

# AlignVE: Visual Entailment Recognition Based on Alignment Relations

Biwei Cao, Jiuxin Cao, Jie Gui, Jiayun Shen, Bo Liu, Lei He, Yuan Yan Tang, and James Tin-Yau Kwok

**Abstract**—Visual entailment (VE) is to recognize whether the semantics of a hypothesis text can be inferred from the given premise image, which is one special task among recent emerged vision and language understanding tasks. Currently, most of the existing VE approaches are derived from the methods of visual question answering. They recognize visual entailment by quantifying the similarity between the hypothesis and premise in the content semantic features from multi modalities. Such approaches, however, ignore the VE’s unique nature of relation inference between the premise and hypothesis. Therefore, in this paper, a new architecture called AlignVE is proposed to solve the visual entailment problem with a relation interaction method. It models the relation between the premise and hypothesis as an alignment matrix. Then it introduces a pooling operation to get feature vectors with a fixed size. Finally, it goes through the fully-connected layer and normalization layer to complete the classification. Experiments show that our alignment-based architecture reaches 72.45% accuracy on SNLI-VE dataset, outperforming previous content-based models under the same settings.

**Index Terms**—Computer vision, visual entailment, alignment relation.

## I. INTRODUCTION

**V**ISUAL entailment (VE) proposed by Xie et al. [1], [2] is a multi-modal inference task derived from the original single-modal textual entailment (TE) [3] in natural language

This work was supported by National Key R&D Project of China under Grant No.2021QY2102; National Natural Science Foundation of China under Grants No.62172089, No.61972087, No.62172090, No.62106045, No.62172458; Natural Science Foundation of Jiangsu Province under Grant No.BK20191258; Jiangsu Provincial Key Laboratory of Computer Networking Technology; Jiangsu Provincial Key Laboratory of Network and Information Security under Grant No. BM2003201; Key Laboratory of Computer Network and Information Integration of Ministry of Education of China under Grant No. 93K-9; CAAI-Huawei MindSpore Open Fund; Alibaba Group through Alibaba Innovative Research Program.

Biwei Cao, Jiuxin Cao, Jie Gui and Jiayun Shen are with the School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China, and also with the Key Laboratory of Computer Network and Information of Ministry of Education of China, Nanjing 211189, China. Jiuxin Cao and Jie Gui are also with Purple Mountain Laboratories, Nanjing 210000, China. E-mail: {caobiwei, jx.cao, guijie, jyshen}@seu.edu.cn.

Bo Liu is with the School of Computer Science and Engineering, Southeast University, Nanjing 211189, China. E-mail: bliu@seu.edu.cn.

Lei He is with Information Engineering University, China and Purple Mountain Laboratories, Nanjing 210000, China. E-mail: helei@pmlabs.com.cn.

Yuan Yan Tang is with the Department of Computer and Information Science, University of Macao, Macao 999078, China. E-mail: yuanyant@gmail.com.

James Tin-Yau Kwok is with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong 999077, China. E-mail: jamesk@cse.ust.hk.

(Corresponding author: Jiuxin Cao, Jie Gui)


processing research and visual question answering (VQA) [4]. In entailment recognition, given a premise  $P$  and a hypothesis  $H$ , the system outputs the *entailment* if  $H$  can be concluded from  $P$ . The result is a *contradiction* if  $H$  contradicts  $P$  and otherwise it is *neutral*, which means  $P$  and  $H$  are not related. For example, in Table I, when the hypothesis text is “Two women are holding packages.”, the entity of “two women” and the action of “holding packages” are able to be matched with the contents of the premise image. Therefore, the hypothesis is included in premise and the entailment relation result is *entailment*. In the hypothesis sentence “The sisters are hugging goodbye while holding to go packages after just eating lunch.”, the conclusion of “sisters” is neither able to be drawn from this premise image nor be verified to contradict the image content. The entailment relation of this pair of premise and hypothesis is thus *neutral*. When the hypothesis becomes “The men are fighting outside a deli.”, the verb “fighting” is contradicted with the action of the figures in the premise image, so that the result of the entailment relation is *contradiction*.

Recently, the majority of VE studies [1], [2], [5] are inspired by VQA. It is found that VQA includes a relative wide range of specific tasks, such as generating a sentence as an answer [6], [7], discriminating yes/no question [4], [8], selecting a word or phrase in predefined answers [4] and filling in the blank [9]. What is more, the answers in VQA datasets [4] mainly consist of objects, colors, categories, scenes, sports and even brands. Thus, it can be concluded that VQA is a broader multi-modal task compared to VE. Because of this, most of the VQA methods are universal multi-modal classification models, which extract and fuse multi-modal features to build enriched content semantic features. While for VE task, its goal is to only classify the relation between the premise and hypothesis into three classes. Although the recent VQA methods are universal and can be easily applied to VE, they are not designed specifically for VE. The works which solve VE tasks by simply adopting VQA methods ignore the characteristic of VE — recognizing relations. On the basis of this observation, we believe that modeling the relation between premise and hypothesis should be meaningful and essential to VE studies.

The main contributions of our work can be summarised as follows.

Firstly, based on the difference between a VQA task and a VE task, an alignment-relation-based visual entailment architecture AlignVE is proposed, particularly considering the characteristics of the VE task. It contains three parts, namely visual feature extraction module, text feature extraction module and the alignment-based classifier construction module. The first two modules aim to extract visual and text features

TABLE I  
A VISUAL ENTAILMENT EXAMPLE WITH A PREMISE IMAGE AND THREE KINDS OF ENTAILMENT RELATION.

	Two women are holding packages.	<i>Entailment</i>
	The sisters are hugging goodbye while holding to go packages after just eating lunch.	<i>Neutral</i>
	The men are fighting outside a deli.	<i>Contradiction</i>
Premise	Hypothesis	Entailment Relation

from premise and hypothesis. The alignment-based classifier construction module attempts to build the relation interaction between these features and utilize the obtained alignment relation to complete the classification task.

Secondly, the relation alignment method is firstly introduced in this field for the enhancement of recognizing the interaction between the premise image and hypothesis text, specifically modeling the entailment relation for VE classification.

