al., 2010; Ghose and Ipeirotis, 2010), extended tensor factorization (Moghaddam et al., 2012), and probabilistic matrix factorization models (Tang et al., 2013) have been applied to integrate complicated constraints into the learning process. With the development of deep learning, deep neural networks (Lee and Choeh, 2014; Fan et al., 2018; Chen et al., 2018) have been utilized to model the sophisticate elements in this task. Recently, Qu et al. (2020) proposed a graph neural network to capture the intrinsic relationship between the products and their reviews. However, most existing studies only focus on the text of reviews, neglecting the images that usually exist in online reviews. This paper takes advantage of the images and proposes a novel contrastive learning framework with a selective attention mechanism to learn expressive multimodal features.

Multimodal Representation Learning The foremost problem of multimodal tasks lies in multimodal representation learning (Baltrušaitis et al., 2018). The concept of multimodal representation learning covers many techniques, such as multimodal fusion (Vielzeuf et al., 2018; Wang et al., 2020; Mai et al., 2020; Han et al., 2021a), multimodal contrastive learning (Yuan et al., 2021; Han et al., 2021b), etc. Attention-based architectures are the basic routine in multimodal fusion, but the formulations are similar. In this paper, knowing about the particularity of MRHP and its dataset, we devise a novel attention mechanism to better aggregate information in textual data. Additionally, we also upgrade the application of contrastive learning. Unlike the ordinary treatment that divides samples into positive and negative groups according to "from myself" or "not from myself" (Cui et al., 2020; Liang et al., 2020), we extract contrastive pairs according to the natural correlation in the dataset and construct the framework of two feature spaces termed as domains.

3 Method

In this section, we first introduce the problem definition of Multimodal Review Helpfulness Prediction (MRHP). Then we elaborate on the model architecture and processing pipeline of our method.

3.1 Problem Definition

Given a collection of product descriptions $\mathcal{P} = \{P_1, P_2, ..., P_N\}$ and associated reviews $\mathcal{R} = \{R_1, R_2, ..., R_N\}$ gleaned from an e-shopping

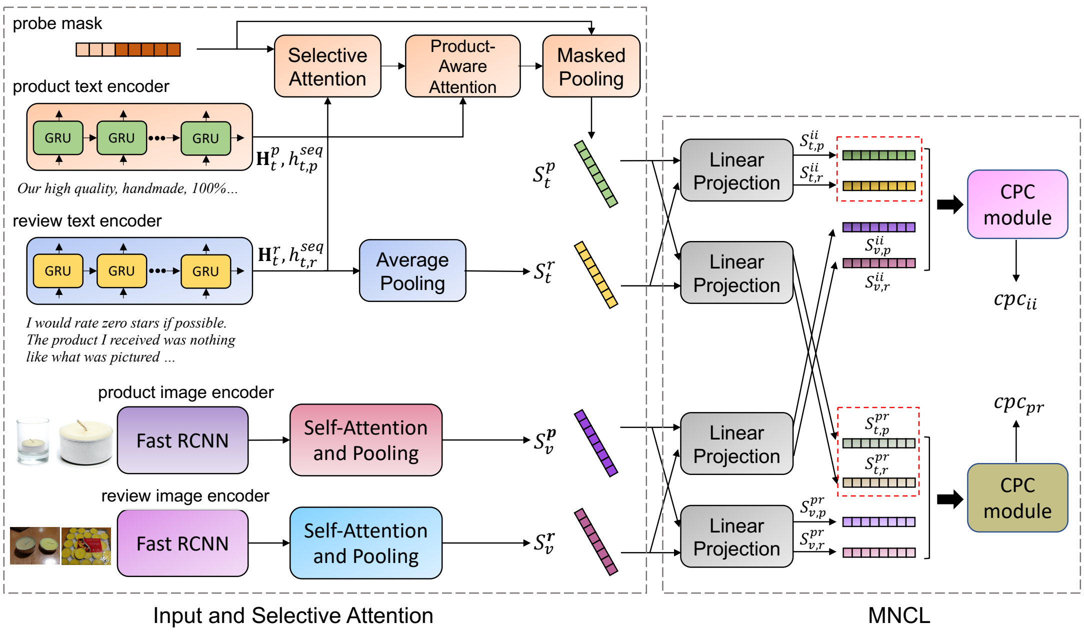


Figure 1: The overview of SANCL. The output layer is omitted. Features in red boxes ($S_{v,r}^{ii}$, $S_{t,r}^{ii}$, $S_{v,r}^{pr}$, $S_{t,r}^{pr}$) are used in final helpfulness score prediction.

