AnANet: Modeling Association and Alignment for Cross-modal Correlation Classification

Man Xu^{1,2,3}, Junyan Wang³, Yuan Tian^{1,2}, Ruike Zhang^{1,2}, and Wenji Mao^{1,2}

Institute of Automation, Chinese Academy of Sciences

School of Artificial Intelligence, University of Chinese Academy of Sciences

Beijing Wenge Technology Co.,Ltd

Beijing 100190, China

Abstract—The explosive increase of multimodal data makes great demand in many cross-modal applications that follow the strict prior related assumption. Thus researchers study the definition of cross-modal correlation category and construct various classification systems and predictive models. However, those systems pay more attention to the fine-grained relevant types of cross-modal correlation, ignoring lots of implicit relevant data which are often divided into irrelevant types. What's worse is that none of previous predictive models manifest the essence of crossmodal correlation according to their definition at the modeling stage. In this paper, we present a comprehensive analysis of the image-text correlation and redefine a new classification system based on implicit association and explicit alignment. To predict the type of image-text correlation, we propose the Association and Alignment Network according to our proposed definition (namely AnANet) which implicitly represents the global discrepancy and commonality between image and text and explicitly captures the cross-modal local relevance. The experimental results on our constructed new image-text correlation dataset show the effectiveness of our model.

Index Terms—cross-modal correlation, explicit relevant, implicit relevant, decomposition, alignment

I. INTRODUCTION

With the explosive increase of multimodal data in social media, multimodal data (e.g. image-text pair) is in great demand in many cross-modal applications, such as multimodel event classification [1], multimodel topic labeling [2], multimodal sentiment analysis [3], multimodal named entity recognition [4], cross-modal retrieval [5], multimodal hashtag recommendation [6], etc. These works follow the prior assumption that the paired image and text are strictly related and map the multimodal data into the common space via cross-modal correlation learning. But, none of them explicitly define the specific categories of cross-modal correlation during the learning process. Therefore, in practical applications such as social media, these assumptions are fragile and difficult to be strictly satisfied.

The category of cross-modal correlation has been studied in a few prior work. Early researchers focus on the definition of cross-modal correlation category and construct various classification systems according to multiple views, such as cross-modal contribution [7], cross-modal similarity [8], logico-semantic and status relations [9], visual distinction [10], cross-modal dependency [11], systemic functional multimodal analyses [12], etc. However, none of these studies

propose any predictive models for cross-modal correlation categories. Recently, researchers pay more attention to the prediction of the cross-modal correlation categories and expand the existing classification system based on image specificity [13], emotion [14], interrelation metrics [15, 16], parallel and non-parallel [17], contextual and semiotic relations [18], visual content contribution [19], etc. They annotate data and train models to predict the cross-modal correlation category around the following tasks: multimodal regression [13] or multimodal classification [14, 15, 16, 17, 18, 19].

Despite the landmark results of these predictive models, there are still two main drawbacks in the existing studies about the definition and classification of cross-modal correlation. Firstly, existing definitions pay more attention to the finegrained relevant types of cross-modal correlation, leading to a considerable part of implicitly relevant data is divided into irrelevant types according to the existing system. Take the tweet (3) and (4) in Fig.1 as an example, these two image-text pairs will be classified to the category irrelevant according to the previous definition that image does not add to meaning and text is not represented in image [19]. In fact, they have implicit associations based on some relationships (e.g. 'a shouting man' expresses protest against 'work' in Fig.1.3) or uses' sentiments (e.g. finished 'exam' is positively related to 'two cartoon smiling faces' In Fig.1.4). Secondly, previous predictive models combine the multimodal feature representations [19], or utilize cross-modal attention mechanism [16] for image-text classification. But, none of them manifest the essence of cross-modal correlation according to their definition at the modeling stage. For example, the latest work is mainly centered on the role of the text to the tweet's semantics and considers the image as the supplement to the text content, but it does not model this kind of inequality of modality importance in the proposed LSTM + InceptionNet model [19].

In this paper, we focus on the most popular cross-modal data, image-text pair, and present a comprehensive analysis about the image-text correlation based on the implicit association and explicit alignment. According to our definition, it requires the predictive model to have the ability to reason about whether there are implicit associations or explicit alignments between the cross-modal situations. To support our argument, we create a new cross-modal correlation annotation protocol to construct a large scale image-text dataset and propose



Explicit Relevant



Implicit Relevant



Fig. 1. Examples of the three types of text-image correlation in this study. Explicit relevant tweets show detailed cross-modal alignments, such as *dog* and *car* in tweet (1), or *tacos* in tweet (2); images and texts in implicit relevant tweets have none cross-modal alignment but are implicitly associated with each other based on implicit relevant, such as providing some relationship between shouting and protesting in tweet (3) or enhancing users' positive sentiment between image and text in tweet (4); images and texts in tweet (5) and (6) are completely unrelated.

the Association and Alignment Network, namely AnANet. It consists of two parallel streams for visual and linguistic processing. The association net implicitly represents the global discrepancy and commonality between image and text via orthogonal decomposition. The alignment net explicitly captures the local relevance between image regions and text words via an interactive attention mechanism. The main contributions of our work are as follows:

- We redefine a new classification system for image-text correlation based on implicit association and explicit alignment and propose a two-stream framework to model the essence of cross-modal correlation according to our definition.
- We construct the association and alignment network to

- represent the global implicit relevance and local explicit relevance between image and text, which allows for visiolinguistic distinguishable representations and enables sparse cross-modal interaction.
- We compare our model with the existing state-of-the-art methods, and the experimental results on our constructed new multimodal dataset demonstrate the effectiveness of our model in cross-modal correlation classification task.