Last but not least, comprehensive experiments are conducted with SNLI-VE dataset to re-evaluate existing VE models and evaluate both migrated TE models and our proposed model. The results indicate that our AlignVE architecture outperforms the previous methods and still keeps the simplicity.

## II. RELATED WORK

Text entailment (TE), a classic natural language processing task, is the predecessor task of VE and has developed greatly in entailment recognition. Derived from TE, VE is defined as a multi-modal classification task, which is closely related to VQA, a typical multi-modal classification task that has been extensively studied in recent years. The VQA study history reflects the development of multi-modal classification methods. Both of TE and VQA have a close relation and inspiration to VE.

### A. Textual Entailment

TE is proposed as the PASCAL recognizing textual entailment challenge [3] in 2005. In early studies, most of them are based on similarity measurement [10]–[12], alignment algorithms [13], [14], machine learning [15] and logic inference [16] methods.

Since the large dataset SNLI [17] was published, deep-learning based methods are used in the majority of recent studies. RNN like LSTM is applied to TE [17], combined with attention mechanism [18], [19]. More LSTM-based methods have been developed like match-LSTM (mLSTM) [20], re-read LSTM (rLSTM) [21] and Enhanced Sequential Inference Model (ESIM) [22]. CNN-based methods have also been developed for TE, e.g. TBCNN (Tree-based CNN) [23], [24]. The Interactive Inference Network (IIN) [25] explicitly models the relation between the premise and hypothesis as an interaction space and utilizes a CNN to extract features in it. The logic-based methods also have progresses recently [26]. The TE studies analyze the entailment recognition task and propose specific methods designed for this, which inspire VE studies.

### B. Visual Question Answering

The VQA is formally defined as providing an accurate natural language answer to a given image and a natural language question about the image [4]. Early studies [7], [26] recognized VQA as a natural language generation task and then handled it with the encoder-decoder framework. However, recent studies mostly model VQA as a classification task which selects an answer from a predefined corpus. The typical paradigm of classification approach is that the VQA system extracts image features with CNN or object detector, then encodes the question with RNN and finally implements several mechanisms to do feature fusion and classification. To improve the quality of multi-modal feature extraction and fusion, many methods based on the attention mechanism have been proposed for VQA, such as region selection attention [27], co-attention [28], stacked attention [29], high-order attention [30], bottom-up and top-down attention [31], dense co-attention [32], modular co-attention [33], dynamic fusion with intra-and inter-modality attention [34] and multi-grained attention [35]. In conclusion, the design of VQA methods becomes more complicated together with the trend of fine-grained, multi-modal and multi-hop attention, thus enriching the representation of image and question features. Also, the model CLIP [36] utilizes the natural language supervision information as the training signal to learn the visual feature, and takes the dot product method to build the interaction between image and text domain in multi-modal embedding space. These methods are analyses of images and texts at the macro level, and lack of a fine-grained analysis into the entailment relation between images and texts.

### C. Visual Entailment

The current number of VE studies is limited [1], [2], [5], [37]. Most methods are deep-learning based. Xie et al. [1], [2] propose the Explainable VE (EVE) architecture, which is based on Bottom-Up/Top-down architecture [31], the first place of the 2017 VQA challenge. Their intuition is that both of VE and VQA can be modeled as multi-modal classification tasks. By modifying typical VQA models for VE task, EVE can benefit from proved multi-modality feature fusion and high classification performance of VQA models. Do et al. [5] re-annotate SNLI-VE dataset, re-evaluate existing models and introduce a new task e-SNLI-VE 2.0 that requires explanation sentence generation for VE recognition. Chen et al. [38]

TABLE II  
RELATIONSHIP AMONG TE, VQA AND VE IN THEIR GOALS, RESULT CATEGORIES AND METHODS.

	TE	VQA	VE
Goal	Recognize the relationship between sentences.	Answer visual questions.	Recognize the relationship between the image and sentence.
Category	is entailment or not (early stage), entailment / neutral / contradiction	Yes/No, object, color, number, position, multiple choices...	entailment / neutral / contradiction
Method	manually crafted features (early stage), logic inference, deep learning classification	deep learning generation, deep learning classification	deep learning classification, logic inference

propose a universal pretrained image-text representation model UNITER and include VE as one of their downstream tasks.

Besides, Suzuki et al. [37] propose a logic-based VE system. This system translates an image into a scene graph or first-order logic (FOL) structure and at the same time, parses sentences into FOL formulae. Then, this system applies Prover9<sup>1</sup> as the inference engine to conduct theorem proofs. However, this method is slow and can only recognize VE as a binary classification task since it relies on a timeout (10s) to decide whether a hypothesis is entailed or not.

Therefore, we can draw a conclusion for the relationship among TE, VQA and VE in Table II.

In all, most VE studies lack analysis on the characteristics of VE, thus lacking the study on the interaction relation between the premise and hypothesis. New methods are needed to fill up this gap. Therefore, our AlignVE architecture is designed to model the premise-hypothesis relation as an alignment matrix and recognizes VE based on it. It is designed specifically for VE and maintains the simplicity of the whole structure. To the best of our knowledge, there is no relation-based VE study currently. It can be viewed as a significant starting point of relation-based VE methods.

### III. THE ALIGNVE ARCHITECTURE

The VE task is defined as a multi-modal task which recognizes the relation between a premise image and hypothesis text. Formally, given a premise image  $P_{image}$  and a hypothesis text  $H_{text}$ , the goal of VE is to determine if  $H_{text}$  can be concluded from  $P_{image}$  and to classify this relation into three classes: *Entailment*, *Contradiction* and *Neutral*.

*Entailment* holds if the evidence in  $P_{image}$  is enough to conclude that  $H_{text}$  is true.

*Contradiction* holds if the evidence in  $P_{image}$  is enough to conclude that  $H_{text}$  is false.

*Neutral* holds otherwise, implying the evidence in  $P_{image}$  is insufficient to draw a conclusion about  $H_{text}$ .

The overall architecture of AlignVE is shown in Figure 1 which mainly contains three modules as follows.

- The visual feature extraction module, which uses a CNN to extract grid features or an object detector to extract RoI features, and then encodes the visual features with an AttEnc.
- The text feature extraction module, which uses GloVe [39] as word embedding and encodes the text features with AttEnc.