website. Each product description $P_i \in \mathcal{P}$ contains the product name n_{p_i} plus the text and image descriptions T_{p_i} and I_{p_i}. The underlying review collection R_i associated with product i contains m review pieces $R_i = \{r_{i1}, r_{i2}, ..., r_{im}\}$. Each review data frame is composed of images $I_{r_{ij}}$ and text $T_{r_{ij}}$ as well. We exhibit an example of input data at the model's input position in Figure 1. All review pieces are annotated with helpfulness scores $s_{ij} \in \{0, 1, 2, 3, 4\}$. Multimodal review helpfulness prediction can be formulated as a regression task that aims to predict the helpfulness score of each review piece, and a ranking task to sort these reviews by their scores in descending order.

3.2 Overview

The overall architecture of SANCL is depicted in Figure 1. We first generate a probe mask for each review according to the corresponding product name and review text as shown in Figure 2. The probe mask highlights the sentences that mention the product, which then participates in the computation of selective attention to produce text representations. For images, we feed the features extracted by pre-trained visual neural networks to two self-attention modules to produce image representations. Then we project these representations of each modality in both product description and customer review into two shared spaces (domains). We finally develop a contrastive learning module to compute the cross-modality and review-product contrastive scores, which further improves the quality of representations output from attention modules.

3.3 Input Encoding

Context-aware Textual Representation For both review and product text, we initialize the token representations with GloVe (Pennington et al., 2014)[1] or pre-trained models as $\mathbf{E}_t = \{e_1^t, e_2^t, ..., e_l^t\} \in \mathbb{R}^{l \times d_e^t}$, where l is the length (number of tokens) of a given sentence and d_e^t is the embedding dimension. We then send these embeddings to a uni-directional Gated Recurrent Unit (GRU) (Cho et al., 2014), yielding token-wise and sequence representations $\mathbf{H}_t = \{h_1^t, ..., h_l^t\}$ and h_t^{seq}:

$$\mathbf{H}_t, h_t^{seq} = \mathbf{GRU}(\mathbf{E}; \theta_t). \tag{1}$$

where θ_t is the parameters in GRU.

Visual Feature Extraction We apply Faster R-CNN (Ren et al., 2015) on raw images and yield the hidden representations $\mathbf{E}_v = \{e_1^v, e_2^v, ..., e_n^v\} \in \mathbb{R}^{n \times d_e^v}$ in the last layer before the classifier to map the Regions of Interest (RoI) in an image to a hidden space, where n is the number of hot regions detected in the image and d_e^v is the vector lengths of hidden representations. Then same as Liu et al. (2021), we feed them into a self-attention module that outputs the encoded image representations $\mathbf{H}_v = \{h_1^v, h_2^v, ..., h_n^v\}$.

[1] We used glove.840B.300d in our experiments.

3.4 Probe-based Selective Attention (PSA)

Having gained elementary encoded multimodal representations, interactions between parallel review pieces and corresponding product descriptions are required to form the product-aware review representations. Previous work primarily formulated these interactions as token-wise description–review attention (Fan et al., 2019; Qu et al., 2020). Though succinct and effective, this token-by-token or sentence-by-sentence computation scheme may neglect distinct the relative importance among sentences and missing really task-related information. Because only through loss back-propagation without any re-weighting operation, it can not always be ensured that larger weights will be put on those key sentences. To mitigate this issue, we propose the selective attention approach. It first generates a special "probe" mask then performs discriminate attention based on that.

Probe Mask Generation The probe mask should reflect the position (i.e., in which sentence) where the product is mentioned in a review. An example of the generation process is displayed in Figure 2. We first retrieve the core words from the product name by looking up its dependency tree and picking the lemmatized form of the words around the root. Next, we use coreference resolution to identify all coreference clusters in the review.

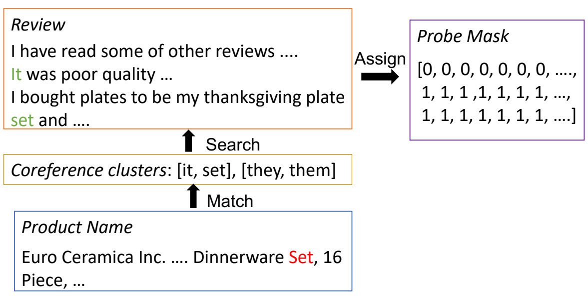


Figure 2: An example of mask generation

There are three possible resolution results: (1) A cluster containing the core word of the product name; (2) At least one cluster exists but the core word is missing in all clusters; (3) No coreference cluster exists. For (1) and (3), we do not require extra steps as the existence of entity clusters can be confirmed. For (2), we are still uncertain whether an entity cluster is in the text. We devise a simple rule to tackle this situation—we regard the first cluster as product name mention cluster, based on our observation that the first repeatedly mentioned pronouns in a review are more likely to refer to the product. After locating these product name mentions, we create the probe mask $M \in \mathbb{R}^{1\times l}$ by assigning 1 to the positions of those mention-found sentences and 0 to others. The process is summarized in Algorithm 1.

