# Optimal Target-Oriented Knowledge Transportation for Aspect-Based Multimodal Sentiment Analysis

Linhao Zhang[1,2,3], Li Jin[1]*, Guangluan Xu[1], Xiaoyu Li[1], Xian Sun[1], Zequn Zhang[1], Yanan Zhang[4], Qi Li [5]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing (China)
[2] University of Chinese Academy of Sciences, Beijing (China)
[3] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing (China)
[4] Sichuan University, Chengdu (China)
[5] Faculty of Psychology, Beijing Normal University, Beijing (China)

* Corresponding author: jinlimails@gmail.com

Received 9 February 2023 | Accepted 6 November 2023 | Early Access 13 February 2024

unir
LA UNIVERSIDAD EN INTERNET

## Abstract

Aspect-based multimodal sentiment analysis under social media scenario aims to identify the sentiment polarities of each aspect term, which are mentioned in a piece of multimodal user-generated content. Previous approaches for this interdisciplinary multimodal task mainly rely on coarse-grained fusion mechanisms from the data-level or decision-level, which have the following three shortcomings:(1) ignoring the category knowledge of the sentiment target (mentioned in the text) in visual information. (2) unable to assess the importance of maintaining target interaction during the unimodal encoding process, which results in indiscriminative representations considering various aspect terms. (3) suffering from the semantic gap between multiple modalities. To tackle the above challenging issues, we propose an optimal target-oriented knowledge transportation network (OtarNet) for this task. Firstly, the visual category knowledge is explicitly transported through input space translation and reformulation. Secondly, with the reformulated knowledge containing the target and category information, the target sensitivity is well maintained in the unimodal representations through a multistage target-oriented interaction mechanism. Finally, to eliminate the distributional modality gap by integrating complementary knowledge, the target-sensitive features of multiple modalities are implicitly transported based on the optimal transport interaction module. Our model achieves state-of-the-art performance on three benchmark datasets: Twitter-15, Twitter-17 and Yelp, together with the extensive ablation study demonstrating the superiority and effectiveness of OtarNet.

## Keywords

Aspect-Based Multimodal Sentiment Analysis, Optimal Transport, Social Media Opinion Mining.

DOI: 10.9781/ijimai.2024.02.005

## I. Introduction

Social media websites provide interactive platforms to facilitate the creation and sharing of individuals' expressions through multiple social activities (for example, 'like', 'reply', 'retweet', '@', 'share' in Twitter) [1]. Fine-grained sentiment analysis over these user generated content (UGC) in social websites (e.g., Twitter, Flickr) are effective in understanding public opinions toward social hotspots or figures, and it has drawn increasing recent attention in both academia and industry [2]. For example, socialists and psychologists have strong interests in understanding individual reactions toward specific social issues. Companies are willing to acquire online evaluations of their products as feedback to make further improvements. Therefore, how to incorporate heterogeneous multimodal information to conduct fine-grained sentiment analysis over the mentioned aspect terms has become an emerging interdisciplinary research problem, proposed as Aspect-Based Multimodal Sentiment Analysis (ABMSA) [3]-[5].

Despite the well-established research fields of multimodal learning and affective computing, there are under-researched challenges for the aspect-based multimodal sentiment analysis (ABMSA) toward social media user-generated content (UGC): (1) Due to the viral nature of internet posts, sentences in social media UGC are always shorter, more informative and informal compared to the well-organized reviews used for traditional affective computing. (2) The visual information is much noisier with multiple objects for UGC than for videos of human speakers commonly leveraged in multimodal sentiment analysis. (3) Except for the modality gap that commonly presents in multimodal learning, there are additional semantic gaps for social media UGC, considering the fact that linguistic information in UGC focuses more on opinions reflecting sentiment polarities, while visuals imply more on the sentiment targets. These peculiarities limited the performance of methods developed for traditional opinion mining tasks [6], [7].

Please cite this article in press as: L. Zhang, L. Jin, G. Xu, X. Li, X. Sun, Z. Zhang, Y. Zhang, Q. Li. Optimal Target-Oriented Knowledge Transportation for Aspect-Based Multimodal Sentiment Analysis, International Journal of Interactive Multimedia and Artificial Intelligence, vol. 9, no. 4, pp. 59-69, 2025, http://dx.doi.org/10.9781/ijimai.2024.02.005

However, most existing research makes little targeted effort to social media aspect-based multimodal sentiment analysis. Traditional multimodal methods with early fusion in data-level [8], [9], or late fusion in decision-level [10],[11], suffer the problem of extra input redundancy and distributional modality gap, which result in the suboptimal performance for multimodal interaction. Although [1],[12], [13] have made attempts to eliminate the modality gap through modified cross-modal attention mechanisms, they neglect the semantic gap in social media user-generated content described above, especially the underlying target category knowledge in visual components. These semantic gaps may finally result in the increasing risk of misalignment in inter-modal interactions. Besides, previous works also neglect the importance of maintaining target sensitivity, which is particularly essential in acquiring discriminative multimodal representations to perform fine-grained analysis considering various aspect terms.

In this paper, we propose OtarNet, a multi-stage knowledge transportation framework based on optimal transport (OT) for ABMSA, which is effective in maintaining target sensitivity to avoid triviality and misalignment, caused by insufficient aspect interaction and semantic gaps respectively. Firstly, we explicitly transport the visual category knowledge through input space translation and reformulation, through which we acquire a synthetic sequence to supply context information. Secondly, the synthetic context sequence is incorporated into the unimodal encoding process, and ensures the good maintenance of target sensitivity through the proposed intra-modality target interaction mechanism, which outputs target-sensitive unimodal representations rich in semantic knowledge. Thirdly, the multiple unimodal representations are fed into the optimal transport interaction module, in which the inter-modality complementary knowledge (i.e. opinion knowledge in text and target knowledge in image) is implicitly transported to the other modality. Our contributions are summarized as follows:

- We propose OtarNet, a multi-stage knowledge transportation framework for aspect-based multimodal sentiment analysis. OtarNet explicitly transports the visual context knowledge before feature fusion to maintain target sensitivity, which is neglected by most multimodal approaches developed for traditional sentiment analysis. The proposed intra-modality target interaction mechanism is effective in avoiding triviality and misalignment.
- We leverage the optimal transport interaction to implicitly transport inter-modality complementary knowledge. OT interaction is effective in eliminating the distributional modality and semantic gap, which puts an extra burden on previous data-level or decision-level fusion techniques.
- We conduct extensive quantitative and qualitative experiments on three benchmark datasets: Twitter-15, Twitter-17 and Yelp. The newly-achieved state-of-the-art performance, together with the extensive ablation studies and visualizations demonstrate the superiority and effectiveness of OtarNet.

## II. Related Work

Despite the well-established field of sentiment analysis, our OtarNet focuses on aspect-based (aspect term) multimodal sentiment analysis, which is a novel challenge proposed firstly in 2019 by [3] and drawing increasing attention. This relatively new task stemmed from two lines of research, namely fine-grained sentiment analysis and multimodal sentiment analysis.

### A. Fine-Grained Sentiment Analysis

Fine-grained sentiment analysis aims to identify the sentiment polarity of a textual sentence on a given aspect or target [14]. Its research methods can be divided into three main groups: traditional feature selection based methods, neural network based methods and adaption of transformer-style models.

Early lexicon-based methods [15], [16] were established on handcrafted features such as lexical, syntactic and semantic features. These studies always demanded for a professional prior knowledge in linguistics [17], [18] and sometimes failed to capture the dependency between the given target and associated context. Later, neural networks with higher capability of encoding original features as continuous vectors were applied. [19]–[21] modified Long Short-Term Memory (LSTM) recurrent networks with stronger expressive power by attention mechanism to incorporate key information in sentence to a target aspect. [22] chose to use Gated Recurrent Unit (GRU) modules to utilize content information, which was able to deal with the syntactically structures of complex sentence. Moreover, sophisticated neural models with subtle intermediate attention were developed. [23] designed a Memory Network with multi-hop attention and external memory, which can explicitly capture the importance of each context word. [20], [24] leveraged multi-layer and multi-grained attention correspondingly to exploit semantic dependencies between opinion words in multi-level modeling for aspects. Recently, since the pre-trained language model [25] has made success in many tasks, [6] utilized BERT with an additional corpus and realized performance improvement in both aspect extraction and sentiment label classification. [26] achieved accurate prediction for this task which has been translated to a sentence-pair classification task by constructing auxiliary sentences. However, these studies fail to consider visual features that may boost these text-based approaches, which are one key factor of this paper.

### B. Multimodal Sentiment Analysis

Multimodal sentiment analysis is an emerging research, the goal of which is to regress or classify the overall sentiment of an utterance integrating textual and non-textual information. Relative methods can also be divided into three groups: feature engineering methods, neural network based methods and modification of large pretrained models.