II. RELATED WORK

This section will detailedly introduce the studies about the learning, definition, and classification of image-text correlation.

A. Cross-modal Correlation Learning

With the popularity of multimodal data in social media, more and more traditional tasks begin to process multimodal data based on cross-modal correlation learning. The image can provide rich visual semantic information for the text and assist in enhancing text content understanding, such as multimodel event classification [1], multimodel topic labeling [2], multimodal named entity recognition [4], and multimodal hashtag recommendation [6]. The theoretical insights of the connection between image and instructional text can be used to estimate automated discourse analysis [20], guide image caption generation [21], support multimodal summarization [22], or learn cross-modal retrieval [5]. The context-related information between the image and text can help analyze user's emotion [3] or sentiment [23], and the context-conflicting information is also useful for multimodal sarcasm detection [24]. These works follow a prior assumption that the paired image and text are strictly related, but none of them explicitly define the specific categories of cross-modal correlation during the learning process. In other words, all these assumption that the text with an image represent similar concepts which are not true in social media (e.g. Twitter) [19].

B. Cross-modal Correlation Definition

Since the modeling of cross-modal correlation is the fundamental component of many multimodal applications, the definition of cross-modal correlation has gradually attracted the attention of researchers. The early definition of crossmodal correlation was established for specific data, such as six types of image-text relations for comic books which in terms of their equal/unequal contributions to meanings of image and text [7], three major categories for web pages according to the similarity between the illustration and text [8], detailed classification system about image-text relations in scientific articles based on the logico-semantic and status relations [9], a taxonomy of four image-text relationships in picture books based on systemic functional multimodal analyses [12]. In recent years, researchers focus on establishing classification systems for open source multimodal data, such as social media data [10, 11]. Chen et al. [10] uncover what people post about and the correlation between the tweet's image and text,

TABLE I
COMPARISON OF DIFFERENT CROSS-MODAL CLASSIFICATION SYSTEMS

Year	Method	Category	Type	Prediction	Domain
1998	McCloud's [7]	1.equal contributions to meanings of image and text 2.unequal contributions to meanings of image and text	Bool	no	Book
2003	Marsh's [8]	1.image expressing little relation to the text 2.image expressing close relation to the text 3.image going beyond the text	Bool	no	Web Pages
2005	Martinec's [9]	1.logico–semantic: expansion or projection 2.status relations: equal or unequal	Bool	no	Scientific Articles
2013	Chen's [10]	1.visually-relevant tweets 2.visually-irrelevant tweets	Bool	no	Social Media
2014	Wu's [12]	1.elaboration 2.extension 3.enhancement 4.divergence	Bool	no	Book
2014	Wang's [11]	1.independent image-text 2.image-text with similar meaning 3.text depending on image 4.image depending on text	Bool	no	Social Media
2015	Chen's [14]	1.visually relevant 2.emotionally relevant 3.emotionally irrelevant	Bool	yes	Social Media
2015	Jas's [13]	the specificity of given multiple descriptions with an image (i.e. 0.0-1.0)	Float	yes	Caption
2017	Henning's [15]	1.cross-modal mutual information 2.semantic correlation	Bool	yes	Caption
2018	Zhang [17]	1.parallel 2.non-parallel	Bool	yes	Ad
2019	Kruk's [18]	1.minimal 2.close 3.transcendent	Bool	yes	Social Media
2019	Kruk's [18]	1.divergent 2.parallel 3.additive	Bool	yes	Social Media
2019	Vempala's [19]	1.image adds to the tweet meaning and text is represented in image 2.image adds to the tweet meaning and text is not represented in image 3.image does not add to meaning and text is represented in image 4.image does not add to meaning and text is not represented in image	Bool	yes	Social Media
2020	Otto's [16]	1.cross-modal mutual information 2.semantic correlation 3.status	Bool	yes	Social Media, Caption, Wikipedia
2021	Our	1.image and text are explicitly relevant 2.image and text are implicitly relevant 3.image and text are irrelevant	Bool	yes	Social Media

showing an important functional distinction between visually-relevant and visually-irrelevant tweets. Wang et al. [11] define the correspondences between image and text in microblog on the basis of their dependency conditions (i.e. dependent or independent), to assist the discovery of multimodal social topics. However, the above research mainly focuses on the construction of classification systems for image-text correlation, none of them propose such predictive models for image-text correlation categories.

C. Cross-modal Correlation Classification

According to the previous definition of image-text correlation, researchers recently pay attention to the prediction of the correlation categories. The early work finds that both visual elements and emotional elements play important roles

in image-text correlation [14], and develop a visual emotional topic model to capture the image-text correlation from visual or emotional perspectives. To measure the similarity between an image with multiple description texts, Jas and Parikh [13] propose the concept of image specificity and turn the cross-modal correlation classification problem into a specificity regression problem. Zhang et al. [17] focus on the parallel and non-parallel relationships between image and text slogan and design 9 features (e.g. topics, slogan specificity and concreteness, etc.) to train an ensemble of SVM classifiers. To explore contextual relations (i.e. minimal, close, or transcendent) and semiotic relations (i.e. divergent, additive, or parallel) behind image-text pair, two taxonomies are also proposed to capture literal and semiotic meanings of image-text pair [18]. According to the role of the image to the semantics of the