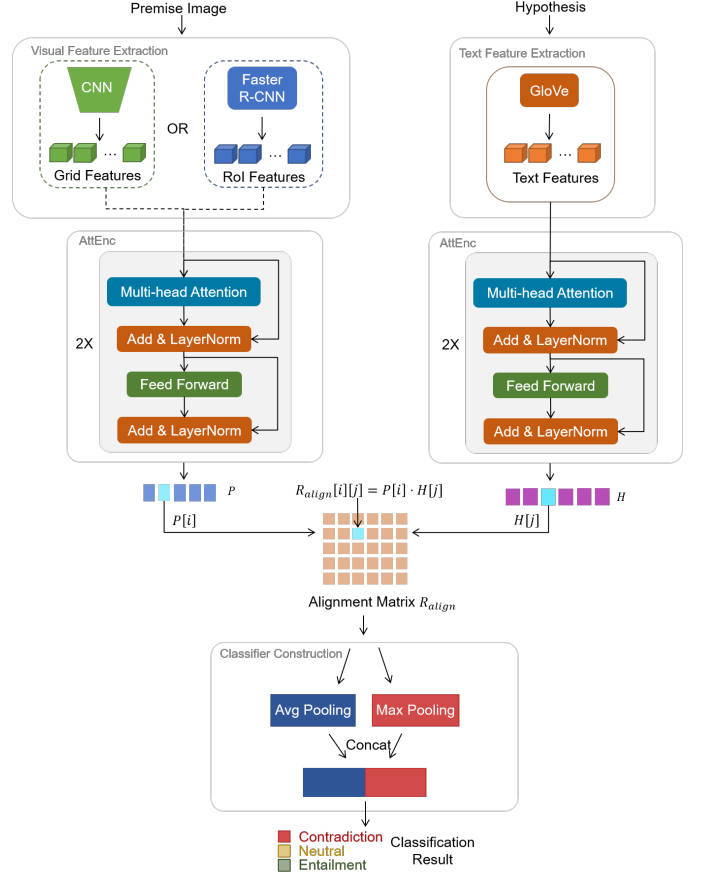


Fig. 1. The Architecture of AlignVE. The main components of the AlignVE are visual extraction module, text feature extraction module and alignment-based classifier construction module.

- The alignment-based classifier construction module, which is to calculate the alignment matrix between the premise and hypothesis, and then apply adaptive pooling to get a fixed-length feature vector for classification.

#### A. Visual Feature Extraction Module

The visual feature extraction module takes an image  $P_{image}$  as the input premise and outputs a sequence of vectors  $P$  to represent the premise image. Two approaches are taken to extract image features. One inputs images to a pretrained CNN and takes the feature map as image grid features, and the other one inputs images to a pretrained object detector (Faster R-CNN) and takes the RoI features as image features. Both approaches are widely applied in visual understanding fields and several related works [31], [40] make a detailed

<sup>1</sup><http://www.cs.unm.edu/mccune/prover9/>

comparison and analysis between the grid and RoI features. This process can be represented as

$$F_p = CNN(P_{image}) \quad (1)$$

or

$$F_p = RCNN(P_{image}). \quad (2)$$

In our study,  $F_p \in \mathbb{R}^{m \times d_p}$  where  $m$  is fixed to 36 for both grid or RoI features and  $d_p$  is 2048 according to the design of ResNet [41].

Then, the extracted sequence of visual features is encoded by a self-attention encoder *AttEnc* from Transformer [42],

$$P = AttEnc(F_p). \quad (3)$$

Here, the dimensionality of each feature vector in  $P \in \mathbb{R}^{m \times d}$  is transformed to a unified feature dimensionality  $d$ .

1) *Grid Features*: Specifically, for image grid features, a pretrained ResNet101 [41] is used as CNN and the output feature map from last layer of conv5\_x before pooling is taken as grid features. Due to the varying spatial sizes of different images, the output feature maps are in different sizes. The bilinear interpolation is selected to resize the feature map with variable spatial size to the one of fixed size  $6 \times 6$ .

2) *RoI Features*: For image RoI features, a pretrained Faster R-CNN [43] is taken as object detector to detect salient objects and RoI features  $F_p \in \mathbb{R}^{m \times d_p}$  is obtained by mapping RoI boxes on the feature map of backbone with RoIAlign. Top-36 RoI features are selected by ranking with the RoI's bounding box score. For the cases of an image containing RoIs less than 36, zero-vector is padded to it.

3) *Self-Attention Encoder*: To achieve better representation of an image, the self-attention encoder proposed by Vaswani et al. [42] in Transformer architecture is applied instead of using the grid/RoI features directly. We call this AttEnc for short. AttEnc is powerful and faster than RNN like LSTM [44] or GRU [45] because its calculation flow is designed in a parallel way, not using iterations. It is applied to enrich the image features with their context information and is responsible to transform the input dimensionality to a unified feature dimensionality  $d$ .

For an AttEnc, it takes a sequence of features  $F_p$  as input:

$$F_{enc} = AttEnc(F_p), \quad (4)$$

where  $F_p \in \mathbb{R}^{m \times d_p}$ .

In a basic building block of AttEnc, the scale-dot product (SDP) attention [42] is taken for the input features,

$$Att_{SDP}(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_p}})V. \quad (5)$$

Based on SDP attention, the multi-head attention [42] is designed to pay attention to information from different representation subspaces at different positions,

$$Att_{MH}(Q, K, V) = concat(head_1, \dots, head_h)W^O, \quad (6)$$

where  $W^O$  is the projection parameter matrix. In self-attention mechanism, the query  $Q$ , key  $K$  and value  $V$  are input features

with different linear transformations as (7). For the  $i$ -th head, it comes from an SDP attention in a representation subspace,

$$head_i = Att_{SDP}(F_p W_i^Q, F_p W_i^K, F_p W_i^V), \quad (7)$$

where  $W_i^Q$ ,  $W_i^K$  and  $W_i^V$  are projection parameter metrics. The attended features  $F_{att} \in \mathbb{R}^{m \times d_p}$ , which are obtained from the self-attention mechanism, are then passed to a feed-forward network with residual connection [41] and layer normalization [46] as

$$F_{enc} = LN(ReLU(LN(F_{att} + F_p)W_f + b_f) + F_{att}). \quad (8)$$

The output of AttEnc is  $F_{enc} \in \mathbb{R}^{m \times d}$ , where  $m$  is the length of input sequence of features and  $d$  is the target feature dimensionality.