Early work mainly focused on feature engineering, which [27] combined adjective-noun pairs with linguistic features to calculate sentiment scores, and [28] proposed to fuse text and image features to obtain similarity of two instances for a new neighborhood classifier. Then motivated by the fusion approaches in feature and score-level [29], [30], pre-trained text and image CNNs [31] were conducted to extract feature and combine these multimodal features to train a logistics regression model. [32] modified LSTM to capture interactions between modalities through time. After that, models modified with attention mechanism [1], [12], [13] were proposed, [33] introduced a novel Attention-Based Modality-Gated Networks (AMGN) to learn the fine-grained correlation and the discriminative features between different modalities. In 2019, [3] introduced a Multi-Interactive Memory Network (MIMN) to supervise the textual and visual information under the given aspect. MIMN learned not only the interactive influences between cross-modality data but also the self influences in single-modality data. And more recently, [4] modified BERT architecture based on target-sensitive cross attention to capture the interaction between modalities. [34] proposed EF-CapTrBERT model to solve this task through input space translation, exploiting generated caption to substitute original images.

However, the existing approaches mainly focus on eliminating the modality gap generated by unimodal encoding procedures. Our OtarNet focuses more on target sensitivity and semantic gaps, which are solved based on multi-stage knowledge transportation and Optimal Transport Interaction.

### C. Optimal Transport

Recently, Optimal Transport (OT) has attracted increasing attention in multiple fields [35]. As one of the research hotspots from optimization theory, OT has excellent performance on sequence alignment and domain adaption problems. By finding the best transportation plan between two data distributions with minimum cost, OT explicitly formulates signals to provide additional guidance [36]. Thus OT has achieved promising results compared to attention-based approaches guided by task-specific loss only [37]. For OT applications related to knowledge transportation, [38] explicitly distilled the knowledge of the monolingual summarization teacher into the student through an OT-based distance, which is effective in estimating the discrepancy and constructing cross-lingual correlation. And VOLT [39] formulated the quest of vocabularization as an optimal transport (OT) problem by finding the optimal transport matrix from the character distribution to the vocabulary token distribution. [40] used the transport plan as an ad-hoc attention score in the context of network embedding to align data modalities. MuLOT [41] utilized OT-based domain adaptation to learn strong cross-modal dependencies for sarcasm and humor detection. [42] innovatively revisited the label assignment from a global perspective and proposed to formulate the assigning procedure as an optimal transport (OT) problem. However, none of these studies have exploited optimal transport to implicitly incorporate complementary knowledge in ABMSA.

## III. Problem Definition

Given a set of multimodal samples (e.g., tweets from Twitter) $\mathcal{D}$. Each piece of user-generated content $C \in \mathcal{D}$ consists of text information $T$ with $n$ words $[w_1, ..., w_n]$ (e.g., [*Taylor Swift drawn with colored pencils!* [*emoji*] ]) and an associated image $I$ (e.g., first picture in Fig. 1). The sentiment target Ttar , as a sub-sequence of words in T is also given (e.g., [*Taylor Swift*]), which is assigned a sentiment label $y_{tar}$, belonging to a given label set, such as {*positive, negative, neutral*} for Twitter and rating scores {1, 2, 3, 4, 5} for YELP. Our problem definition can be stated as follows: given $\mathcal{D}$ as training corpus, the task goal is to learn a target-oriented sentiment classifier, so that it can correctly predict sentiment labels $y_{tar}$ for sentiment targets $T_{tar}$ when encountering unseen samples. Note that there may be one or more targets mentioned in one sentence $T$, and the model needs to predict a single sentiment label for each associated target.

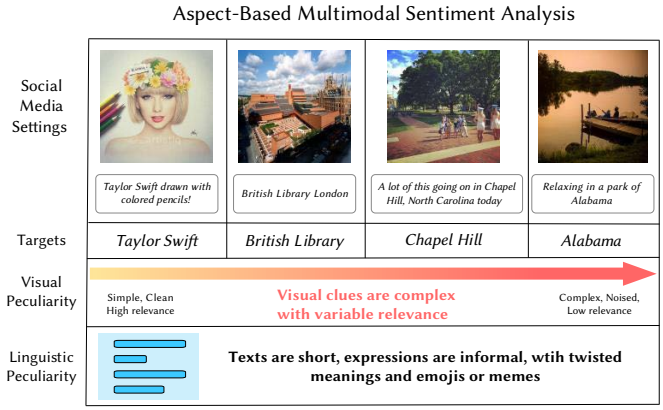



Fig. 1. An example of aspect-based multimodal sentiment analysis (ABMSA) from Twitter. One or more target aspect terms will be mentioned in the text for one piece of user-generated content (UGC). The difficulties of analyzing social media UGC lie in their convenient non-standard writing and network vocabulary.

## IV. Proposed Methodology

In this section, we formulate our task firstly and then decompose OtarNet, as Fig. 2 shows, into three main components: (1) Intra-modality target knowledge transportation, which combines the semantic fusion and implicit fusion through incorporating constructed bridge sentence and bridge feature to calculate the target-sensitive feature. (2) Inter-modality complementary knowledge translation, which enhances the interactivity across image and text modalities. (3) Multimodal feature fusion, which conducts the final stage of fusion to capture the interior relationship between modalities. (4) Optimization process with a classifier is finally leveraged to minimize the standard

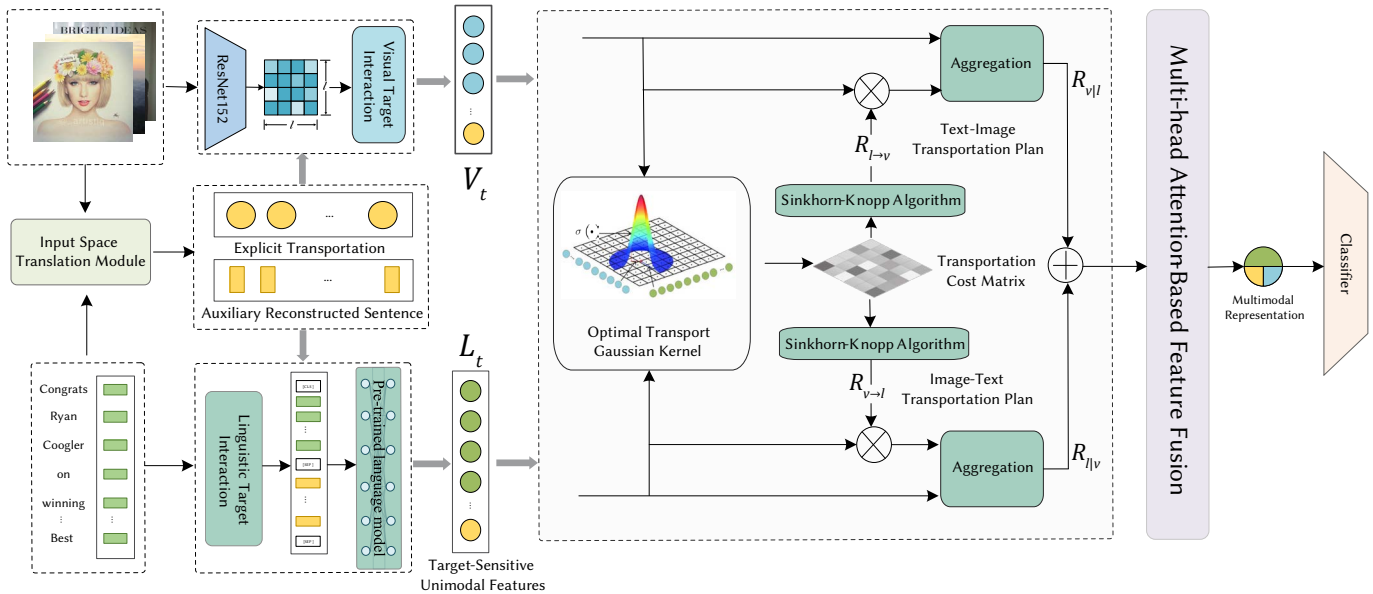



Fig. 2. The workflow of the proposed model OtarNet. Firstly, the auxiliary sentence with its representations are reconstructed through the Input Space Translation Module. Secondly, the auxiliary sequences containing target information are explicitly transported into the two unimnodal encoding processes, which generate two target-sensitive features. Thirdly, the target-sensitive features are intergrated through the OIptimal Transport-Based Interaction Module, which are designed to capture the complementary knowledge in multiple modalities. Finally, the multi-head attention-based fusion layer is leveraged to integrate the multimodal features, which are then used for classification.

cross-entropy loss function as the objective. The aforementioned modules are introduced in the following subsections, respectively.