Algorithm 1: Probe Mask Generation

Input: Review sentences R, product name P
Output: Probe mask M

 # core words and coreferece clusters extraction:
$\hat{R} \leftarrow$ `Lemmatise(`R`)`, $\hat{P} \leftarrow$ `Lemmatise(`P`)`;
 # core words extraction:
$T \leftarrow$ `DependencyParse(`$\hat{P}$`)`;
$W \leftarrow$ `FindWordsNearRoot(`T`)`;
clusters $\leftarrow$ `FindCoreferenceCluster(`$\hat{R}$`)`
 # mask generation:
$M \leftarrow$ `ZeroInit(`R`.size)`
if $C = \emptyset$ **then**
  return M
end
foreach c **in** clusters **do**
  if any $w \in W$ **in** c **then**
    gold_cluster = c
  end
end
if gold_cluster $= \emptyset$ **then**
  gold_cluster = clusters[0]
end
foreach sent $\in \hat{R}$ **do**
  if any $w \in$ gold_cluster **in** sent **then**
    M[sent.`start` : sent.`end`] $\leftarrow True$
  end
end
return M

Selective Attention with Probe Mask There are three steps to acquire product-aware review representations—self-attention, cross-text attention, and pooling, among which the first and last steps take advantage of probe masks generated. We first transform the probe mask to a new form:

$$M' = \alpha M + \beta(1 - M), \tag{2}$$

where $\alpha > \beta > 0$ since we expect the mask could help focus more on the sentences where the product is mentioned. This effect embodies in the self attention computation of the review text $\mathbf{H}_t^r$, where the fundamental attention weights are computed as:

$$\mathbf{A} = \mathbf{softmax}(\mathbf{W}\mathbf{H}_t^r), \tag{3}$$

We renew the original attention matrix $A \in \mathbb{R}^{l\times l}$:

$$\mathbf{A}' = (M')^T M' \odot \mathbf{A}, \tag{4}$$

In this process, the attention weights are actually re-weighted as

$$a'_{ij} = \begin{cases} \alpha^2 a_{ij}, \text{ if } m_i = m_j = 1 \\ \alpha\beta a_{ij}, \text{ if } m_i = 1, m_j = 0 \\ \beta^2 a_{ij}, \text{ if } m_i = m_j = 0 \end{cases}, \quad (5)$$

An intuitive explanation to this would be that a token receives more information from hot regions (whose mask value is 1) than non-hot regions, and the strength ratio of these two regions is α/β. Naturally we set α to 1.0 while generating β individually for each review from its sequence representation:

$$\beta = \mathbf{sigmoid}(\mathbf{W}_{gen} h^{seq}_{t,r} + b_{gen}), \quad (6)$$

where $\mathbf{W} \in \mathbb{R}^{d_h \times 1}$ is the weight matrix and $b \in \mathbb{R}$ is the bias. The sigmoid function ensures that $\beta \in (0, 1)$. Then we acquire the self-attention results as in common practice:

$$\mathbf{H}^{r'}_t = \mathbf{H}^r_t + \mathbf{A}'\mathbf{H}^r_t, \quad (7)$$

In cross-text attention, since weighted-sum is performed on the product text, we do not utilize probe masks in this stage and obtain $\mathbf{H}^{r''}_t$. Finally, we average the result with the probe mask by a weighted sum to aggregate these sentence representations:

$$S^r_t = \mathbf{weighted\_sum}(\mathbf{H}^{r''}_t, M'). \quad (8)$$

Note that for review image representations there are only cross-image attention and average pooling to yield S^r_v.