### B. Text Feature Extraction Module

The text feature extraction module takes a sequence of text  $T$  as the input hypothesis and outputs a sequence of vectors  $H$  to represent the hypothesis text. At the beginning, the text is tokenized and the GloVe [39] is used as word embedding to represent each token,

$$F_h = GloVe(T). \quad (9)$$

For an  $n$ -token text, the primary representation sequence  $F_h \in \mathbb{R}^{n \times d_h}$  means each token of the text is represented by a vector whose size is  $d_h$ .

Next, with the aim of injecting the relative position information into the text sequence, the positional encoding  $PE$  is added to text,

$$F_h = F_h + PE. \quad (10)$$

Specifically, the position encoding is referenced from the sine and cosine functions in Transformer [42],

$$PE_{(pos, 2i)} = \sin(\frac{pos}{10000^{\frac{2i}{d}}}), \quad (11)$$

$$PE_{(pos, 2i+1)} = \cos(\frac{pos}{10000^{\frac{2i}{d}}}), \quad (12)$$

where  $pos$  is the position and  $i$  is the dimension. In the end, the self-attention encoder is utilized to encode the word representations with their context information and transform the input dimensionality to a unified feature dimensionality  $d$ ,

$$H = AttEnc(F_h). \quad (13)$$

Here, the  $H \in \mathbb{R}^{n \times d}$  is transformed in a unified feature dimensionality  $d$ .

### C. Alignment Relations

The alignment relations are modeled as an alignment matrix  $R_{align}$ . Given the premise image features  $P$  from the visual feature extraction module and hypothesis text features  $H$  from the text feature extraction module, the alignment value of the  $i$ -th premise vector  $P[i]$  and the  $j$ -th hypothesis vector  $H[j]$  is the dot-product of both  $d$ -dim vector,

$$R_{align}[i][j] = P[i] \cdot H[j], \quad (14)$$

where  $\cdot$  represents vector dot-product.

TABLE III  
SNLI-VE DATASET DISTRIBUTION IN IMAGE AND THREE KINDS OF ENTAILMENT RELATIONS AS WELL AS THE SPLITTING WAY.

	Training	Validation	Testing	Total
Image	29,783	1,000	1,000	<b>31,783</b>
#Entailment	176,932	5,959	5,973	188,864
#Neutral	176,045	5,960	5,964	187,969
#Contradiction	176,550	5,939	5,964	188,453
#Total	529,527	17,858	17,901	<b>565,286</b>

TABLE IV  
SNLI-VE-2.0 DATASET DISTRIBUTION IN IMAGE AND THREE KINDS OF ENTAILMENT RELATIONS AS WELL AS THE SPLITTING WAY.

	Training	Validation	Testing	Total
Image	29,783	1,000	1,000	<b>31,783</b>
#Entailment	131,023	5,254	5,218	141,495
#Neutral	125,902	3,442	3,801	133,145
#Contradiction	144,792	5,643	5,721	156,156
#Total	401,717	14,339	14,740	<b>430,796</b>

The alignment matrix can be written in a matrix computation way,

$$R_{align} = PH^T, \quad (15)$$

where  $R_{align} \in \mathbb{R}^{m \times n}$  is the alignment matrix representing the alignment relations between  $m$  premise image features and  $n$  hypothesis text features.

#### D. Classifier Construction

Considering the shape of the alignment matrix  $m \times n$  is variable for different premise-hypothesis pairs, it has to be down-sampled into a fixed size for the further classification task. Therefore, we choose to use the adaptive pooling method, which is able to obtain a specified output size no matter what the size of the input is, on this 2-D matrix. Here, we use the concatenation of adaptive average pooling and adaptive max pooling, with the purpose of enriching the feature information. To be more specific, given the 2-D alignment matrix, we do the adaptive average pooling and max pooling operations concurrently to achieve two fixed-size feature vectors respectively. Each pooling method pools the alignment matrix to a 150-D vector, and then the two pooled 150-D vectors are concatenated to form a 300-D feature vector, which is used as the input of the fully-connected layer for the classification. This process is formulated in Eq. (16) - (19).

$$V_{avg} = avgpool(R_{align}), \quad (16)$$

$$V_{max} = maxpool(R_{align}), \quad (17)$$

$$V = concat(V_{avg}, V_{max}), \quad (18)$$

$$S = softmax(VW_v + b_v), \quad (19)$$

where  $avgpool$  and  $maxpool$  are adaptive average pooling and adaptive max pooling respectively.  $S \in \mathbb{R}^C$  is the predicted score for VE classification,  $C$  is the number of classes and both  $W_v$  and  $b_v$  are parameters learned through training process.

## IV. EXPERIMENTS

### A. Dataset

Models are evaluated on SNLI-VE<sup>2</sup> [1], [2] dataset and SNLI-VE-2.0<sup>3</sup> [5]. The SNLI-VE dataset proposed by Xie et al. is a VE dataset created by finding the corresponding image in Flickr30K dataset for each text premise-hypothesis pair in SNLI through the annotation of SNLI dataset. As shown in Table III, 31,783 images and 565,286 premise-hypothesis pairs are labeled for 3 classes as *entailment*, *neutral* and *contradiction*. We use the same dataset SNLI-VE published by Xie et al. for fair comparison. The SNLI-VE-2.0 dataset is built by Do et al. based on the SNLI-VE dataset with the improvement on the classification accuracy by re-annotating the incorrect data in *neutral* class of validation set and test set. This dataset distribution is shown in Table IV, with 430,796 premise-hypothesis pairs for the 3-class classification. Do et al. [5] mention that they try to see the difference in performance on the corrected validation and test sets by reproduction of BUTD model. However, the statistics of Do et al.'s work [5] shows that the corrected dataset has no effect on the BUTD model performance by a drop of test accuracy from SNLI-VE dataset to SNLI-VE-2.0 dataset. Since Do et al. have completed related experiments using SNLI-VE-2.0, we also include this dataset and implement models on it for a better comparison.