### A. Intra-Modality Target Knowledge Transportation

This section introduces the process of transporting target knowledge during the unimodal encoding procedure.

#### 1. Input Space Translation & Reformulation

This module is leveraged to distill the object-level target information in complex visuals, and generate a synthetic context sequence with its features learned by a pretrained language model[1]. The detailed information of this module is displayed in Fig. 3.

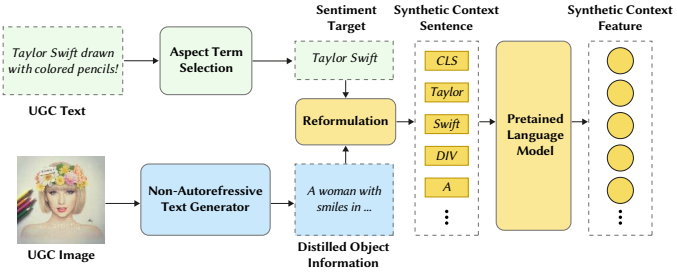



Fig. 3. Procedure of Input Space Translation & Reformulation.

Let $C_u$ denote the content generated by user u, which contains a piece of text information (e.g. tweets, retweets or comments) $T_u$ and an associated image $I_u$. To follow the work of [34], we exploit a transformed-style architecture to implement a non-autoregressive text generator $G_{A_u}(I_u)$, through which the object-level target information is distilled into caption sentences. To leverage the aspect term information in the text modality, as well as the distilled visual information, we reconstruct the auxiliary sentence $S_{A_u}$:

$$\begin{aligned} S_{Au} &= \text{S\_trans}(I_u, T_u) \\ &= \{[\text{CLS}], [T_{as}], [\text{DIV}], [G_{Au}(I_u)]\} \\ &= \{[\text{CLS}], [t_1, \ldots, t_k], [\text{DIV}], [c_1, \ldots, c_m]\} \end{aligned} \tag{1}$$

where $T_{as}$ denotes the aspect term containing a sub-sequence of $k$ words {$t_1$, ..., $t_k$} from text information $T_u$, $G_{A_u}(I_u)$ denotes the caption generator [34], implemented by utilizing a carefully designed variant model from DETR (DEtection TRansformer) [43] to get the $m$ caption words {$c_1$, ..., $c_m$}. Parameters of the caption generator $G_{A_u}$ in (1) are pretrained well and frozen during the whole experiment.

With the obtained auxiliary sentence, we can explicitly transport the context knowledge into the linguistic encoding process in the input space. And for the visual stream, the context knowledge can be transported in the feature space, thus we adopt a pretrained language model, which shares the same parameters with the linguistic encoder, to get the auxiliary context features $F_{au}$.

#### 2. Linguistic Knowledge Transportation & Encoding

This module is designed to incorporate target information while implementing the linguistic encoding process. As shown in Fig. 2, linguistic target interaction is leveraged before the encoding stage. As the pre-trained language models (like BERT, RoBERTa [6], [25]) can help acquire contextualized word representations with initialized parameters, which get well-trained over a large corpus. Thus the transformer-style encoders in sentence-pair classification mode are leveraged to integrate the input sentence and the reconstructed auxiliary sequence. The target-sensitive linguistic feature $\mathbf{L}_t$ can be achieved as:

$$\mathbf{L}_t = \text{LM}([\text{CLS}], T_u, [\text{SEP}], \text{S\_trans}(I_u, T_u)) \tag{2}$$

In (2), LM denotes the pretrained language models like BERT, RoBERTa, etc. '[CLS]' and '[SEP]' are the special tokens in the vocabulary used for classification and separation, and S_trans is the operation of input space translation introduced earlier in Fig. 2.

#### 3. Visual Knowledge Transportation & Encoding

Dually, this module is designed to incorporate target information into visual feature space during the encoding process. Encoded visual features are firstly extracted from an input image $I_u$ by ResNet [44]. The output size of the last convolutional layer in ResNet is $l \times l \times d_v$, where $l \times l$ denotes the $l^2$ block regions of an input image, $d_v$ denotes the depth of feature map. The extracted visual feature of block regions $\{\mathbf{f}^i\}_i^{l \times l}$ is fed into a linear transformation with matrix $W_v \in \mathbb{R}^{\text{dv} \times \text{d\_h}}$), in which $d_h$ denotes the dimension of hidden states from BERT encoder. Thus the visual feature $\mathbf{V}$ is projected into the same space as the linguistic feature to match the embedding size of BERT:

$$\begin{aligned} \mathbf{V} &= \mathbf{W}_v \cdot \{\mathbf{f}^i\}_{i=1}^{l \times l} \\ &= \mathbf{W}_v \cdot \text{ResNet}(I_u) \end{aligned} \tag{3}$$

With the obtained auxiliary context features $\mathbf{F}_{au} \in \mathbb{R}^{N_2 \times d_h}$ and visual features $\mathbf{V}$ in (3), the target-sensitive visual representations can be achieved through visual target interaction, which conducts attentive interaction. Specifically, for the output of $i$-th head $\mathbf{O}_{head}^i$, we acquire the necessary query, key-value vectors through linear feature projection: $\mathbf{Q}_b^i = \mathbf{W}_q^i \mathbf{O}_b$, $\mathbf{K}_b^i = \mathbf{W}_k^i \mathbf{V}$, $\mathbf{V}_b^i = \mathbf{W}_v^i \mathbf{V}$. The vectors are then used for calculating the attention output of the i-th head:

$$\begin{aligned} \mathbf{O}_{head}^i &= \text{softmax}\left(\frac{\mathbf{Q}_b^{i\,T} \cdot \mathbf{K}_b^i}{\sqrt{d_h}}\right)\mathbf{V}_\mathbf{b}^\mathbf{i} \\ &= \text{softmax}\left(\frac{[\mathbf{W}_q^i \mathbf{O}_b]^T \cdot [\mathbf{W}_k^i \mathbf{V}]}{\sqrt{d_h}}\right)\mathbf{W}_v^i \mathbf{V} \end{aligned} \tag{4}$$

All the attention outputs of m such heads $\mathbf{O}_{head}^1, \mathbf{O}_{head}^2, \ldots, \mathbf{O}_{head}^m$ in (4) are concatenated together, followed by a projection matrix $\mathbf{W}_j$ to get aggregated representation of $m$ heads with residual connection and layer normalization (denoted as LN):

$$\mathbf{M}_t = \text{LN}(\mathbf{V} + \mathbf{W}_j[\mathbf{O}_{head}^1, \mathbf{O}_{head}^2, \ldots, \mathbf{O}_{head}^m + b_j]) \tag{5}$$

Moreover, a dense layer and another residual connection are utilized from input $\mathbf{V}$ to the non-linear activated output feature of $\mathbf{M}_t$ in (5), followed by layer normalization to acquire the final target-sensitive visual feature $\mathbf{V}_t$:

$$\begin{aligned} \mathbf{V}_t &= \mathbf{V} + \sigma(\mathbf{M}_t + b_t) \\ &= \mathbf{V} + \sigma(\mathbf{W}_m \cdot \mathbf{W}_j[\mathbf{O}_{head}^1, \mathbf{O}_{head}^2, \ldots, \mathbf{O}_{head}^m] + b_t) \end{aligned} \tag{6}$$

where [·] denotes the concatenation operation in feature dimension, $\sigma$ is the non-linear activation function GELU [45], $\mathbf{W}_j \in \mathbb{R}^{(m \times d_h) \times d_h}$, $\mathbf{W}_m \in \mathbb{R}^{d_h \times d_h}$ are trainable parameters, $\mathbf{V}_t$ in (6) denotes target-sensitive visual feature, the final output of visual knowledge transportation.

### B. Inter-Modality Complementary Knowledge Transportation

To enhance the interactivity across image and text modality, we exploit a recently proposed technique viz. optimal transport kernel (OTK) to incorporate information between heterogeneous modalities. OTK incorporates the idea of the optimal transport plan and kernel methods to fuse the unimodal features with varying dimensions and dependencies.

Let $\mathbf{V}_t = (v_1, v_1, ..., v_n)$ be the target-sensitive visual feature, $\mathbf{L}_t = (l_1, l_1, ..., l_p)$ denotes the target-sensitive linguistic feature obtained in (4) ($n = p$ is not necessary). Let $\kappa$ be the Gaussian kernel with reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ and its associated kernel embedding $\varphi$: $\mathbb{R}^d \to \mathcal{H}$. Then we can get the $n \times p$ cost matrix $\mathbf{K}$ which carries the comparisons $\kappa(v_i, l_j)$ before alignment.