3.5 Multi-domain Natural Contrastive Learning (MNCL)

From the theory of mutual information, training to distinguish positive samples from negative ones in terms of their similarity can enrich the learned representations and enhance downstream tasks' performance. In our work, we are concerned about the natural relations and split them into two domains: the inner-instance domain (ii) and product-review (pr) domain. Before forwarding the input representations into the MNCL module, all pooled representations are projected to the domain-specific shared spaces through two linear layers and an activation layer in between. We denote them as $S^d_{m,f}$, where $m \in \{t, v\}$ is the modality type, $f \in \{r, p\}$ is the field (review or product description) and $d \in \{ii, pr\}$ is the domain name:

$$S^d_{m,f} = W^d_{m,f,2}\mathbf{Tanh}(W^d_{m,f,1} S^d_m + b^d_{m,f,1}) + b^d_{m,f,2} \quad (9)$$

where $W^d_{m,*,i}$ and $b^d_{m,*,i}$ are weights and biases in the i-th layer of the projection network. Note that the data in the same modality and domain share the same network parameters. In the succeeding content, we are going to describe details of the two contrastive-learning domains, mainly concerning how to pick positive and negative samples for contrastive learning and training.

Inner Instance (II) Domain In the inner instance domain, we separate positive and negative pairs according to how similar the representations between image and text are in a single training instance. First, from the sellers' perspective, the text and image of a product should match well so as to attract customers. Thus we mark text-image pairs of product descriptions as positive ones (the set of these pairs is denoted as $\mathbb{S}^p_{ii}$). Therefore, we mark the former as positive (the set is denoted as $\mathbb{S}^+_{ii}$) and the latter as negative ones (the collection is denoted as $\mathbb{S}^-_{ii}$). Besides, from our observation, reviews that achieve high helpfulness scores possess a high similarity between its text and the attached image.

Product-Review (PR) Domain The semantic matching property also exists between product descriptions and their associated reviews. As helpfulness is dependent on how well a review is pertinent to the theme of the product, we argue that review pieces of high helpfulness scores ($\mathbb{S}^+_{pr}$) should match the product introduction both visually and literally, while those low-score pieces ($\mathbb{S}^-_{pr}$) match the introduction poorly in both modalities.

Multi-domain Contrastive Predictive Coding (MCPC) In contrastive predictive coding (Oord et al., 2018), we need to compute contrastive scores for every sample pair. According to the common approach (Yuan et al., 2021; Han et al., 2021b), exponential function is chosen as the score function:

$$\varphi(\mathbf{A}, \mathbf{B}) = \mathbf{exp}\left(\frac{\mathbf{norm}(\mathbf{A}^T)\mathbf{norm}(\mathbf{B})}{\tau}\right), \quad (10)$$

where $\mathbf{norm}(*)$ is the l2-norm function, τ is the temperature hyper-parameter, for simplicity we keep its value 1.0 in our experiments. By noise contrastive estimation (Gutmann and Hyvärinen, 2010), in the inner instance domain the score is computed as:

$$cpc_{ii} = - \sum_{(S_{t,j}, S_{v,j}) \in (\mathbb{S}_{ii}^{+} \cup \mathbb{S}_{ii}^{P})} \log \frac{\varphi(S_{t,j}, S_{v,j})}{\sum_{S_k \in (\mathbb{S}_{ii}^{+} \cup \mathbb{S}_{ii}^{-} \cup \mathbb{S}_{ii}^{P})} \varphi(S_{t,k}, S_{v,k})}, \quad (11)$$

where $(S_{t,j}, S_{v,j})$ are the text-image pair from the instance, i.e., a review piece or product description. The summation is over $\mathbb{S}_{ii}^{+}$ and $\mathbb{S}_{ii}^{P}$ because instances counted here are from both product descriptions and review pieces. Similarly in product–review domain the score is:

$$cpc_{pr}^{m} = - \sum_{S_{m,j}^{r} \in \mathbb{S}_{pr}^{+}} \log \frac{\varphi(S_{m,j}^{r}, S_{m,j}^{p})}{\sum_{S_k \in (\mathbb{S}_{pr}^{+} \cup \mathbb{S}_{pr}^{-})} \varphi(S_{m,k}^{r}, S_{m,k}^{p})}. \quad (12)$$

$$cpc_{pr} = cpc_{pr}^{t} + cpc_{pr}^{v} \quad (13)$$

where $S_{m,j}^{r}$ is the representation of modality m in review r from the positive review set $\mathbb{S}_{pr}^{+}$ and $S_{m,j}^{p}$ is the counterpart of the corresponding product.