### B. Baselines

We select several models in recent VE studies [1], [2], [5] as baseline models and migrate several TE architectures [21], [25] to the VE task for a better comparison.

1) *TD/BUTD*: This Bottom-Up Top-Down (BUTD) attention architecture is originally proposed by Anderson et al. [31] for VQA and image caption. The ‘bottom-up’ attention stands for RoI feature from object detection in practice since the object detection can be considered as a hard attention on images. For convenience, the grid-feature-based Top-Down attention is named as TD and the RoI feature based Bottom-Up

<sup>2</sup><https://github.com/necla-ml/SNLI-VE>

<sup>3</sup><https://github.com/maximek3/e-ViL>

TABLE V  
EXPERIMENT RESULTS ON SNLI-VE DATASET PROVIDED BY PREVIOUS PAPERS [1], [2] AND THIS PAPER’S IMPLEMENTATION, INCLUDING THE RE-IMPLEMENTED MODELS, TRANSFERRED MODELS AND OUR PROPOSED MODEL ALIGNVE.

Model Type	Model	Validation Accuracy (%)	Test Accuracy (%)
Original Model	TD	70.53	70.30
	BUTD	69.34	68.90
	EVE-Image	71.56	71.16
	EVE-RoI	70.81	70.47
Re-implemented Model	TD	71.00	71.04
	BUTD	70.52	70.97
	EVE-Image	71.52	71.43
	EVE-RoI	70.43	70.43
Transferred Model	rLSTM-Grid	70.92	71.25
	rLSTM-RoI	71.24	71.26
	IIN-Grid	71.02	71.34
	IIN-RoI	71.33	71.30
Our Proposed Model	AlignVE-Grid	<b>72.31</b>	<b>72.45</b>
	AlignVE-RoI	72.02	72.20

TABLE VI  
TEST ACCURAY (%) OF BUTD MODEL AND OUR PROPOSED MODEL ALIGNVE ON SNLI-VE AND SNLI-VE-2.0 DATASET.

Model	SNLI-VE dataset	SNLI-VE-2.0 dataset
BUTD	70.97	69.26
AlignVE-Grid	72.45	72.67

TABLE VII  
ABLATION EXPERIMENT RESULTS OF ATTENC WITH MLP AND GRU IN ACCURACY (%) ON SNLI-VE DATASET.

Model (replace AttEnc)	Before ablation		After ablation	
	Validation	Test	Validation	Test
rLSTM-Grid	70.92	71.25	69.65	69.34
rLSTM-RoI	71.24	71.26	69.73	70.07
IIN-Grid	71.02	71.34	69.99	70.30
IIN-RoI	71.33	71.30	69.55	70.11
AlignVE-Grid	<b>72.31</b>	<b>72.45</b>	<b>71.25</b>	70.63
AlignVE-RoI	72.02	72.20	70.73	<b>71.01</b>

TABLE VIII  
ABLATION EXPERIMENT RESULTS OF ALIGNMENT MATRIX WITH A CO-ATTENTION LAYER IN ACCURACY (%) ON SNLI-VE DATASET.

Model (replace Alignment Matrix)	Before ablation		After ablation	
	Validation	Test	Validation	Test
AlignVE-Grid	<b>72.31</b>	<b>72.45</b>	66.3	66.01
AlignVE-RoI	72.02	72.20	<b>66.64</b>	<b>66.78</b>

TABLE IX  
RESNET SERIES EXPERIMENT RESULTS OF THE ALIGNVE-GRID MODEL IN ACCURACY (%) ON SNLI-VE DATASET.

Model	Validation	Test
ResNet-18	70.27	70.25
ResNet-34	70.49	70.48
ResNet-50	70.46	70.79
ResNet-101	<b>72.31</b>	<b>72.45</b>
ResNet-152	71.02	71.3

Top-Down attention as BUTD. Xie et al. [1], [2] evaluate both TD and BUTD as baseline models. Do et al. [5] re-evaluate BUTD only. In our experiment, we implement this architecture as one of the baselines and re-evaluated both TD and BUTD in the experiment environment.

2) *EVE*: The Explainable Visual Entailment (EVE) model is a VE architecture proposed by Xie et al. [1] by modifying the BUTD Attention model. Two branches are composed of EVE architecture: the EVE-Image and EVE-RoI. The EVE-

Image takes grid features as image features while the EVE-RoI takes RoI features as image features. Both EVE-Image and EVE-RoI are re-evaluated by using our experiment settings.

3) *rLSTM*: The re-read LSTM (rLSTM) architecture is originally proposed by Sha et al. [21] as a LSTM model specifically designed for TE. The rLSTM architecture is designed to enrich the content feature of hypotheses, by re-reading the premise while reading the hypothesis in LSTM iterations. In experiment, we migrate the rLSTM to VE by replacing the original text-only input with the our image model and text model.

4) *IIN*: The Interactive Inference Network (IIN) architecture is originally proposed by Gong et al. [25] for TE. The IIN architecture is a typical relation-based entailment recognition architecture which models the interaction relation between each pair of the premise token and hypothesis token by building the interaction space. IIN is altered to apply to VE by using our image model and text model as embedding layer and encoding layer of IIN architecture. The interaction



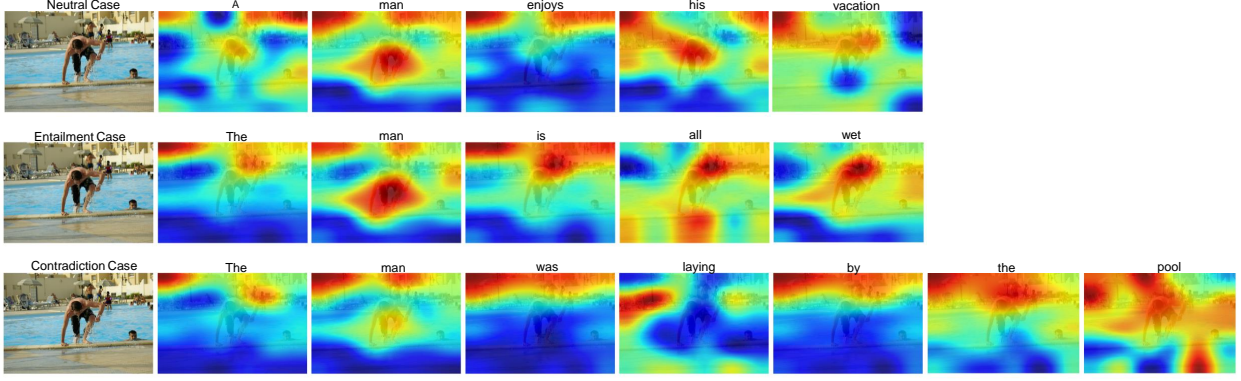


Fig. 2. Alignment matrix visualization of AlignVE in the “neutral”, “entailment” and “contradiction” cases. Premise images are on the leftmost part and the visual attention images according to each word of the hypotheses are on the right. In a visual attention image, the color from blue to red represents the alignment relation from weak to strong.