[1] https://github.com/saahiluppal/catr/

Then the transport plan between $\mathbf{V}_t$ and $\mathbf{L}_t$, denoted by the $n \times p$ matrix $\mathbf{P}(\mathbf{V}_t, \mathbf{L}_t)$ is defined as the unique solution of:

$$\min_{\mathbf{P} \in U} \sum_{ij} \mathbf{C}_{ij} \mathbf{P}_{ij} - \varepsilon H(\mathbf{P}) \tag{7}$$

$$H(\mathbf{P}) = -\sum_{ij} \mathbf{P}_{ij}(log(\mathbf{P}_{ij}) - 1) \tag{8}$$

where $C_{ij}$ in (7) represents the pairwise costs for aligning the elements of $\mathbf{V}_t$ and $\mathbf{L}_t$. Equation (8) is the optimizing objective in the space of admissible couplings. To follow the recent work of [46], we choose $\mathbf{C} = -\mathbf{K}$ in our implementation, then the interaction based on transport matrix $\mathbf{P}(\mathbf{V}_t, \mathbf{L}_t)$ is defined as:

$$\begin{aligned} \Phi_l(v) &= \sqrt{p} \times \left( \sum_{i=1}^{n} \mathbf{P}_{i1} \varphi(v_i), \ldots, \mathbf{P}_{ip} \varphi(v_i) \right) \\ &= \sqrt{p} \times \mathbf{P}(\mathbf{V_t}, \mathbf{L_t})^T \varphi(v) \end{aligned} \tag{9}$$

The hybrid linguistic feature $\mathbf{L}_h$ is obtained by aggregating the original target-sensitive linguistic feature $\mathbf{L}_t$ and the visual transported interaction feature $\mathbf{R}_{v \to l} = \Phi_l(v)$ from (9):

$$\mathbf{L}_h = \mathbf{L}_t \oplus \mathbf{R}_{v \to l} = \mathbf{L}_t \oplus \Phi_l(v) \tag{10}$$

Similar to (10), we obtain the final hybrid visual feature $\mathbf{V}_h$ through (11), which is also a result of concatenation:

$$\mathbf{V}_h = \mathbf{V}_t \oplus \mathbf{R}_{l \to v} = \mathbf{V}_t \oplus \Phi_v(l) \tag{11}$$

### C. Multimodal Feature Fusion

Multimodal feature transfusion is designed to conduct the final stage of fusion, which captures the correlation between elements from different modalities. Through the above phase of OTI, two hybrid features $\mathbf{V}_h$ and $\mathbf{L}_h$ are obtained. OTI module adopts different vectors as queries to produce weighted representations, which are sensitive to features from different information streams. However, the features across modalities are involved in interactions through transportation weights, while the direct fusion of element values is still missing. So we propose to use multimodal feature transfusion based mainly on multi-head self-attention to capture the missing correlation in the element level. The input of multimodal feature transfusion is organized based on the two obtained hybrid features $\mathbf{V}_h$, $\mathbf{L}_h$ by OTI, and the target-sensitive visual feature $\mathbf{V}_t$.

The pooling operation is leveraged on visual features by taking the transformation of the first token. Then the output products are concatenated with linguistic features $\mathbf{I}_m$:

$$\mathbf{v}_{pooling} = \tanh(\mathbf{W}_p \cdot \mathbf{V}_t[1] + b_p) \tag{12}$$

$$\mathbf{v}'_{pooling} = \tanh(\mathbf{W}'_p \cdot \mathbf{V}_h[1] + b'_p) \tag{13}$$

$$\mathbf{I}_m = \mathbf{v}_{pooling} \oplus \mathbf{v}'_{pooling} \oplus \mathbf{L}_h \tag{14}$$

where $\mathbf{X}[1]$ in (12) and (13)denotes the first element of $\mathbf{X}$. And $\mathbf{I}_m \in \mathbb{R}^{(N_1+2) \times d_h}$ in (14) are fed into the multimodal feature transfusion module, which outputs $\mathbf{O}_m$ as logits fed into the classifier.

### D. Optimization Process

After the forward process of multimodal feature transfusion, we get the final multimodal hidden states $\mathbf{O}_m$ for sentiment classification. Following previous work [4], [47], the pooled output of the first token is adopted. which is denoted as $\mathbf{H}_p \in \mathbb{R}^{d_h}$ and fed into a linear function followed by a softmax function for classification:

$$p(y|\mathbf{H}_p) = \text{softmax}(\mathbf{W}_c^t \cdot \text{dropout}(\mathbf{H}_p)) \tag{15}$$

In (15) $\mathbf{W}_c \in \mathbb{R}^{c \times d_h}$, c is the category number of dataset. All the parameters in OtarNet are optimized through back propagation while minimizing the standard cross-entropy loss function defined in (16):

$$\mathcal{L} = -\frac{1}{|\mathcal{D}|} \sum_{k=1}^{|\mathcal{D}|} \textbf{log} \quad p(y^k|\mathbf{H}_p^k) \tag{16}$$

The overall training process is displayed in the following algorithm 1, in which line 1-6 initializes model parameters and input data, line 7-10 includes the process of input space translation based on input data, line 11-12 represents for the forward process of OtarNet to acquire final representation $\mathbf{H}_p$ and line 14-17 refers to the optimization method in details.

**Algorithm 1.** Training Process of OtarNet

**Input**: Training Set $\mathcal{D}$, max number of epochs $N_{epoch}$, batch size $\beta$, learning rate $\eta$, parameters of caption generator $\theta_{cg}$,

**Output**: $\theta_{OtarNet}$

1: Initialize caption generator $CG(\theta_{cg})$
2: repeat
3: **for** $i = 1 \to \left[\frac{\mathcal{D}}{\beta}\right]$ **do**
4: $mini\_batch \leftarrow sample(T, \beta)$
5: $L \leftarrow 0$
6: **for** $S \in mini\_batch$ **do**
7: Forward image through encoder: $\mathbf{V} \leftarrow ResNet(I)$
8: Forward image through caption generator: $caption \leftarrow CG(I)$
9: Tokenize caption and tweet sentence $S_t$
10: Obtain $S_{au}$ and $F_{au}$ via the input space translation module: $\{S_{au}, F_{au}\} \leftarrow MB(caption, S_t)$
11: Forward $\{S_t, S_{au}, F_{au}, V\}$ to get features: $\mathbf{L}_t \leftarrow \text{BERT}(S_t + S_{au})$ $\mathbf{V}_t \leftarrow \text{ImplicitFusion}(\mathbf{V}, \mathbf{F}_{au})$
12: Forward $\{\mathbf{L}_t, \mathbf{V}_t\}$ to get the final $\mathbf{H}_p$
13: $\mathcal{L}(\mathbf{H}_p) \leftarrow -\frac{1}{|\mathcal{D}|} \sum_{k=1}^{|\mathcal{D}|} \textbf{log} \quad p(y^k|\mathbf{H}_p^k)$
14: $\mathcal{L} \leftarrow \mathcal{L} + \mathcal{L}(\mathbf{H}_p)$
15: **end for**
16: Update $\theta_{OtarNet}$ using $\triangle \mathcal{L}$
17: **end for**
18: **until** the evaluation results on the validation set drop continuously or this process has been iterated for $N_{epoch}$ times

## V. Experiments

### A. Datasets

We evaluate the OtarNet on three widely used benchmarks, including Twitter-15 [48], Twitter-17 [49] and Yelp[2]. Their introduction are shown in Table I, with details displayed as follows:

TABLE I. Differences of Benchmarks

| Datasets | Twitter-15 | Twitter-17 | Yelp |
|---|---|---|---|
| Data Source | Tweets in 2014-2015 | Tweets in 2016-2017 | Yelp Reviews |
| ContextSource | NA Text Generation | NA Text Generation | CrowdSourcing |
| Annotation Method | Nichesourcing | Nichesourcing | CrowdSourcing |
| Aspect Categories | Open Domain | Open Domain | Services |
| Class Number | 3 | 3 | 5 |

[2] https://www.yelp.com/dataset

### 1. Twitter-15 and Twitter-17

These two sets consist of tweets including text and images posted during 2014-2015 and 2016-2017 respectively, whose sentiment labels over targets(i.e., entities in text), assigned from set {negative, neutral, positive} were supplemented later by [4]. The context information is collected by [34], which leveraged an object-aware transformer followed by a single-pass non-autoregressive text generation approach. The sentiment polarities toward each target were labeled by taking the majority label among three domain experts (Nichesourcing). The aspect categories contain various internet figures or events. Their statistics are displayed in Table II.