3.6 Prediction and Training

We select all review-related representations in the common spaces of two domains $(S_{t,r}^{ii}, S_{t,r}^{pr}, S_{v,r}^{ii}, S_{v,r}^{pr})$ and concatenate them as features for final prediction ($\mathbf{F}$). After concatenating these features, a linear layer takes them as input and outputs the helpfulness score predictions ξ_r:

$$\mathbf{F} = \mathbf{concat}([S_{t,r}^{ii}, S_{t,r}^{pr}, S_{v,r}^{ii}, S_{v,r}^{pr}]) \quad (14)$$

$$\xi_r = \mathbf{W}_o \mathbf{F} + b_o, \quad (15)$$

where $\mathbf{W_o}$ and b_o are the weight matrix and bias in the output layer. Same as Liu et al. (2021), we adopt the standard pairwise ranking loss as the task loss:

$$\mathcal{L}_{task} = \sum_{i} \mathbf{max}(0, \gamma - \xi_{r^{+},i} + \xi_{r^{-},i}), \quad (16)$$

where r^{+}, r^{-} are an arbitrary pair of review pieces under product P_i, γ is a scaling factor. Contrastive losses make up the auxiliary loss:

$$\mathcal{L}_{aux} = cpc_{ii} + cpc_{pr} \quad (17)$$

Hence the total loss for training is (κ is a hyper-parameter to adjust the effect of auxiliary loss):

$$\mathcal{L} = \mathcal{L}_{task} + \kappa \mathcal{L}_{aux} \quad (18)$$

4 Experimental Settings

This section presents the datasets used in the experiments, baseline models, and metrics.

4.1 Datasets

We conduct experiments on three MRHP datasets (Liu et al., 2021) in different categories: *Clothing, Shoes & Jewelry, Home & Kitchen* and *Electronics*. The text and images in these datasets are crawled from Amazon online shops through 2018 until 2019. The helpfulness scores are annotated according to the number of helpfulness votes to the reviews. More specifically the scores equal to $\lfloor \log_2 n_{\text{votes}} \rfloor$ and are clipped into $[0, 4]$. Details of datasets are provided in Appendix.

4.2 Baseline Models

Following previous work, we first compare our model with a bunch of baselines in the text-only setting, which examines and signifies the enhancement by our selective attention mechanism and text-related contrastive learning modules. The baseline candidates contain Multi-Perspective Matching (BiMPM) network (Wang et al., 2017), Embedding-gated CNN (EG-CNN) (Chen et al., 2018), Convolutional Kernel-based Neural Ranking Model (Conv-KNRM) (Dai et al., 2018) and Product-aware Helpfulness Prediction Network (PRHNet) (Fan et al., 2019). In multimodal settings, we pick a collection of state-of-the-art multimodal helpfulness prediction models for comparison:

- **SSE-Cross** (Abavisani et al., 2020): The Stochastic Shared Embeddings (SSE) Cross-modal Attention Network introduces a novel cross-attention mechanism that can filter noise components from weak modalities which may mislead the model to make wrong predictions on a sample. SSE is adopted as the regularization technique to alleviate over-fitting in the fusion process to further prompt the prediction accuracy.

- **D&R Net** (Xu et al., 2020): The Decomposition and Relation Network learns the commonality and discrepancy between image and text in decomposition network and the multi-view semantic association information in relation network.

- **MCR** (Liu et al., 2021): The Multi-perspective Coherent reasoning method in-

Setting	Model	Cloth. & Jew			Electronics			Home & Kitchen		
		MAP	**N-3**	**N-5**	**MAP**	**N-3**	**N-5**	**MAP**	**N-3**	**N-5**
Text-only	BiMPM* (Wang et al., 2017)	57.7	41.8	46.0	52.3	40.5	44.1	56.6	43.6	47.6
	EG-CNN* (Chen et al., 2018)	56.4	40.6	44.7	51.5	39.4	42.1	55.3	42.4	46.7
	Conv-KNRM* (Dai et al., 2018)	57.2	41.2	45.6	52.6	40.5	44.2	57.4	44.5	48.4
	PRHNet† (Fan et al., 2019)	58.23	43.36	47.21	52.31	40.43	43.88	57.11	44.46	48.27
	SANCL (Ours)	**58.98**$^\natural$	**44.75**$^\natural$	**48.57**$^\natural$	**53.03**$^\natural$	**41.03**$^\natural$	**44.77**$^\natural$	**58.03**$^\natural$	**45.59**$^\natural$	**49.31**$^\natural$
	BERT (Devlin et al., 2018)	56.47	42.98	46.84	51.95	39.77	43.11	56.62	42.12	46.87
	PRHNet+BERT† (Fan et al., 2019)	57.51	43.65	47.74	52.28	40.66	44.02	57.32	44.74	48.42
	SANCL+BERT (Ours)	**58.49**$^\natural$	**44.91**$^\natural$	**48.69**$^\natural$	**53.13**$^\natural$	**41.77**$^\natural$	**45.01**$^\natural$	**58.20**$^\natural$	**45.83**$^\natural$	**49.65**$^\natural$
Multimodal	SSE-Cross* (Abavisani et al., 2020)	65.0	56.0	59.1	53.7	43.8	47.2	60.8	51.0	54.0
	D&R Net* (Xu et al., 2020)	65.2	56.1	59.2	53.9	44.2	47.5	61.2	51.8	54.6
	MCR† (Liu et al., 2021)	66.96	58.03	61.06	55.86	46.32	49.45	63.17	53.85	57.14
	SANCL (Ours)	**67.26**	**58.61**$^\natural$	**61.48**$^\natural$	**56.19**	**46.98**$^\natural$	**49.92**$^\natural$	**63.35**	**54.28**$^\natural$	**57.40**
	MCR+BERT (Liu et al., 2021)	65.81	55.94	58.75	55.15	45.67	48.62	62.39	52.91	56.09
	SANCL+BERT (Ours)	**66.52**$^\natural$	**56.73**$^\natural$	**59.90**$^\natural$	**56.04**$^\natural$	**46.77**$^\natural$	**49.95**$^\natural$	**62.74**	**53.65**$^\natural$	**56.91**$^\natural$