layer is implemented as element-wise product. The feature extraction layer is implemented as a plain ResNet and the last fully connected layer of ResNet is directly used as prediction output layer.

### C. Experiment Configurations

The number of image features in each premise image is fixed to  $m = 36$  for both grid features and RoI features. The text is tokenized by spaCy<sup>4</sup> and the GloVe.840B.300d<sup>5</sup> is used as word embedding which represents each token as a 300-d vector and is frozen during training.

For dataset SNLI-VE, as we can see in Table V, Xie et al. provide the experiment result of TD/BUTD and the EVE model they proposed.

The experiment is implemented with PyTorch [47]. Models are trained with cross entropy loss optimized by SGD with momentum 0.9 or Adam [48]. The max training epoch is set to 100 and the batch size is 64. The learning rate is set as  $1e-4$  by default. We apply a learning rate decay strategy that the learning rate is decayed by a coefficient 0.1 whenever the loss on validation set does not decrease for 2 epochs. The training dumps a checkpoint after an epoch and the checkpoint with best validation set accuracy is evaluated for test set accuracy.

For TD/BUTD, we re-implement the model following the existing implementation<sup>6</sup> and set major configurations according to Do et al. [5]. For EVE, we implement the model and set major configurations according to the paper [2] since no code has been released yet. The migrated rLSTM, IIN and our AlignVE are implemented with the same image model and text model. The configurations of AttEnc in them are tuned to 6 heads and 2 layers.

To control randomness and enhance experiment reproducibility, the random seeds of both Python and PyTorch are set to 12345 and the cuDNN deterministic mode is turned on.

### D. Results and Analyses

1) *Comparative Experiment:* The experiment results on SNLI-VE dataset are shown in Table V. The results reported by Xie et al. [2] are also included for a better comparison.

We realize the fairness of experiments by controlling the variables of data processing and feature extraction sections, thus ensuring the experimental rigor. In experiment, the test accuracy of TD and BUTD are 71.04% and 70.97%, higher than Xie et al.’s work by 0.74% and 2.07% separately. The test accuracy of EVE-Image and EVE-RoI are 71.43% and 70.43% respectively. Compared with Xie et al.’s results, our EVE-Image is higher by 0.27% and our EVE-RoI is lower by 0.04%. In all, the re-evaluation experiments generally reproduce the results reported by Xie et al. It is acceptable to have some differences within a reasonable range because of the differences in data preprocessing, feature extraction, model and training configurations, randomness control and so on. Thus, our re-implementation of previous experiments meets the expectation and the accuracy is higher in general, indicating it is reasonable to use the reproduced experimental results as the benchmark model for the comparative experiment.

The rLSTM and IIN are selected as typical examples of content-based and relation-based entailment recognition architectures respectively. With grid image features, the test accuracy of our migrated rLSTM and IIN are 71.25% and 71.34%, both are slightly lower than EVE-Image and higher than TD. With RoI image features, the test accuracy of rLSTM and IIN are 71.26% and 71.30%, both higher than those of BUTD and EVE-RoI. As for the migrated rLSTM and IIN, IIN performs better than rLSTM by a very slight margin.

Our AlignVE achieves 72.45% with grid image features and 72.20% with RoI image features. Compared to the results of models in previous VE researches, our AlignVE outperforms TD by 1.41%, BUTD by 1.23%, EVE-Image by 1.02% and EVE-RoI by 1.77%. Compared with the migrated IIN, our AlignVE has the same embedding layer and encoding layer, but the interaction layer, feature extraction layer and output

<sup>4</sup><https://spacy.io/>

<sup>5</sup><http://nlp.stanford.edu/data/glove.840B.300d.zip>

<sup>6</sup><https://github.com/claudiogreco/coling18-gte>

layer are simpler. Specifically, our AlignVE models the relation between the premise-hypothesis pair as an alignment matrix rather than an interaction space. Additionally, the feature extraction layer and output layer of our AlignVE are simply implemented with a pooling and a fully connected layer instead of a deep CNN. Although the structure is simpler, the experiment results proves that AlignVE still outperforms migrated IIN by 1.11% and 0.90% with grid and RoI image features respectively.

The results of experiments on SNLI-VE-2.0 dataset is shown in Table VI. Our re-implementation on this dataset verifies that the corrected dataset has no significant effect on the model performance by a drop of BUTD model from 70.97% on SNLI-VE dataset to 69.26% on SNLI-VE-2.0 dataset. Furthermore, our model AlignVE also shows that the improved dataset has no explicit difference in model performance with a slight increase of AlignVE model from 72.45% to 72.67%. Therefore, the SNLI-VE-2.0 dataset may not be meaningful from the current experiment data. Recall that only the validation and test set are corrected for data in *neutral* class. The training set is needed to be corrected as well if a better model performance is required.

2) *Ablation Experiment*: We do the ablation experiment to show the effectiveness of the AttEnc applied in our study. To replace the AttEnc in the model, MLP is used in visual feature extraction module to achieve visual features and GRU [45] is taken in text feature extraction module to obtain textual features. In this ablation experiment, we keep the other settings the same as described in Experiment Configurations and do the implementation based on the SNLI-VE dataset. As shown in Table VII, the test accuracy of various models all drop by 1% approximately and the accuracy of the model AlignVE is still better than the accuracy of other compared models after the ablation, which proves AttEnc is effective and indicates that the proposed alignment-based VE architecture is meaningful as well.