TABLE II. STATISTICS OF TWITTER DATASET

| DataSet | Split | Positive | Neutral | Negative | Total |
|---|---|---|---|---|---|
| Twitter15 | Training | 928 | 1883 | 368 | 3179 |
| | Validation | 303 | 670 | 149 | 1122 |
| | Test | 317 | 607 | 113 | 1037 |
| Twitter17 | Training | 1508 | 1638 | 416 | 3562 |
| | Validation | 515 | 517 | 144 | 1176 |
| | Test | 493 | 573 | 168 | 1234 |

### 2. Yelp

The third dataset we use is Yelp corpora obtained from Yelp Dataset Challenge. This corpora contains elaborate information on businesses across 10 cities, where we leverage complement reviews, photos, and corresponding captions. The sentiment polarities are labeled by directly taking the user ratings (from 1-5), and the task is a standard five-class classification problem. The evaluated categories are restricted to the provided services, such as their food, drink, environment, etc. Compared to Twitter sets, Yelp performs more fine-grained classification (5-class) based on well-organized reviews for specific domains. The statistics of Yelp are displayed in Table III.

TABLE III. STATISTICS OF YELP DATASET

| Ratings | Training Set | Testing Set | Validation Set | Total |
|---|---|---|---|---|
| 1 | 1248 | 384 | 387 | 2019 |
| 2 | 774 | 291 | 256 | 1321 |
| 3 | 1926 | 699 | 628 | 3253 |
| 4 | 504 | 164 | 177 | 845 |
| 5 | 1737 | 531 | 597 | 2865 |
| Total | 6189 | 2069 | 2045 | 10303 |

## B. Evaluation Metrics

Following the previous work [4], [34], [50], we adopt Accuracy and Macro-F1 score (M-F1 for short) as our evaluation metrics. Accuracy can be calculated as follows:

$$\text{Accuracy} = \sum_{i=1}^{C}(\text{TP}_i + \text{TN}_i)/|\mathcal{D}_{test}|$$

where $C$ is the category numbers, $|D_{test}|$ is the total sample numbers in the test set. $\text{TP}_i$, $\text{TN}_i$ are the numbers of True Positive and True Negative samples for the i-th category. To calculate Macro-F1, we firstly need to calculate the F1 score for each category based on their scores of precision and recall. The calculation of F1 for the $i$-th category $\text{F1}_i$ is defined as:

$$\text{Precision}_i = \text{TP}_i/(\text{TP}_i + \text{FP}_i)$$

$$\text{Recall}_i = \text{TP}_i/(\text{TP}_i + \text{FN}_i)$$

$$\text{F1}_i = \frac{2*\text{Precision}_i*\text{Recall}_i}{\text{Precision}_i+\text{Recall}_i}$$

where $\text{TP}_i$, $\text{FP}_i$, $\text{FN}_i$, $\text{TN}_i$ are the numbers of True Positive, False Positive, False Negative and True Negative samples for the i-th category. Based on the obtained F1 scores of each class, the Macro-F1 is defined as an average based on the categories:

$$\text{Macro} - \text{F1} = \sum_{i=1}^{C}\text{F1}_i/C$$

## C. Experimental Settings

In our implementation, all the experiments are conducted with Pytorch on one 32G Tesla V100 GPU. We initialized pretrained weights of language models from HuggingFace[3]. As displayed in Table IV, the batch size is set as 32, and the maximum number of training epochs is set to 9. We apply the early stop strategy to avoid over-fitting. We train the models with an Adam weight decay optimizer with an initial learning rate of 5e-5. The optimal hyper-parameters are obtained by grid search. To ensure further reliability of our results and facilitate later explorations, we make our codes publicly available at https://github.com/TomatoNLPer/OTarNet

For the input images, we adopt a pre-trained ResNet-152[4], which outputs a feature map of size $7 \times 7 \times 2048$, indicating 49 block regions with depth $d_v$ as 2048. For the input text, we maintain the standard configuration of BERT/RoBERTa and stack 12 BERT layers. The feature dimension of hidden state output by one BERT layer dh is 768, which is calculated by inner multi-head attention with $m = 12$ heads. We then truncate the max input length N1 and the max bridge sentence length $N_2$ to 125. Besides, the other settings of hyperparameters during the training process are displayed in Table IV.

TABLE IV. HYPERPARAMETER SETTINGS

| Hyperparameter | Symbol | Value |
|---|---|---|
| Epochs | E | 9 |
| Batchsize | B | 32 |
| Dropout | d | 0.15 |
| Learning Rate | lr | 5e-5 |
| OTI layer | LOTI | 1 |
| ME Layer | LME | 1 |
| Weight Decay | Wd | 0.01 |
| Optimizer | - | AdamW |
| BERT Weights | - | bert-base-uncased |
| RoBERTa Weights | - | bertweet-base |

## D. Model Zoo

In this subsection, we will give comprehensive introductions to the leveraged baseline models, including:

- **EF-Net**: An attention capsule extraction and multi-head fusion network for MABSA, which is established based on multi-head attention (MHA) and the ResNet-152
- **Res-MGAN**: A combination of textual and visual contents from ResNet and MGAN. It is implemented by concatenating the pooling results of ResNet and MGAN, which is a multi-grained attention network proposed in [51] for fusing the target and the context.
- **Res-BERT + BL**: Similar to Res-MGAN, Res-BERT + BL is a combination of textual and visual content from ResNet and BERT. BL denotes another BERT layer on the top, which is leveraged for feature fusion.

[3] https://huggingface.co/models

[4] https://download.pytorch.org/models/resnet152-b121ed2d.pth

- **mPBERT**: A variant of mBERT, which uses the max pooling of visual features and first token pooling ([CLS]) to obtain the final output.
- **RelConsTransLG**: A constituent-based transformer, which applies meta auxiliary learning to generate labels on edges between tokens, and can induce constituents without constituent parsers for MABSA.
- **TomBERT**[5]: A multimodal backbone leveraging a target attention mechanism to perform target-image matching, which is helpful for deriving target-sensitive visual representations.
- **EF-CapTrBERT**[6]: A two-stream multimodal backbone, which leverages space translation to construct an auxiliary sentence for language models.

## E. Overall Performance

We compare OtarNet against a collection of neural network based or modified transformer models designed for multimodal aspect-based sentiment analysis. The model performance on three benchmark datasets is displayed in Table V, in which the Twitter results of compared models are directly quoted from published articles, and the Yelp results are obtained through our reimplementation [52].

Based on the displayed experimental results we can make a couple of observations: (1) Our model OtarNet outperforms former multimodal models and achieve the new SOTA performance on three benchmark datasets, demonstrating the effectiveness of our work. (2) Compared to the second-best model, our OtarNet further enhance the performance by a margin and achieve an average of 2.91%, 5.56%, and 2.04% performance improvement respectively on the three datasets from the perspective of Macro-F1, which further indicates the effectiveness of our framework.(3) Since the corpora of Twitter datasets inevitably contains noisy UGC on Twitter websites, the model performance is relatively lower than those in well organized Yelp dataset.

## F. Further Analysis

The previous SOTA model EF-CapTrBERT is sub-optimal, which leverages distilled visual knowledge to perform the early fusion. The degradation may come from the noise during distillation and the lack of complementary details from the visual stream. Thus EF-CapTrBERT finally achieved competitive results to a variation of our model OtarNet + T-Trans, which adopt similar settings to merely transport target knowledge to the text modality. And another competitive baseline TomBERT leverage cross-model attention mechanisms to realize target-sensitive visual representations. Their methods neglect the process of maintaining target sensitivity in linguistic encoding, which is a main difference that leads to limited performance. Compared to previous SOTA TomBERT, EF-CapTrBERT, and our variations, the performance improvements of OtarNet + M-Trans confirm the achievements of our goals, including maintaining target sensitivity and complementary knowledge transportation, which get further verification and analysis in the following ablations and visualizations.

## G. Effectiveness of Target Knowledge Transportation

To achieve the objective of transporting target category knowledge in visual components, we exploit a transformer-style architecture to distill the object-level target information in complex visuals. The overall procedure is displayed in Fig. 3. Contrastively, we conduct ablation experiments on three test sets by decomposing the Input Space Translation Module. Specifically, different transportation plans are closely attempted, including T-Trans, I-Trans and M-Trans, which means transporting the target information into different modalities (Text, Image, Multimodal) to explore the effectiveness of bidirectional target knowledge transportation.

As shown in Table V, all the settings of multimodal data with different transportation settings achieve better performance compared to those without target knowledge, which can well confirm the bridge effect in fusing information from multiple modalities. For the transportation plans of target knowledge, M-Trans achieves expected predominant results compared to models with other settings, which turns out that bidirectional integration can achieve better accomplishment for the goal of incorporating target knowledge, and function better bridge effect compared to unimodal transportation plans.

To provide a more intuitive comparison, we provide attention maps in Fig. 4 for the weights in the matrix of optimal transportation plans. With the transportation of distilled visual information (b), OtarNet pays more attention to the sentiment target in visuals compared to (a). The comparison results are explicitly met with our goal of maintaining target sensitivity.