tweet, Vempala and Preotiuc-Pietro [19] aim to identify if the image's content contributes with additional information to the meaning of the tweet beyond the text via jointing image-text neural networks. Although the above researches have proposed a variety of image-text correlation classification methods, they still treat this task as the standard image-text classification task via simply combining the multimodal feature encoders for image-text classification at the prediction stage, such as SVM [17], InceptionNet+LSTM [19], ResNet+GRU [18]. In fact, the joint encoding between image and text is very helpful for cross-modal modeling [15, 16]. To describe different interrelations of textual and visual information, Heening and Ewerth [15] propose two metrics, namely cross-modal mutual information(CMI) and semantic correlation(SC). They utilize an Multimodal Autoencoder (InceptionNet+LSTM) to gather a compact embedding, where the first input of the text LSTM layer is the image embedding in order to emulate a natural article processing (reading the text under consideration of the enclosed image). Furthermore, Otto et al. [16] construct a new taxonomy of eight semantic image-text relationships via leveraging another metric describing status relation of imagetext pair into CMI and SC [15] and propose a image guided textual attention layer to ensure that the neural network reads the textual information under the consideration of the visual features, which forces it to interpret the features in unison. Only a small amount of work has considered the interactive relationship between graphics and text in the process of modeling graphics and text embedding

The Table I shows a detailed comparison of existing crossmodal classification systems. There are two main types of image-text correlation, relevant and irrelevant, and the existing work pays more attention to fine-grained relevant correlation classification. However, in practical applications, a considerable part of the data is divided into the irrelevant types according to the existing system which is actually implicitly relevant. And the worst is that, for those predictive methods, little is known about how textual content is related to the images with which they appear. That means none of them manifest the essence of cross-modal correlation according to their definition at the modeling stage. Thus, in this paper, we first define an annotation scheme that focuses on the explicit or implicit correlation between image and text and then propose a multimodal classifier to model cross-modal correlation according to our definition.

III. DEFINITION AND DATA COLLECTION

A. New Definition of Cross-modal Correlation

Different from the various definitions of cross-modal correlation in previous work which pay more attention to the fine-grained relevant types of cross-modal correlation, in this paper, we focus on finding those implicitly relevant data which are often divided into irrelevant types according to the existing classification systems.

We focus on the most popular cross-modal data, imagetext pair, and define the new classification system of imagetext correlation based on implicit association and explicit

TABLE II STATISTICS OF OUR CONSTRUCTED DATASET

	Train	Dev	Test	All
Explicit Relevant	5779	542	558	6879
Implicit Relevant	1070	110	106	1286
Irrelevant	3445	323	319	4087
All	10294	975	983	12252

alignment, including explicit relevant, implicit relevant, and irrelevant. To research the cross-modal correlation and support our argument, we create a new cross-modal correlation annotation protocol according to our definition, which uses the following guidelines:

Explicit relevant. The image and text have the same meaning with explicitly alignments:

- Some or all of the visual regions are depicted in the text (e.g. Fig.1.1).
- Some or all of the textual words are shown in the image (e.g. Fig.1.2).

Implicit relevant. The image and text have the different meanings, but:

- One modality makes a reference to something or depicts something that adds information to another modality (e.g. Fig.1.3).
- One modality enhances the sentiment or expresses a feeling about the content of another modality (e.g. Fig.1.4).

Irrelevant. The information of each modality is independent and does not add any useful content to represent the whole meaning of the image-text pair (e.g. Fig.1.5 and Fig.1.6).

B. Data Collection and Annotation

In this paper, we focus the text-image relationship in social media. The latest research studies how the meaning of the entire tweet is composed through the relationship between its textual content and its image and conduct a Text-Image Relationship dataset [19], including 4,471 multimodal tweets. We expand this dataset and also use Twitter as our data source to collect more image-text pairs, getting a larger scale Crossmodal Correlation Dataset, namely CCD (will be released soon).

We hired 3 annotators to judge the type of each image-text pair independently, and also use the majority vote strategy [19] to aggregate those judgments as the final label. We randomly divide this dataset into the training set, development set and test set, and the detailed statistics are given in Table II.

IV. PROPOSED MODEL

Fig.2 illustrates the overall architecture of our proposed model for cross-modal correlation classification. We first learn the single-modality representations via two independent encoders, and then design the AnANet to model the cross-modal correlation through two parallel streams: an association net to globally weaken the discrepancy and strengthen the commonality; a alignment net to find relevant anchors to capture cross-modal local relevance.

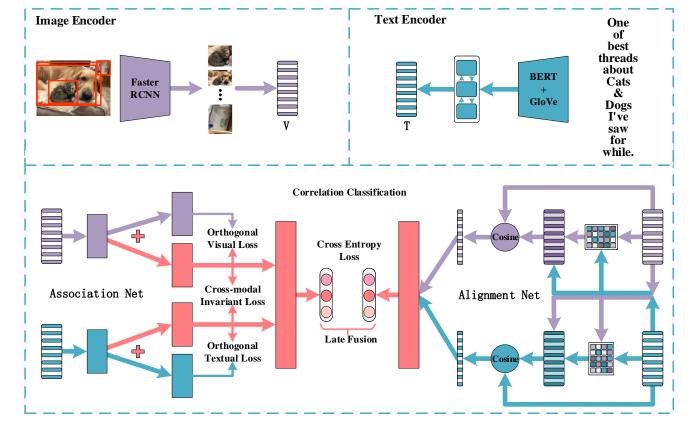


Fig. 2. Overall architecture of our proposed AnANet for cross-modal correlation classification.