Table 1: Results on three datasets; all reported metrics are the average of five runs; "$*$" are from Liu et al. (2021) and "$\dagger$" are from the open-source code in Liu et al. (2021); "$\natural$" represent the results significantly outperforms PRHNet and MCR with p-value < 0.05 based on paired t-test.

corporates the joint reasoning across textual and visual modalities from both the product and the review. Three types of coherence are modeled to learn effective modality representations for the helpfulness prediction.

In both settings, we also test our method with BERT (Devlin et al., 2018) encoder and compare that to the respective SOTA models on BERT. In addition, we test and record basic BERT performance (BERT+a double linear layers).

4.3 Metrics

As MRHP is a ranking task, the metrics for comparison are as well ranking-customized. After sorting all prediction-truth scores in descending order, the Mean Average Precision (MAP) computes the mean precision of top-1 to top-K samples. K is usually large enough to ensure top-K can encompass the entire collection of reviews under every product. The Normalized Discounted Cumulative Gain (NDCG-N) (Järvelin and Kekäläinen, 2017; Diaz and Ng, 2018) purely reckons the gain value over top-N predictions (N is 3 and 5 in our experiments), which simulates the real circumstances of a typical customer who would always read the topmost reviews.

5 Results and Analysis

In this section, we will compare our approach with several advanced baselines and explore how it improves in the multimodal helpfulness prediction task.

5.1 Performance Comparison

We list the performance of our model and baselines in Table 1. Notably, SANCL consistently outperforms all the baselines in both text-only and multimodal, BERT and Glove initialization settings. These outcomes initially demonstrate the efficacy of our method in MRHP tasks. It is surprising that we cannot gain significant performance boost by replacing Glove with BERT as the text encoder. We speculate the reason is that Glove embeddings are expressive enough for this task.

Moreover, it can be claimed that SANCL is a lightweight model compared to the multimodal SOTA, since the model size and GPU memory consumption of SANCL are much lower than MCR. The total number of parameters is 2.63M in MCR and 1.41M (exclude the embedding layer) in SANCL respectively, which indicates a double efficiency. The average GPU memory usage of SANCL during the training on Amazon-MRHP Home & Kitchen is around 2.4G, while MCR occupies an average of 13.7G GPU memory during training, which is 4.7 times higher than SANCL.

5.2 Ablation Study

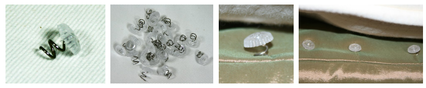
To verify the benefits of our proposed method, we carry out comprehensive ablation experiments on the Amazon electronics dataset, including the selective attention and contrastive learning components. In selective attention, we first replace learned β with a fixed value of 0.5, since we find most β values in our experiments are around 0.5. Next, we remove the entire selective attention module and only preserve the primitive attention computation. The

Description	MAP	N-3	N-5
SANCL	**56.19**	**46.98**	**49.92**
Attention			
w/o learned β (fixed at 0.5)	55.61	46.37	49.58
w/o probe mask	55.43	46.11	49.45
Contrastive learning			
w/o cpc_{ii}	55.54	46.29	49.23
w/o cpc_{pr}	55.81	46.40	49.47
w/o cpc_{ii} and cpc_{pr}	55.35	46.28	49.09

Table 2: Ablation study of SANCL on the Electronics dataset.