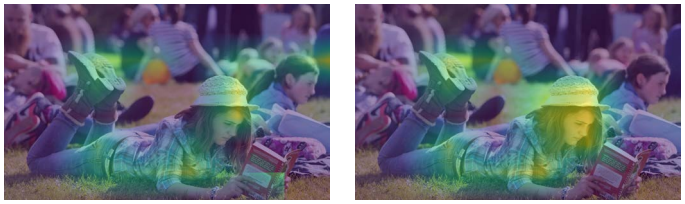
[5] https://github.com/jefferyYu/TomBERT

[6] https://github.com/codezakh/exploiting-BERT-thru-translation

TABLE V. Overall Performance on Two Twitter Datasets and Yelp

| Comparisons | Model | Twitter-15 | | Twitter-17 | | Yelp | |
|---|---|---|---|---|---|---|---|
| | | Accuracy | Macro-F1 | Accuracy | Macro-F1 | Accuracy | Macro-F1 |
| Baselines | Res-MGAN [4] | 71.65 | 63.88 | 66.37 | 63.04 | 80.62 | 71.37 |
| | EF-Net [50] | 73.65 | 67.90 | 67.77 | 65.32 | 81.33 | 71.50 |
| | Res-BERT+BL [4] | 75.02 | 69.21 | 69.20 | 66.48 | 80.93 | 71.75 |
| | mPBERT(CLS) [48] | 75.79 | 71.07 | 68.80 | 67.06 | 79.81 | 68.52 |
| | RelConsTransLG [53] | 76.80 | 73.30 | 69.80 | 68.50 | 80.15 | 68.25 |
| | TomBERT(FIRST) [4] | 77.25 | 71.75 | **70.34** | 68.03 | 81.46 | 73.44 |
| | EF-CapTrBERT [34] | **78.35** | **73.61** | 69.93 | **68.90** | **82.14** | **74.15** |
| Ours | OtarNet - TG | 74.67 | 69.33 | 68.34 | 67.52 | 78.87 | 71.67 |
| | OtarNet + T-Trans | 78.23 | 72.94 | 72.54 | 70.68 | 81.67 | 74.68 |
| | OtarNet + I-Trans | 76.48 | 70.33 | 69.68 | 68.27 | 80.85 | 73.11 |
| | OtarNet + M-Trans | **80.63** | **76.32** | **74.57** | **72.73** | **84.83** | **76.66** |
| Margin | $\delta_{\text{ours-second\_best}}$ | △**2.28** | △**2.71** | △**4.23** | △**3.83** | △**2.69** | △**1.51** |

The numbers in **bold face** denotes the best results, The numbers in **bold face** denotes the best results in Baselines
**TG** denotes the Target Knowledge obtained through the input space translation module.



(a) Attention weights without target knowledge transportation

(b) Attention weights with target knowledge transportation

Fig. 4. In this sample, the distilled target knowledge through **S_trans** in (1) is: ***A girl sitting on the ground with a baseball bat***. As the attention map in (a), the visual features without target knowledge are less discriminative. By integrating the distilled context information, OtarNet captures better visual semantics as shown in (b). More visualizations are provided in the later section of **Qualitative Analysis**.

### *H. Effectiveness of Optimal Transport Interaction*

We decompose the optimal transport interaction into two single parts, including Image_to_Text transportation (IOT) and Text_to_Image transportation (TOI), then add one or both of them to OtarNet to demonstrate the effectiveness of transport modules. To make a further step, we explore two ways of input to produce different query vectors for transport plans, which further explains the combined effect of the target knowledge and optimal transport interaction.

Depicted by the results of accuracy (ACC.) and Macro-F1 (F1) in Table VI, outputs with bidirectional optimal transport interaction achieve the best performance improvement due to the bidirectional complementary knowledge transportation. The superiority of OtarNet + Bi-CKT indicates that, the proposed Optimal Transport Interaction method succeeds in transporting the inter-modal complementary knowledge, and produces hybrid features with more comprehensive information for analyzing targets' sentiment. Besides, we also perform an ablation study and deepen the block architecture by removing or stacking the same transport layer, as Fig. 5 shows. The ablation of the OTI layer naturally brings performance degradation. Nevertheless, we observe little or tiny performance boost with a deeper interaction module, indicating that two layers of optimal transport interaction are sufficient for transporting complementary knowledge. This may be because the caption carrying part of the target information, as enhanced image attributes, has interacted with linguistic information through input space translation. [34].

TABLE VI. Effectiveness of Optimal Transport Interaction

| Datasets &Methods | Twitter-15 | | Twitter-17 | | Yelp | |
|---|---|---|---|---|---|---|
| | ACC. | F1 | ACC. | F1 | ACC. | F1 |
| OtarNet - CKT | 76.82 | 73.74 | 72.01 | 69.84 | 83.24 | 75.53 |
| OtarNet + TOI-CKT | 78.25 | 74.32 | 72.37 | 71.25 | 83.75 | 76.14 |
| OtarNet + IOT-CKT | 78.73 | 74.57 | 72.98 | 71.65 | 83.55 | 75.74 |
| OtarNet + Bi-CKT | **80.63** | **76.32** | **74.57** | **72.73** | **84.83** | **76.66** |

CKT is the abbreviation for complementary knowledge transportation. TOI-CKT refers to Text_to_Image knowledge transportation, IOT-CKT refers to Image_to_Text transportation, and Bi-CKT refers to the bidirectional optimal transport interaction.

### *I. Effectiveness of Multimodal Feature Transfusion*

In this section, we conduct two main experiments on the Yelp dataset for the multi-head attention-based fusion (MHAF) module to validate the effectiveness and find the best settings. We ablate the MHAF module and feed multiple types of features to the classifier. The results are shown in Fig. 6.

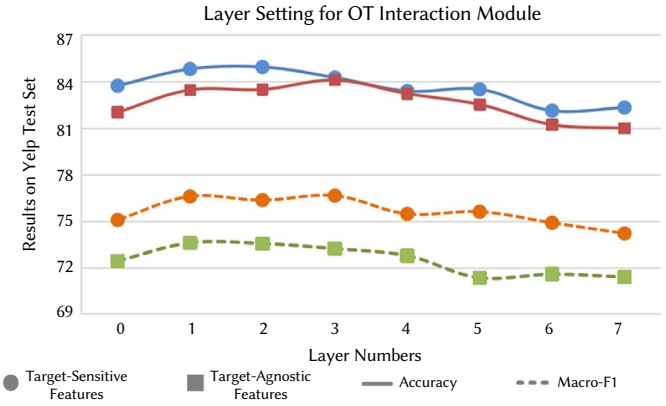
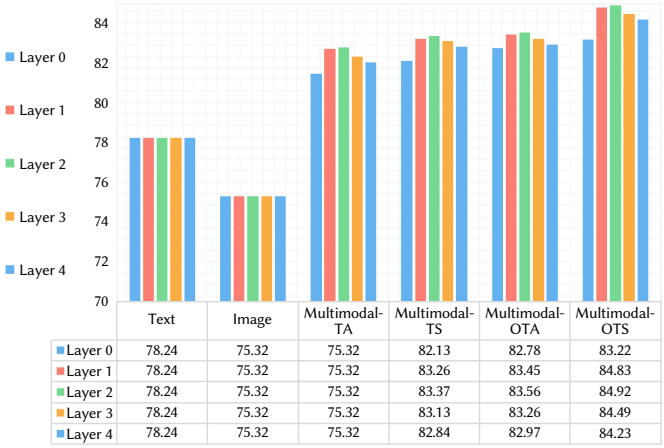



Fig. 5. Layer settings of optimal transport interaction module and results on Yelp test set.




| | Text | Image | Multimodal-TA | Multimodal-TS | Multimodal-OTA | Multimodal-OTS |
|---|---|---|---|---|---|---|
| Layer 0 | 78.24 | 75.32 | 75.32 | 82.13 | 82.78 | 83.22 |
| Layer 1 | 78.24 | 75.32 | 75.32 | 83.26 | 83.45 | 84.83 |
| Layer 2 | 78.24 | 75.32 | 75.32 | 83.37 | 83.56 | 84.92 |
| Layer 3 | 78.24 | 75.32 | 75.32 | 83.13 | 83.26 | 84.49 |
| Layer 4 | 78.24 | 75.32 | 75.32 | 82.84 | 82.97 | 84.23 |

Fig. 6. Ablation studies for multimodal feature fusion layer. The accuracy results on the Yelp test set are displayed. **TA** denotes the Target-Agnostic feature, acquired by base encoders without target-oriented knowledge transportation. **TS** denotes the Target-Sensitive feature, acquired by the target interaction modules without complementary knowledge transportation. And **OTA**, **OTS** are the **TA**, **TS** features after OT interaction.