A. Encoders

Given an image-text pair (Image, Text), we first use two independent encoders to map them into single-modality feature representations: a set of image features V = $\{v_1, v_2, \cdots, v_K\} \in \mathbb{R}^{K \times d}$ and a set of word features T = $\{t_1, t_2, \cdots, t_N\} \in \mathbb{R}^{N \times d}$, where each image feature v_i encodes a region in the Image, each word feature t_i represents the word w_i in the Text, K and N are the lengths of image region sequence and text word sentence, d is the dimensions of the encoded image and text features.

Image Encoder uses the famous image object detection framework, Faster RCNN [25], to select K image regions and encode them as sequential features $V = \{v_1, v_2, \cdots, v_K\}.$ Unlike prior work [19] directly utilizing InceptionNet to extract the output of the last fully-connected layer as the global image feature, this structure allows for more detailed local objects and enables more visual semantic information.

$$V = \sigma(w_v * FasterRCNN(Image) + b_v) \in \mathbb{R}^{K \times d} \quad (1)$$

where $\sigma(\cdot)$ is the ReLU activation function, w_v, b_v are the parameters of the fully-connected layer.

Text Encoder transforms each word token w_i in the text into continuous representation via both word embedding w_i^G and contextualized embedding w_j^B . The former is embedded by GloVe [26] pretrained on large social media corpus, and

the latter is embedded by BERT [27] which leverages prior linguistic knowledge and has been shown success in capturing syntactic, semantic, and world knowledge. We concatenate these two embeddings and utilize the Bi-directional GRU layer to model the contextual information and generate the final word features T.

$$w_j^G = Embed_{GloVe}(w_j) \in \mathbb{R}^{d_G}$$
 (2)

$$w_j^B = Embed_{Bert}(w_j) \in \mathbb{R}^{d_B} \tag{3}$$

$$w_j^G = Embed_{GloVe}(w_j) \in \mathbb{R}^{d_G}$$

$$w_j^B = Embed_{Bert}(w_j) \in \mathbb{R}^{d_B}$$

$$T = BiGRU([w_j^G \oplus w_j^B]) \in \mathbb{R}^{N \times d}$$
(4)

where \oplus denotes the operation of vector concatenation, d_G and d_B are the dimensions of word embedding and contextualized embedding respectively.

B. Association Net

To focus on the implicit association between image and text, we design the association net to implicitly decompose the image and text into cross-modal discrepancy and commonality, and then reinforce this kind of commonality as the global relevant representation.

Cross-modal Decomposition. It consists of a common layer, visual decomposition layer, and textual decomposition layer. We feed the overall visual and textual features f^* into these three layers and break down them into shared invariant features f_{inv}^* and unique variant features f_{var}^* .

$$f_{inv}^* = Wf^* \in \mathbb{R}^{d_{inv}} \tag{5}$$

$$f_{var}^* = P^* f^* \in \mathbb{R}^{d_{var}} \tag{6}$$

where * indicates the input modality image or text, f^{image} is the average of image feature sequence V, f^{text} is the average of text feature sequence T. To ensure the conflict of different subspaces, we further enforce an additional orthogonal constraint on the unique modality projection matrices $P^* \in \mathbb{R}^{d_{inv} \times d}$ and the shared projection matrix $W \in \mathbb{R}^{d_{inv} \times d}$.

$$W^T P^* = 0 \ (* \in \{image, text\}) \tag{7}$$

Commonality Strengthen. We regard the shared invariant features f_{inv}^* as the representation of cross-modal commonality, and the unique variant features f_{var}^* as the representation of cross-modal discrepancy. In cross-modal correlation classification task, we expect our model to focus more on the relevance between image and text. Thus, we reinforce their cross-modal commonality and weaken their cross-modal discrepancy. Specifically, we abandon the unique variant features and directly concatenate the above shared invariant features as the global relevant representation r_g .

$$r_g = [f_{inv}^{image} \oplus f_{inv}^{text}] \in \mathbb{R}^{2d_{inv}}$$
 (8)

C. Alignment Net

To further represent the explicit alignments between different modality information, we design the alignment net to absorb image features V and text features T, and infer the local relevant anchors by aligning image region and text word.

Interactive Attention. Our interactive attention mechanism firstly calculates the similarity matrix to measure how image regions and text words relate, which is defined with two complementary formulations below.

$$A_{t2v} = Att(V, T) \in \mathbb{R}^{K \times N} \tag{9}$$

$$A_{v2t} = Att(T, V) \in \mathbb{R}^{N \times K} \tag{10}$$

where the attention function $Att(\cdot)$ specified by the input matrix $H \in \mathbb{R}^{L \times d}$ and $Q \in \mathbb{R}^{M \times d}$ is defined as:

$$Att(H,Q) = \left[\frac{\exp(s_{ij})}{\sum_{i} \exp(s_{ij})}\right]_{i,j} \in \mathbb{R}^{L \times M}$$
 (11)

$$s_{ij} = H_{i:}Q_{:j}^T \tag{12}$$

To attend on words with respect to each image region, we calculate the weighted combination of text word features T as the text guided visual attention feature \hat{V} .

$$\hat{V} = A_{t2v} * T \in \mathbb{R}^{K \times d} \tag{13}$$

Similarly, we also calculate the weighted average of image region features V according to the attention matrix A_{v2t} to get the image guided textual attention feature \hat{T} .