Product Name: Twisty Pins for Upholstery, Slipcovers and Bedskirts 50/pkg

Product description: Package of 50 Clear Twisty Pins for securing fabrics and accent trims. Nickel plated steel pin 1/2" in diameter clear top, wire twist 3/8" long. Perfect for Medium to light weight fabrics, bed skirts, bed ruffles, slipcovers and upholstery.



Review (Helpfulness Score: 4): I bought *these* to pin the loose material on a sofa cover and they worked like a charm. The sofa cover definitely looks form fitting now.

Predictions: SANCL: 4.5291 MCR: -1.0832

CPC score: $cpc_{ii} = 0.82, cpc^{t}_{pr} = 0.76, cpc^{v}_{pr} = 0.21$

Table 3: Examples from the Amazon Home & Kitchen test set.

decline in the outcome of both situations manifests that the probe-based selective attention amends the cross-text information exchange between text fields. For multi-domain contrastive learning, we delete the CPC losses of a single or both domains in training. The results indicate that both domains have a positive impact on performance. Moreover, the effect of the two domains does not counteract their collaboration, as we observe accumulated benefits when they operate together.

5.3 Case Study

To understand how our model deals with samples in-depth, we randomly picks up a test product-review instance from the test set of Amazon Home & Kitchen to explain how SANCL works and, as shown in Table 3.

In this example, the customer bought the pins to fix the edge of his sofa. Instead of photoing pins themselves, the customer only presented the tidy sofa after installing the pins. We first visualize the attention weights in test time, as shown in Fig. 3. Note that only the first sentence in the review contains the elements in the coreference clusters, which we have emphasized with italics and underline in Table 3. Consistently, we observe the significant larger weights in the region of first sentence (row/column 1-19) while the rest region's weights are much smaller. We also ran MCR and collect its prediction on this example, and it is clear that MCR commits a severe error here, probably caused by the direct classification on the unimodal cosine similarity. In our approach, as we carefully analyze and classify the positive and negative pairs in the multi-domain contrastive learning framework, the huge semantic similarity between review text and image and between product description and review text, indicated by the high CPC scores $S^{ii}_{v,r}$, assists the model to correctly predict the score.

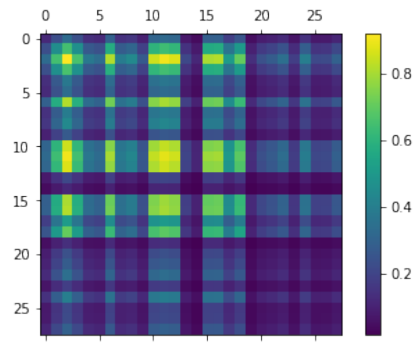


Figure 3: The self-attention weights of the review text in the given example at test time (β=0.57)

6 Conclusion

We propose a novel framework, SANCL, for the task of multimodal review helpfulness prediction (MRHP) in this paper. We first present a selective attention mechanism, which purposefully aggregates information from these crucial sentences in the review text by generating the probe mask that exerts re-normalization on the attention weights and pooling stage. We then build up a multi-domain natural contrastive learning framework in our model. It exploits the natural relations among the data from different fields and modalities in the dataset to enhance the model's capacity of multi-modal representation learning. Results of comprehensive experiments and analyses demonstrate the superiority of our model over the comparable baselines and the efficacy of the novel components.

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.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

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.

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*.

Wanyun Cui, Guangyu Zheng, and Wei Wang. 2020. Unsupervised natural language inference via decoupled multimodal contrastive learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5511–5520.

Zhuyun Dai, Chenyan Xiong, Jamie Callan, and Zhiyuan Liu. 2018. Convolutional neural networks for soft-matching n-grams in ad-hoc search. In *Proceedings of the eleventh ACM international conference on web search and data mining*, pages 126–134.

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.

Miao Fan, Chao Feng, Lin Guo, Mingming Sun, and Ping Li. 2019. Product-aware helpfulness prediction of online reviews. In *The World Wide Web Conference*, pages 2715–2721.

Miao Fan, Yue Feng, Mingming Sun, Ping Li, Haifeng Wang, and Jianmin Wang. 2018. Multi-task neural learning architecture for end-to-end identification of helpful reviews. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 343–350. IEEE.

Iftah Gamzu, Hila Gonen, Gilad Kutiel, Ran Levy, and Eugene Agichtein. 2021. Identifying helpful sentences in product reviews. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 678–691, Online. Association for Computational Linguistics.