Conclusions can be made that the Multimodal-OTS feature is definitely more representative to provide discriminative details for classification. The multi-head attention-based feature fusion (MHAF) module effectively receives an average performance boost of 2.90% due to the inner interactive mechanism. To make a further exploration for the depth of the MHAF module, we add the inner attention layers to conduct another experiment, the results of which are shown in Fig. 6. However, as the same result in OTI, we find the model performance drops slightly or grows at a very slow pace.

### *J. Qualitative Analysis*

In this section, we present some examples from trained models to provide several qualitative analyses, including a case study and attention maps to better understand what OtarNet has learned.

#### *1. Case Study*

Fig. 7 shows two predictions with the text attention visualizations of OtarNet. The displayed representative samples confirm the peculiarity that images focus more on targets while sentences express opinions. This peculiarity emphasizes the necessity of transporting complementary knowledge in different modalities of information. This objective is achieved by OtarNet through the Optimal Transport Kernel method, and can be reflected by text attention visualizations, in which key opinion words in sentences are successfully captured. Results are obtained through Text-BERT, Multimodal-BERT, and

our OtarNet, which incorporate different input information. For the confusing information (e.g., 'Best Independent Film for Fruitvale') in unimodal, Text-BERT fails to leverage the combined effect from multiple sources and Multimodal-BERT may misunderstand the key target in visuals. However, our OtarNet with multiple knowledge transportation leveraging context information succeeds to catch the key differences and make accurate predictions.

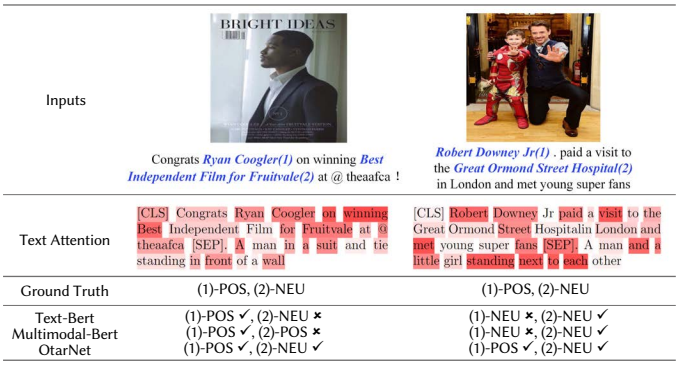



Fig. 7. Examples that text-only or classic multimodal BERT with text and image make the wrong predictions, but our proposed approach with context information insertion gets correct. The text attention indicates the importance of different words computed by the OtarNet.

### 2. Visualizations

In order to further validate the combined effect between images and text, we visualize the obtained attention weights of block regions in the visual feature map as shown in Fig. 8. And it is obvious that some key factors in pictures like emotions or actions are necessary to infer the implicit sentiment in text. Specifically, for the target 'NFL' in the first example, it's nearly impossible to infer the implied negative orientation based only on text modality. Nevertheless, with the action of clutching her chest and the pained expression on her face in the picture, our OtarNet can make the correct prediction. Similar examples are displayed in Fig. 8, which demonstrates that our OtarNet has the ability to capture key details in visual features helpful for judgment.

## VI. Conclusion

This paper proposes a novel OtarNet for multimodal aspect-based sentiment analysis. Different from previous works, which are suffering from the problems of lacking target interaction and distributional modality gap, our OtarNet leverage multi-stage interaction mechanisms to transport knowledge from multiple perspectives for solving the issues above. To maintain interactions with aspect terms for target sensitivity, we leverage an input space translation and multistage interaction method to capture the intra-modality target knowledge of social media content. To capture the inter-modality complementary knowledge, OtarNet exploits a novel approach of the Optimal Transport Kernel method. Compared to attention mechanisms guided by task-specific loss only, OtarNet based on Optimal Transport offers additional signals by reformulating the multimodal fusion as a transportation problem. Experiments on three real-world datasets demonstrate the effectiveness and superiority of our model, which gets further indicated in the ablation study and visualizations. In general, OtarNet exhibits excellent effectiveness in the fine-grained sentiment analysis of open-domain social media content and cross-modal complementation. In the future, we will consider designing models with other ways of interaction, including graph aggregation or loss regulations. We also plan to apply our model to solve related problems such as fine-grained multimodal aligning in hate detection,

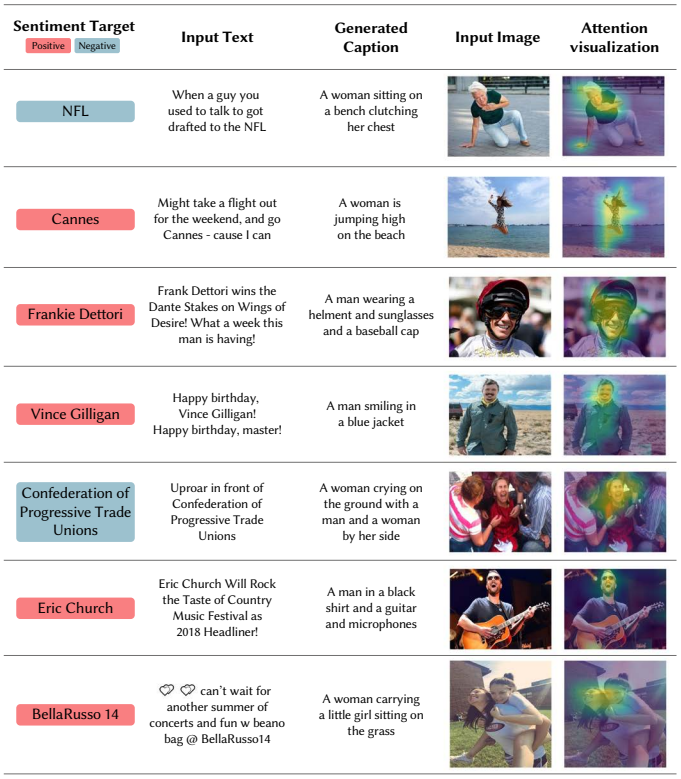



Fig. 8. Attention map of several examples in the Twitter dataset. The auxiliary sentences provided by the non-autoregressive text generator are also provided. We select the top-K values in the optimal transportation plans to make visualizations, which can partly reflect what the OtarNet has learned to integrate. The displayed results confirm the model's capacity of capturing key details in visual features.

as well as multi-lingual applications under low-resource language scenarios. We are also interested in incorporating data from other sources such as speeches or videos.

## Acknowledgement

The work is supported by the National Natural Science Foundation of China (62206267).

## References

[1] F. Huang, X. Zhang, Z. Zhao, J. Xu, Z. Li, "Image– text sentiment analysis via deep multimodal attentive fusion," *Knowledge-Based Systems*, vol. 167, pp. 26–37, 2019.

[2] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," *Knowledge- based systems*, vol. 161, pp. 124–133, 2018.

[3] N. Xu, W. Mao, G. Chen, "Multi-interactive memory network for aspect based multimodal sentiment analysis," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 371–378, AAAI Press.

[4] J. YU, J. JIANG, "Adapting bert for target-oriented multimodal sentiment classification," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019, pp. 5408–5414.

[5] M. E. Basiri, M. Abdar, M. A. Cifci, S. Nemati, U. R. Acharya, "A novel method for sentiment classification of drug reviews using fusion of deep and machine learning techniques," *Knowledge-Based Systems*, vol. 198, p. 105949, 2020.

[6] H. Xu, B. Liu, L. Shu, S. Y. Philip, "Bert post-training for review reading comprehension and aspect-based sentiment analysis," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 2324–2335.

[7] M. Li, L. Chen, J. Zhao, Q. Li, "Sentiment analysis of chinese stock reviews based on bert model," *Applied Intelligence*, vol. 51, no. 7, pp. 5016–5024, 2021.

[8] V. Pérez-Rosas, R. Mihalcea, L.-P. Morency, "Utterance- level multimodal sentiment analysis," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013, pp. 973–982.

[9] V. P. Rosas, R. Mihalcea, L.-P. Morency, "Multimodal sentiment analysis of spanish online videos," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 38–45, 2013.

[10] F. Celli, B. Lepri, J.-I. Biel, D. Gatica-Perez, G. Riccardi, F. Pianesi, "The workshop on computational personality recognition 2014," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1245–1246.

[11] J. G. Ellis, B. Jou, S.-F. Chang, "Why we watch the news: a dataset for exploring sentiment in broadcast video news," in *Proceedings of the 16th international conference on multimodal interaction*, 2014, pp. 104–111.

[12] X. Yang, S. Feng, D. Wang, Y. Zhang, "Image- text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021.