$$\hat{T} = A_{v2t} * V \in \mathbb{R}^{N \times d} \tag{14}$$

Pair-wise Comparison. After the interactive attention layer, the attended multimodal features \hat{V} and \hat{T} have aggregated the one modality information with respect to another modality information. Though the strategy of directly combining multimodal features is very common in existing multimodal work [4, 19], it can not measure the degree of local association between image regions and text words in this task. Hence, we design the pair-wise comparison strategy to quantify the crossmodal local relevance. The pair-wise comparison function $f(\cdot)$ uses cosine function to compare the original modality feature V and T with the cross-modal emphasized attention feature \hat{V} and \hat{T} term by term.

$$\overline{v} = f(V, \hat{V}) = [cosine(V_{i:}, \hat{V}_{i:})]_{i=1}^{K} \in \mathbb{R}^{K}$$
 (15)

$$\bar{t} = f(T, \hat{T}) = [cosine(T_i, \hat{T}_i)]_{i=1}^N \in \mathbb{R}^N$$
 (16)

This structure enables our model to measure the degree of local relevance between image regions and text words according to the similarity items pointed by the two indicator sequences \overline{v} and \overline{t} . Then, we concatenate these sparse crossmodal interaction features as the local relevant representation r_l for cross-modal correlation classification.

$$r_l = [\overline{t} \oplus \overline{v}] \in \mathbb{R}^{K+N} \tag{17}$$

D. Classification

Due to the heterogeneity of global and local cross-modal relevant representations r_g and r_l , we utilize the late fusion strategy for final classification. It first uses two fully-connected layers to predict the correlation type of image-text respectively, and then aggregates these two results as the final prediction \hat{y} .

$$\hat{y} = softmax \left(\lambda(w_q r_q + b_q) + \eta(w_l r_l + b_l)\right) \tag{18}$$

where λ and η are hyperparameters to control the contributions of association net and alignment net for cross-modal correlation classification.

E. Optimization

We optimize our model with three losses: classification loss, cross-modal invariant loss, and modality orthogonal loss.

Classification Loss. To optimize the final classification layer, we use the cross entropy loss function which is the most commonly used in classification tasks as the cross-modal correlation classification loss \mathcal{L}_c .

$$\mathcal{L}_c = -\sum_i y_i \log \hat{y}_i \tag{19}$$

where y_i is the ground truth of *i*-th image-text pair (i.e. 0 for irrelevant, 1 for implicit relevant, and 2 for explicit relevant), and \hat{y}_i is the predicted label of our model.

Cross-modal Invariant Loss. In association net, we share the same matrix W to ensure projecting image and text into the same subspace and regard the shared invariant features f_{inv}^* as the representation of cross-modal commonality. To ensure the similarity of decomposed shared invariant features, we propose the cross-modal invariant loss \mathcal{L}_i , denoted as:

$$\mathcal{L}_{i} = \left\| f_{inv}^{image} - f_{inv}^{text} \right\|_{2} \tag{20}$$

TABLE III
COMPARATIVE RESULTS WITH MULTIMODAL BASELINES

Group	Method	Encoder	Acc	F1
_	Random	-	35.45	38.30
1	Majority [19]	-	56.19	40.43
	TFN [3]	FN, LSTM(GloVe)	79.76	79.46
2.	CoMemory [23]	InceptionV3, Memory Network(GloVe)	77.82	77.42
-	SCAN [28]	Faster RCNN, GRU(GloVe)	80.87	79.49
	InceptionNet+LSTM [19]	InceptionNet, LSTM(GloVe)	76.93	75.45
	DCNN(GloVe) [18]	ResNet18, GRU(GloVe)	79.87	78.92
3	DCNN(GloVe+BERT) [18]	ResNet18, GRU(GloVe+BERT)	82.70	82.34
	Deep Classifier [16]	InceptionV2, GRU(GloVe)	80.47	79.14
	AnANet(GloVe)	Faster RCNN, GRU(GloVe)	81.58	81.09
Ours	AnANet(GloVe+BERT)	Faster RCNN, GRU(GloVe+BERT)	84.84	83.84

where $\|\cdot\|_2$ denotes the 2 norm.

Modality Orthogonal Loss. In association net, we also use two unique modality projection matrices P^{image} and P^{text} to decompose the image and text into unique variant subspace under the additional orthogonal constraint (Eq.(7)). Here, we convert this constraint into the following modality orthogonal loss \mathcal{L}_o .

$$\mathcal{L}_{o} = \sum_{* \in \{image, text\}} \|W^{T} P^{*}\|_{F}$$

$$\mathbf{s.t} \quad W^{T} P^{*} = 0 \ (* \in \{image, text\})$$
(21)

where $\|\cdot\|_F$ denotes the Frobenius norm.

According to Eq.(19-21), we finally minimize the combined loss function to optimize our whole network:

$$\mathcal{L} = \mathcal{L}_c + \alpha \mathcal{L}_i + \beta \mathcal{L}_o \tag{22}$$

where α and β are hyperparameters to control the distributions of the cross-modal invariant loss and modality orthogonal loss.

V. EXPERIMENT

A. Implementation Details

We evaluate the proposed AnANet on our constructed CCD dataset for cross-model correlation classification. Each sample in this dataset is an image-text pair. For each image, we first resize it into 224*224 and utilize pretrained Faster RCNN¹ to select 36 image regions and encode them into 1024 image region features. For each text, we embed each word into 200-dimensional word embedding by GloVe² pretrained on the Twitter data and 768-dimensional contextual embedding by BERT³. We set the maximum length of the text sequence to 100 and encode the text into 1024-dimensional hidden space via BiGRU. We map multimodal features into three projection subspaces and set the dimensions of them to 200. We optimize our model by Adam update rule with 64 mini-batch and 0.001 learning rate.