Anindya Ghose and Panagiotis G Ipeirotis. 2010. Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE transactions on knowledge and data engineering*, 23(10):1498–1512.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings.

Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021a. 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, and Soujanya Poria. 2021b. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192.

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.

Soo-Min Kim, Patrick Pantel, Timothy Chklovski, and Marco Pennacchiotti. 2006. Automatically assessing review helpfulness. In *Proceedings of the 2006 Conference on empirical methods in natural language processing*, pages 423–430.

Sangjae Lee and Joon Yeon Choeh. 2014. Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, 41(6):3041–3046.

Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. 2020. Learning to contrast the counterfactual samples for robust visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3285–3292.

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.

Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. 2010. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World wide web*, pages 691–700.

Sijie Mai, Haifeng Hu, and Songlong Xing. 2020. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 164–172.

Samaneh Moghaddam, Mohsen Jamali, and Martin Ester. 2012. Etf: extended tensor factorization model for personalizing prediction of review helpfulness. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 163–172.

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.

Xiaoru Qu, Zhao Li, Jialin Wang, Zhipeng Zhang, Pengcheng Zou, Junxiao Jiang, Jiaming Huang, Rong Xiao, Ji Zhang, and Jun Gao. 2020. Category-aware graph neural networks for improving e-commerce review helpfulness prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2693–2700.

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.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China National Conference on Chinese Computational Linguistics*, pages 194–206. Springer.

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.

Oren Tsur and Ari Rappoport. 2009. Revrank: A fully unsupervised algorithm for selecting the most helpful book reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 3.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0.

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

Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33.

Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. In *IJCAI*.

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.

Xin Yuan, Zhe Lin, Jason Kuen, Jianming Zhang, Yilin Wang, Michael Maire, Ajinkya Kale, and Baldo Faieta. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6995–7004.

Zhu Zhang and Balaji Varadarajan. 2006. Utility scoring of product reviews. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 51–57.

Feng Zhu and Xiaoquan Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of marketing*, 74(2):133–148.

A Dataset Specification

Specifications of the two datasets are in Table 4 and 5 below.

Amazon-MRHP (Products/Reviews)			
Category	Cloth. & Jew.	Elec.	Home & Kitch.
Train	12074/277308	10564/240505	14570/369518
Dev	3019/122148	2641/84402	3616/92707
Test	3966/87492	3327/79750	4529/111593

Table 4: Statistics of the Amazon-MRHP dataset.

Lazada-MRHP (Products/Reviews)			
Category	Cloth. & Jew.	Elec.	Home & Kitch.
Train	6596/104093	3848/41828	2939/36991
Dev	1649/26139	963/10565	736/9611
Test	2062/32274	1204/12661	920/12551

Table 5: Statistics of the Lazada-MRHP dataset.

B Hyperparameter Search

The optimal hyperparameter settings are provided in Table 6 and 7.

Glove Hyperparameters			
	Cloth. & Jew.	Elec.	Home & Kitch.
learning rate	$1e^{-4}$	$5e^{-5}$	$1e^{-4}$
text embedding dim	300	300	300
text embedding dropout	0.5	0.5	0.2
image embedding dim	128	128	128
LSTM hidden dim	128	128	128
shared space hidden	64	64	64
κ	0.25	0.1	0.1
batch size	32	32	32

Table 6: Hyperparameters for all categories using glove-300d embeddings.

Amazon-MRHP Hyperparameters			
	Cloth. & Jew.	Elec.	Home & Kitch.
learning rate	$2e^{-5}$	$2e^{-5}$	$2e^{-5}$
text embedding dim	768	768	768
text embedding dropout	0.5	0.5	0.5
LSTM hidden dim	128	128	128
image embedding dim	128	128	128
shared space hidden	64	64	64
κ	0.3	0.25	0.25
batch size	32	32	32

Table 7: Hyperparameters for all categories using BERT as encoder

We use the same set of settings for text-only and multimodal modes for the same category dataset. The search space of these hyperparameters are: learning rate in $\{1e^{-4}, 2e^{-5}\}$, text embedding dropout in $\{0.2, 0.5\}$, κ in $\{0.1, 0.25, 0.3, 0.5\}$, shared space hidden dimension in $\{64, 128\}$. We train and test each dataset on a single Tesla V100 GPU. In BERT experiments, we use shared a BERT encoder for both product description and review text. To balance the computation cost and model performance, following Sun et al. (2019), we fine-tune the last four layers of the BERT encoder.

C Language Tools

For coreference resolution, we use neuralcoref, an extension that can be placed on SpaCy processors. For BERT model, we use the huggingface transformers package to load.