[13] F. Chen, Z. Yuan, Y. Huang, "Multi-source data fusion for aspect-level sentiment classification," *Knowledge- Based Systems*, vol. 187, p. 104831, 2020.

[14] W. An, F. Tian, P. Chen, Q. Zheng, "Aspect-based sentiment analysis with heterogeneous graph neural network," *IEEE Transactions on Computational Social Systems*, 2022.

[15] J. Wagner, P. Arora, S. Cortes, U. Barman, D. Bogdanova, J. Foster, L. Tounsi, "Dcu: Aspect- based polarity classification for semeval task 4," *SemEval 2014*, p. 223, 2014.

[16] B. Pang, L. Lee, S. Vaithyanathan, "Thumbs up? sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 79–86.

[17] M. S. Akhtar, D. Gupta, A. Ekbal, P. Bhattacharyya, "Feature selection and ensemble construction: A two- step method for aspect based sentiment analysis," *Knowledge-Based Systems*, vol. 125, pp. 116–135, 2017.

[18] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, "Target- dependent twitter sentiment classification," in *Proceedings of the 49th annual meeting of the association for computational linguistics*, 2011, pp. 151–160.

[19] N. Liu, B. Shen, "Aspect-based sentiment analysis with gated alternate neural network," *Knowledge-Based Systems*, vol. 188, p. 105010, 2020.

[20] P. Chen, Z. Sun, L. Bing, W. Yang, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 452–461.

[21] Y. Wang, M. Huang, X. Zhu, L. Zhao, "Attention- based lstm for aspect-level sentiment classification," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 606–615.

[22] Q. Liu, H. Zhang, Y. Zeng, Z. Huang, Z. Wu, "Content attention model for aspect based sentiment analysis," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1023–1032.

[23] D. Tang, B. Qin, T. Liu, "Aspect level sentiment classification with deep memory network," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 214–224.

[24] F. Fan, Y. Feng, D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, 2018, pp. 3433–3442.

[25] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[26] C. Sun, L. Huang, X. Qiu, "Utilizing bert for aspect- based sentiment analysis via constructing auxiliary sentence," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 380–385.

[27] D. Borth, R. Ji, T. Chen, T. Breuel, S.-F. Chang, "Large- scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 223– 232.

[28] Y. Zhang, L. Shang, X. Jia, "Sentiment analysis on microblogging by integrating text and image features," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2015, pp. 52–63, Springer.

[29] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 163–171.

[30] H. Wang, A. Meghawat, L.-P. Morency, E. P. Xing, "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 949–954, IEEE.

[31] Y. Yu, H. Lin, J. Meng, Z. Zhao, "Visual and textual sentiment analysis of a microblog using deep convolutional neural networks," *Algorithms*, vol. 9, no. 2, p. 41, 2016.

[32] A. Zadeh, P. P. Liang, S. Poria, P. Vij, E. Cambria, L. Morency, "Multi-attention recurrent network for human communication comprehension," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 5642–5649, AAAI Press.

[33] F. Huang, K. Wei, J. Weng, Z. Li, "Attention-based modality-gated networks for image-text sentiment analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 3, pp. 1–19, 2020.

[34] Z. Khan, Y. Fu, "Exploiting bert for multimodal target sentiment classification through input space translation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3034–3042.

[35] H.-Y. Lin, H.-H. Tseng, X. Lu, Y. Tsao, "Unsupervised noise adaptive speech enhancement by discriminator- constrained optimal transport," *Advances in Neural Information Processing Systems*, vol. 34, pp. 19935–19946, 2021.

[36] L. Chen, Z. Gan, Y. Cheng, L. Li, L. Carin, J. Liu, "Graph optimal transport for cross-domain alignment," in *International Conference on Machine Learning*, 2020, pp. 1542–1553.

[37] E. Grave, A. Joulin, Q. Berthet, "Unsupervised alignment of embeddings with wasserstein procrustes," in *The 22nd International Conference on Artificial Intelligence and Statistics*, vol. 89, 2019, pp. 1880–1890.

[38] T. T. Nguyen, A. T. Luu, "Improving neural cross- lingual abstractive summarization via employing optimal transport distance for knowledge distillation," in *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, 2022, pp. 11103–11111, AAAI Press.

[39] J. Xu, H. Zhou, C. Gan, Z. Zheng, L. Li, "Vocabulary learning via optimal transport for neural machine translation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021, pp. 7361–7373.

[40] L. Chen, G. Wang, C. Tao, D. Shen, P. Cheng, X. Zhang, W. Wang, Y. Zhang, L. Carin, "Improving textual network embedding with global attention via optimal transport," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 5193– 5202.

[41] S. Pramanick, A. Roy, V. M. Patel, "Multimodal learning using optimal transport for sarcasm and humor detection," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 546–556.

[42] Z. Ge, S. Liu, Z. Li, O. Yoshie, J. Sun, "OTA: optimal transport assignment for object detection," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 303-312.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, "End-to-end object detection with transformers," in *16th European Conference on Computer Vision*, vol. 12346, 2020, pp. 213–229, Springer.

[44] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[45] D. Hendrycks, K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," *CoRR*, vol. abs/1606.08415, 2016.

[46] G. Mialon, D. Chen, A. d'Aspremont, J. Mairal, "A trainable optimal transport embedding for feature aggregation and its relationship to attention," in *ICLR 2021-The Ninth International Conference on Learning Representations*, 2021.

[47] X. Li, L. Bing, W. Zhang, W. Lam, "Exploiting BERT for end-to-end aspect-based sentiment analysis," in *Proceedings of the 5th Workshop on Noisy User-generated Text*, 2019, pp. 34–41.

[48] D. Lu, L. Neves, V. Carvalho, N. Zhang, H. Ji, "Visual attention model for name tagging in multimodal social media," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 1990– 1999.

[49] Q. Zhang, J. Fu, X. Liu, X. Huang, "Adaptive co- attention network for named entity recognition in tweets," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, vol. 32, 2018, AAAI Press.

[50] D. Gu, J. Wang, S. Cai, C. Yang, Z. Song, H. Zhao, L. Xiao, H. Wang, "Targeted aspect-based multimodal sentiment analysis: An attention capsule extraction and multi-head fusion network," *IEEE Access*, vol. 9, pp. 157329–157336, 2021.

[51] F. Fan, Y. Feng, D. Zhao, "Multi-grained attention network for aspect-level sentiment classification," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 3433–3442.

[52] D. Q. Nguyen, T. Vu, A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 9–14.

[53] L. M. S. Khoo, H. L. Chieu, "Meta auxiliary labels with constituent-based transformer for aspect-based sentiment analysis," 2020.

Linhao Zhang

Linhao Zhang received a B.S. degree from Xidian University, Xi'an, China, in 2020. He is working towards the PhD degree in Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include affective computing and multimodal learning.

Li Jin

Li Jin received the B.S degree from Xidian University, Xi'an, China, in 2012 and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2017. He is currently an Associate Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences. His research interests include machine learning, knowledge graph and geographic information processing.

Guangluan Xu

Guangluan Xu received the B.Sc. degree from Beijing Information Science and Technology University, Beijing, China, in 2000, and the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2005. He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include data mining and machine learning.

Xiaoyu Li

Xiaoyu Li received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2016 and M.E. degree from Beijing University of Posts and Telecommunications in 2019. He is currently a Research Assistant Fellow at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His research interests include data mining, information extraction, event logic graph and natural language processing.

Xian Sun

Xian Sun received the B.Sc. degree from Beijing University of Aeronautics and Astronautics, Beijing, China, in 2004. He received the M.Sc. and Ph.D. degrees from the Institute of Electronics, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently a Professor with the Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include computer vision, geospatial data mining, and remote sensing image understanding.

Zequn Zhang

Zequn Zhang received a B.Sc. degree from Peking University, Beijing, China, in 2012, and a Ph.D. degree from Peking University in 2017. He is currently a Research Assistant at the Aerospace Information Innovation Institute, Chinese Academy of Science, Beijing, China. His research interests include information fusion knowledge graphs and natural language processing.

Yanan Zhang

Yanan Zhang received the B.S. degree in communication engineering from Sichuan University, Chengdu, China, in 2016, and the Ph.D. degree from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2022. She is currently an Assistant Professor with Sichuan University. Her research interests include multimodal sentiment analysis and question answering.

Qi Li

Qi Li received the B.S degree from Dalian University of Technology, Dalian, China, in 2013 and the Ph.D. degree from the Department of Computer Science and Technology, Tsinghua University, Beijing, China, in 2018. She is currently a Senior Engineer in the Faculty of Psychology, Beijing Normal University. Her research interests include computational psychology and psychoanalysis in social networks.