B. Baseline Methods

We compare our model with the following multimodal baseline methods, including the lastest methods in cross-modal correlation classification task (i.e. InceptionNet+LSTM, DCNN, Deep Classifier), several representative methods in multimodal learning tasks which are modified for multimodal classification (i.e. TFN, CoMemory, SCAN), and two simple baseline methods (i.e. Random and Majority).

- **Random** predicts the cross-modal correlation category of the image-text pair randomly.
- **Majority** always predicts the most frequent class in experimental dataset (i.e. explicit relevant category).
- InceptionNet+LSTM [19] concatenates multimodal features generated by InceptionNet and LSTM for multimodal classification.
- **Deep Classifier** [16] force the GRU model to read the textual information under the consideration of the visual features via a image guided attention mechanism.
- DCNN [18] takes the ResNet-18 and bidirectional GRU
 for cross-modal correlation classification, which utilizes
 both word2vec type and pre-trained character-based contextual embeddings (ELMo) for textual embedding. For
 fair comparison, we replace its ELMo embedding with
 the stronger language model BERT used in our model.
- TFN [3] uses a joint tensor fusion network to model the intra-modality and inter-modality dynamics of the multiple information.
- CoMemory [23] models the interactions between image and text iteratively via co-memory mechanism.
- SCAN [28] discovers the full latent alignments using both image regions and words to infer image-text similarity via stacked cross attention.

C. Main Result

We use the accuracy and weighted F1-score as our experimental metrics to evaluate the performance of different methods. Table III shows the comparative results on our constructed dataset and our AnANet achieves the best performance among all compared baseline methods.

¹https://github.com/jwyang/faster-rcnn.pytorch

²https://github.com/stanfordnlp/GloVe

³https://github.com/google-research/bert

TABLE IV
ABLATION STUDY: DIFFERENT COMPONENTS OF ANANET

Me	thod	Module	Encoder	Acc	F1	△Acc	△ F 1
1 AnANet		Association Net, Alignment Net	Faster RCNN, GRU(GloVe+BERT)	84.84	83.84	-	-
2	-BERT	Association Net, Alignment Net	Faster RCNN, GRU(GloVe)	81.58	81.09	-3.26	-2.75
3	-Association Net	Alignment Net	Faster RCNN, GRU(GloVe+BERT)	83.21	82.32	-1.63	-1.52
4	-BERT	Alignment Net	Faster RCNN, GRU(GloVe)	80.87	79.49	-3.97	-4.35
5	-Alignment Net	Association Net	Faster RCNN, GRU(GloVe+BERT)	83.52	83.20	-1.32	-0.64
6	-BERT	Association Net	Faster RCNN, GRU(GloVe)	81.18	80.09	-3.66	-3.75
7	-Association Net, -Alignment Net	AvgPooling, Concatnation	Faster RCNN, GRU(GloVe+BERT)	83.01	82.28	-1.83	-1.56
8	-BERT	AvgPooling, Concatnation	Faster RCNN, GRU(GloVe)	81.28	80.18	-3.56	-3.66

By comparing baseline methods in cross-modal correlation classification task (Group 3 of Table III), our AnANet(GloVe) performs better than those two combination based methods, DCNN(GloVe) and InceptionNet+LSTM, which extract multimodal features via combining different encoders. It indicates that our proposed association net and alignment net can effectively model the cross correlation and learn highquality multimodal representations. Deep Classifier performs better than DCNN(GloVe) and InceptionNet+LSTM but is also weaker than our AnANet(GloVe), which shows that the image guided attention attention mechanism used in Deep Classifier is not enough to capture the bidirectional correlation between image and text, which is modeled by our interactive attention of alignment net. It is important that when we introduce the strong language model BERT into DCNN and our AnANet models, they achieve great performance improvements and our AnANet(GloVe+BERT) performs better than all cross-modal correlation classification baselines. It demonstrates that BERT model which has learned common language representations via pre-training on large-scale corpus is very helpful for downstream tasks, including cross-modal correlation classification task.

We also compare our model with those representative multimodal classification methods in Group 2 of Table III. The SCAN method outperforms all baseline methods, indicating the selection of image encoder is crucial for this task. The SCAN method and our AnANet(GloVe) both use the Faster RCNN as image encoder, which is able to discover the local clues between the image region and the text word than global encoder used in other baseline methods (e.g. InceptionNet, ResNet, FN, etc.). Further, compared with SCAN which also utilizes the interactive attention architecture to capture crossmodal correlation and infers the image-text similarity via modeling the full latent alignments of image regions and words (similar to our alignment net), our AnANet still performs better. It is because that we also introduce an extra association net to focus more on the commonality between image and text and weaken their cross-modal discrepancy, which further indicates the effectiveness of our two-stream framework.

D. Ablation Study

In this section, we carry out extensive ablation studies on different variants of AnANet to evaluate the performance of its components.