What is more, to further demonstrate the model effectiveness, we replace the alignment matrix with a co-attention layer [28], where we take the text feature and image feature as inputs and send the output of the co-attention layer to the classification layer in the same way. This co-attention layer uses the parallel co-attention mechanism which generates image and text attention simultaneously. The parameters used are the same as what we use in AlignVE model and the dataset used is SNLI-VE. The results in Table VIII show that the test accuracy of our model has 5.42% and 6.44% higher than replacing with a co-attention layer by using grid image features and RoI image features respectively, which indicates that the alignment matrix is more effective compared to the co-attention layer.

3) *ResNet Series Experiment*: In this paper, we also explore the influence of applying different pre-training models (i.e. ResNet series [41]) on the AlignVE model (with grid image features) performance. The experiment results can be found in Table IX. From the table, it can be concluded that the performance of the AlignVE model with grid image features becomes higher when the feature dimensionality becomes larger. Meanwhile, with the feature dimensionality fixed, the

performance of the model increases at first as the layer number increases and then decreases. The possible reason is that the image dataset is relatively simple and small to make the features easy to identify. So the ResNet-101 is a better fit than ResNet-152, which proves that deeper is not necessarily the better for deep learning models as well.

4) *Visualization Analysis*: Our AlignVE model can also support the visualization analysis since the alignment matrix is 2-dimensional which can be expressed in the form of visual attention. Thus, by applying the alignment matrix into visualization, we can directly obtain and examine the cross-modal alignment between the premise and hypothesis the model has learned.

Considering each row of the alignment matrix corresponds to a section of a premise image and each column corresponds to a hypothesis word, we do the visualization analysis in the unit of column. The alignment value of every column is counted as the visual attention of each word and is painted on the premise image with the jet color mapping. In this case, the alignment can be seen through the mask rendering. The more red in the image means the stronger alignment and the blue indicates the weak alignment relation.

In Figure 2, it shows a picture with three different types of hypotheses which correspond to the three classification result *neutral*, *entailment* and *contradiction*. The semantic of the premise image is that a man is about to go out of a swimming pool. The first hypothesis is “A man enjoys his vacation”. It can be seen from this case that, the visual attention of “man” focuses on the middle of the image which is the area that indicates the man. Since the words “enjoys” and “vacation” are hardly to conclude from this image, the most of the scene corresponding to these two words are blue and green. Therefore, this case demonstrates the model’s judgement ability when there is no evidence to prove the hypothesis matches the premise image. The second text hypothesis is “The man is all wet”. From the visualization results, it can be discovered that the middle of the image is colored in red when the words come to “man” and “wet”. Therefore, the alignment results imply that the model is able to learn the alignments of the figure and environment between premise and hypothesis. Also, the model’s understanding is corresponding to the human intuition. In the third case, the hypothesis is “The man was laying by the pool”. From the visualization, the words “man” and “pool” correspond to the person and the swimming pool in the image, while the visual attention of words “laying” and “by” shows the weak alignment relation to the person’s action and the environment. In summary, the model has the capability of understanding the *contradiction* classification given a pair of a premise image and a hypothesis text.

## V. CONCLUSION

In this paper, we propose a new alignment-based visual entailment (AlignVE) model. With the intuition that entailment recognition tasks like VE should focus on understanding the relation between the premise and hypothesis, the AlignVE calculates the relation of the premise-hypothesis pair as an alignment matrix and recognizes VE based on it. Experiments



show that AlignVE outperforms the previous VE models and the migrated typical TE models on SNLI-VE dataset in the same experiment settings, indicating that AlignVE architecture is simple and powerful.