TABLE V
COMPARATIVE RESULTS WITH DIFFERENT MODULES

Method	Acc	F1	F_{exr}	F_{imr}	F_{irr}	
AnANet	84.84	83.84	90.22	42.86	87.97	
Association Net Alignment Net	83.52 83.21	83.20 82.32	90.02 89.68	43.65 37.71	86.82 86.59	

1) Different Components of AnANet: The detailed ablation results about different components of AnANet are shown in Table IV. The most obvious declines come from the direct removal of BERT model (see rows 2, 4, 6, 8). Each of these variants which only use the GloVe word vector in the text has a significant performance degradation, indicating that the pre-trained language model can provide better word representations for the downstream task and bring about 3.26% accuracy gain and 2.75% F1 gain on this task. Comparing these two core modules (i.e. Association Net and Alignment Net) used in our model, we find that removing Association Net has greater performance drop than removing Alignment Net (see rows 3, 5), indicating global implicit association captured by Association Net plays a more important role than local explicit alignment extracted by Alignment Net for this task. We also remove all modules proposed in this paper and directly concatenate the multimodal representations for classification (see rows 7, 8), similar to those combination based baselines (e.g. DCNN, InceptionNet+LSTM, TFN), these variants of our AnaNet are still dominant, further demonstrating the superiority of Faster RCNN as the image encoder for depicting the local details of the image in this task.

2) Module Bias for Different Types of Data: In this paper, we introduce a new classification system of image-text correlation, and design a two-stream framework according to our definition, expecting it to be good at identifying such kinds of data. In this section, we further explore the preferences of the two core modules of our model for different categories of data. Detailed comparison results are shown in Table V. F_{exr} , F_{imr} and F_{irr} denote the F1 scores of explicit relevant, implicit relevant, irrelevant classes respectively. It can be seen that the two modules have the same ability to identify explicit relevant samples and irrelevant samples. But, for the implicit correlation samples, Association Net shows strong dominance, reaching 43.65% in F_{imr} , far surpassing Alignment Net's

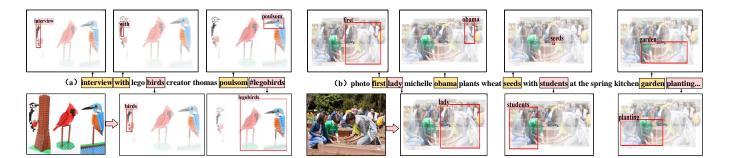


Fig. 3. Visualization of the attended image regions with respect to related word in different sub-figures. We use different regional brightness to represent the attentional weight of both region and word estimated by our interactive attention layer, and outline the maximum attentional region of each word in red.

TABLE VI Ablation Study: Different Feature Fusions of Association Net

Variant	Acc	F1	△Acc	△F1
$1 \ f_{inv}^{image}, f_{inv}^{text}$ (AnANet Used)	84.84	83.84	-	-
$\frac{1}{2} f_{var}^{image}, f_{var}^{text}$	84.02	83.49	-0.82	-0.35
$3 f^{image}, f^{text}$	83.92	82.84	-0.92	-1.00
$4 f_{var}^{image}, f_{var}^{text}, f_{var}^{image}, f_{var}^{text}$	84.02	83.38	-0.82	-0.46
$5 \ f_{inv}^{image}, f_{inv}^{text}, f^{image}, f^{text}$	84.13	83.36	-0.71	-0.48
$6 f_{var}^{image}, f_{inv}^{image}, f_{inv}^{text}, f_{var}^{text}$	83.32	83.05	-1.52	-0.79
$7 \ f_{var}^{image}, f_{inv}^{image}, f_{inv}^{text}, f_{var}^{text}, f^{image}, f^{text}$	83.42	82.91	-1.42	-0.93

37.71%. Being good at identifying this kind of implicit relevant sample is an important factor of our model's superiority over other kinds of baseline methods.

3) Different Feature Fusions of Association Net: We also conduct the ablation studies on different features used in Association Net to evaluate the effectiveness of our commonality strengthen operation. The detailed results of different feature combination strategies are shown in Table VI. In various permutations and combinations of all features, only using the shared invariant features $f_{inv}^{image}, f_{inv}^{text}$ achieves the best performance (see row 1). It is obvious that using our decomposed invariant or variant features for multimodal feature fusion is more effective than directly using the original image and text features (see rows 1, 2, 3). It also works well when we combine these decomposed features with the original features (see rows 3, 4, 5). The results indicate that our Association Net has the ability to learn more distinguishable and high-quality multimodal features. Interestingly, when we combine invariant features and variant features for feature fusion (see row 6), it seems that they are mutually exclusive and even better to directly use original image and text features. Moreover, it is not that the more features used, the stronger the representation ability of the model will be, which not only confuses the model but also increases its complexity (see rows 4, 5, 6, 7).

VI. VISUALIZATION AND ANALYSIS

A. Visualizing Attention

Our proposed Alignment Net utilizes interactive attention mechanism to explicitly calculate the local relevance between

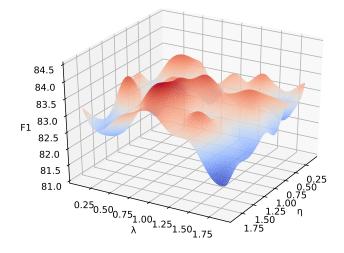


Fig. 4. Different performance with different values of λ and η .

image regions and text words. By visualizing the attention component learned by our model on two relevant multimodal tweets in Fig.3, we are able to provide the interpretability of our Alignment Net. We can observe that the words "birds", "legobirds", "lady", "students" and "planning" receive strong attention on the relatively precise locations, but other words, like "interview", "with", "poulsom", "first", "obama", "seed", "garden" etc., are less focused in sub-figures. The results show that our proposed attention mechanism works well for crossmodal correlation classification via providing the explainable reasoning cross-modal alignments between image regions and text words.

B. Hyperparameter Analysis

In our experiments, we empirically set and fix the hyperparameters λ to 0.1 and η to 2 (see Eq.22), and then tune the hyperparameters λ and η (see Eq.18) using grid search. Fig.4 shows the different performance with different values of λ and η . We can see that our AnANet achieves the best performance when $\lambda=0.7$ and $\eta=1.3$.

VII. CONCLUSIONS

In this paper, we define a new classification system (i.e. explicit relevant, implicit relevant, and irrelevant) for image-

text correlation based on implicit association and explicit alignment. According to our proposed definition, we construct a two-stream method, AnANet, to model the essence of crossmodal correlation. Our model implicitly represents the global discrepancy and commonality between image and text and explicitly captures the cross-modal local relevance. The experimental results on our constructed new dataset demonstrate the effectiveness of our model.

REFERENCES

- [1] M. Zeppelzauer and D. Schopfhauser, "Multimodal classification of events in social media," *Image and Vision Computing*, pp. 45–56, 2016.
- [2] I. Sorodoc, J. H. Lau, N. Aletras, and T. Baldwin, "Multimodal topic labelling," in *Proceedings of the EACL*, 2017, pp. 701–706.
- [3] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the EMNLP*, 2017, pp. 1103–1114.
- [4] Q. Zhang, J. Fu, X. Liu, and X. Huang, "Adaptive coattention network for named entity recognition in tweets." in *Proceedings of the AAAI*, 2018, pp. 5674–5681.
- [5] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the CVPR*, 2019, pp. 10394–10403.
- [6] R. Ma, X. Qiu, Q. Zhang, X. Hu, Y.-G. Jiang, and X. Huang, "Co-attention memory network for multimodal microblog's hashtag recommendation," *IEEE Transactions on Knowledge and Data Engineering*, 2019.
- [7] S. McCloud and A. Manning, "Understanding comics: The invisible art," *IEEE Transactions on Professional Communications*, vol. 41, no. 1, pp. 66–69, 1998.
- [8] E. E. Marsh and M. D. White, "A taxonomy of relationships between images and text," *Journal of Documentation*, 2003.
- [9] R. Martinec and A. Salway, "A system for image–text relations in new (and old) media," *Visual communication*, vol. 4, no. 3, pp. 337–371, 2005.
- [10] T. Chen, D. Lu, M.-Y. Kan, and P. Cui, "Understanding and classifying image tweets," in *Proceedings of the ACM MM*, 2013, pp. 781–784.
- [11] Z. Wang, P. Cui, L. Xie, W. Zhu, Y. Rui, and S. Yang, "Bilateral correspondence model for words-and-pictures association in multimedia-rich microblogs," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 10, no. 4, pp. 1–21, 2014.
- [12] S. Wu, "A multimodal analysis of image-text relations in picture books," *Theory and Practice in Language Studies*, vol. 4, no. 7, p. 1415, 2014.
- [13] M. Jas and D. Parikh, "Image specificity," in *Proceedings* of the CVPR, 2015, pp. 2727–2736.
- [14] T. Chen, H. M. SalahEldeen, X. He, M.-Y. Kan, and D. Lu, "Velda: Relating an image tweet's text and images." in *Proceedings of the AAAI*, 2015, pp. 30–36.

- [15] C. A. Henning and R. Ewerth, "Estimating the information gap between textual and visual representations," in *Proceedings of the ICMR*, 2017, p. 14–22.
- [16] C. Otto, M. Springstein, A. Anand, and R. Ewerth, "Characterization and classification of semantic imagetext relations," *International Journal of Multimedia In*formation Retrieval, vol. 9, no. 1, pp. 31–45, 2020.
- [17] M. Zhang, R. Hwa, and A. Kovashka, "Equal but not the same: Understanding the implicit relationship between persuasive images and text," in *Proceedings of the BMVC*, 2018, pp. 1–14.
- [18] J. Kruk, J. Lubin, K. Sikka, X. Lin, D. Jurafsky, and A. Divakaran, "Integrating text and image: Determining multimodal document intent in instagram posts," in *In Proceedings of the EMNLP-IJCNLP*, 2019, pp. 4614– 4624.
- [19] A. Vempala and D. Preoţiuc-Pietro, "Categorizing and inferring the relationship between the text and image of twitter posts," in *Proceedings of the ACL*, 2019, pp. 2830–2840.
- [20] M. Alikhani, S. N. Chowdhury, G. de Melo, and M. Stone, "Cite: A corpus of image-text discourse relations," in *Proceedings of the NAACL*, 2019, pp. 570–575.
- [21] M. Alikhani, P. Sharma, S. Li, R. Soricut, and M. Stone, "Cross-modal coherence modeling for caption generation," in *Proceedings of the ACL*, 2020, pp. 6525–6535.
- [22] J. Zhu, H. Li, T. Liu, Y. Zhou, J. Zhang, and C. Zong, "Msmo: Multimodal summarization with multimodal output," in *Proceedings of the EMNLP*, 2018, pp. 4154– 4164.
- [23] N. Xu, W. Mao, and G. Chen, "A co-memory network for multimodal sentiment analysis," in *Proceedings of the SIGIR*, 2018, pp. 929–932.
- [24] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proceedings of the ACL*, 2020, pp. 3777–3786.
- [25] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, 2015.
- [26] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings* of the EMNLP, 2014.
- [27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the NAACL*, 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423
- [28] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the ECCV*, 2018.