## REFERENCES

- [1] N. Xie, F. Lai, D. Doran, and A. Kadav, “Visual entailment task for visually-grounded language learning,” *ArXiv*, vol. abs/1811.10582, 2018.
- [2] —, “Visual entailment: A novel task for fine-grained image understanding,” *ArXiv*, vol. abs/1901.06706, 2019.
- [3] I. Dagan, O. Glickman, and B. Magnini, “The pascal recognising textual entailment challenge,” in *MLCW*, 2005.
- [4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, “Vqa: Visual question answering,” in *ICCV*, 2015.
- [5] V. Do, O.-M. Camburu, Z. Akata, and T. Lukasiewicz, “e-snli-ve-2.0: Corrected visual-textual entailment with natural language explanations,” *ArXiv*, vol. abs/2004.03744, 2020.
- [6] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, “Are you talking to a machine? dataset and methods for multilingual image question,” in *NIPS*, 2015.
- [7] M. Malinowski, M. Rohrbach, and M. Fritz, “Ask your neurons: A neural-based approach to answering questions about images,” *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, 2015.
- [8] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, “Yin and yang: Balancing and answering binary visual questions,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5014–5022, 2016.
- [9] L. Yu, E. Park, A. Berg, and T. L. Berg, “Visual madlibs: Fill in the blank description generation and question answering,” *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2461–2469, 2015.
- [10] O. Glickman, I. Dagan, and M. Koppel, “Web based probabilistic textual entailment,” 2005.
- [11] L. Romano, M. Kouylekov, I. Szepietor, I. Dagan, and A. Lavelli, “Investigating a generic paraphrase-based approach for relation extraction,” in *EACL*, 2006.
- [12] B. MacCartney, T. Grenager, M.-C. de Marneffe, D. M. Cer, and C. D. Manning, “Learning to recognize features of valid textual entailments,” in *NAACL*, 2006.
- [13] D. Roth and M. Sammons, “A unified representation and inference paradigm for natural language,” 2008.
- [14] M. Sammons, V. Vydiswaran, T. Vieira, N. Johri, M.-W. Chang, D. Goldwasser, V. Srikumar, G. Kundu, Y. Tu, K. Small, J. Rule, Q. Do, and D. Roth, “Relation alignment for textual entailment recognition,” *Theory and Applications of Categories*, 2009.
- [15] D. Inkpen, D. Kipp, and V. Nastase, “Machine learning experiments for textual entailment,” 2006.
- [16] C. D. Manning and B. MacCartney, “Natural language inference,” 2009.
- [17] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in *EMNLP*, 2015.
- [18] T. Rocktäschel, E. Grefenstette, K. Hermann, T. Kociský, and P. Blunsom, “Reasoning about entailment with neural attention,” *CoRR*, vol. abs/1509.06664, 2016.
- [19] Y. Liu, C. Sun, L. Lin, and X. Wang, “Learning natural language inference using bidirectional lstm model and inner-attention,” *ArXiv*, vol. abs/1605.09090, 2016.
- [20] S. Wang and J. Jiang, “Learning natural language inference with lstm,” in *NAACL*, 2016.
- [21] L. Sha, B. Chang, Z. Sui, and S. Li, “Reading and thinking: Re-read lstm unit for textual entailment recognition,” in *COLING*, 2016.
- [22] Q. Chen, X.-D. Zhu, Z. Ling, S. Wei, H. Jiang, and D. Inkpen, “Enhanced lstm for natural language inference,” in *ACL*, 2017.
- [23] L. Mou, R. Men, G. Li, Y. Xu, L. Zhang, R. Yan, and Z. Jin, “Natural language inference by tree-based convolution and heuristic matching,” *arXiv: Computation and Language*, 2016.
- [24] L. Mou, H. Peng, G. Li, Y. Xu, L. Zhang, and Z. Jin, “Discriminative neural sentence modeling by tree-based convolution,” in *EMNLP*, 2015.
- [25] Y. Gong, H. Luo, and J. Zhang, “Natural language inference over interaction space,” *ArXiv*, vol. abs/1709.04348, 2018.
- [26] P. Martínez-Gómez, K. Mineshima, Y. Miyao, and D. Bekki, “On-demand injection of lexical knowledge for recognising textual entailment,” in *EACL*, 2017.
- [27] K. J. Shih, S. Singh, and D. Hoiem, “Where to look: Focus regions for visual question answering,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4613–4621, 2016.
- [28] J. Lu, J. Yang, D. Batra, and D. Parikh, “Hierarchical question-image co-attention for visual question answering,” in *NIPS*, 2016.
- [29] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21–29, 2016.
- [30] I. Schwartz, A. Schwing, and T. Hazan, “High-order attention models for visual question answering,” in *NIPS*, 2017.
- [31] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.
- [32] D.-K. Nguyen and T. Okatani, “Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering,” *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6087–6096, 2018.
- [33] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, “Deep modular co-attention networks for visual question answering,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6274–6283, 2019.
- [34] P. Gao, H. Li, H. You, Z. Jiang, P. Lu, S. Hoi, and X. Wang, “Dynamic fusion with intra- and inter-modality attention flow for visual question answering,” *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6632–6641, 2019.
- [35] P. Huang, J. Huang, Y. Guo, M. Qiao, and Y. Zhu, “Multi-grained attention with object-level grounding for visual question answering,” in *ACL*, 2019.
- [36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *ICML*, ser. Proceedings of Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.
- [37] R. Suzuki, H. Yanaka, M. Yoshikawa, K. Mineshima, and D. Bekki, “Multimodal logical inference system for visual-textual entailment,” in *ACL*, 2019.
- [38] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “Uniter: Universal image-text representation learning,” in *ECCV*, 2020.
- [39] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *EMNLP*, 2014.
- [40] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen, “In defense of grid features for visual question answering,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10 264–10 273, 2020.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [42] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *ArXiv*, vol. abs/1706.03762, 2017.
- [43] S. Ren, K. He, R. B. 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, pp. 1137–1149, 2015.
- [44] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
- [45] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2015.
- [46] J. Ba, J. Kiros, and G. E. Hinton, “Layer normalization,” *ArXiv*, vol. abs/1607.06450, 2016.
- [47] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, “Pytorch: An imperative style, high-performance deep learning library,” in *NeurIPS*, 2019.
- [48] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2015.



**Biwei Cao** received the B.S. degree in Software Engineering from Australian National University, in 2019, the M.S. degree in Computing from Australian National University, in 2020. She is now working toward the Ph.D degree from the School of Cyber Science and Engineering, Southeast University, Nanjing, China. Her research interests include affective computing, gesture recognition, social computing and language generation.



**Lei He** received the Ph.D. degree from Information Engineering University, China, in 2008. He is an Associate Researcher of Information Engineering University, China. His current research interests include Cyberspace security and Mimic defense.



**Jiuxin Cao** received the Ph.D degree from Xi'an Jiaotong University in 2003. He is currently a professor at the School of Cyber Science and Engineering in Southeast University, Nanjing, China. His research interests include computer networks, social computing, affective computing, behavior analysis, and big data security and privacy preservation. He is the Director of Jiangsu Provincial Key Laboratory of Computer Network Technology, the Senior Member of China Computer Federation, the Member of Chinese Information Processing Society of China, the Fellow of Jiangsu Computer Society, the Member of Jiangsu Information Security Standardization Committee, and the Member of JSAT-ISA.



**Yuan Yan Tang** is an IEEE Life Fellow, IAPR Fellow, and AAIA Fellow. He currently is the Director of Smart City Research Center in Zhuhai UM Science & Technology Research Institute, is also the Emeritus Chair Professor at University of Macau and Hong Kong Baptist University, Adjunct Professor at Concordia University, Canada. His current research interests include artificial intelligence, wavelets, pattern recognition, and image processing. He has published more than 600 academic papers and is the author (or coauthor) of over 25 monographs, books and bookchapters. He is the Founder and Editor-in-Chief of SCI journal "International Journal on Wavelets, Multiresolution, and Information Processing (IJWMIP)". Dr. Tang is the Founder and General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition (ICWAPRs). He is the Founder and Chair of the Macau Branch of International Associate of Pattern Recognition (IAPR). He has serviced as general chair, program chair, and committee member for many international conferences. Dr. Tang served as the Chairman of 18th ICPR, which is the first time that the ICPR was hosted in China.



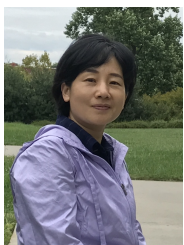
**Jie Gui** received the M.S. degree in computer applied technology from Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China, in 2007, and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2010. He is currently a Professor in the School of Cyber Science and Engineering, Southeast University. He has published more than 40 papers in international journals and conferences such as IEEE TPAMI, IEEE TNNLS, IEEE TCYB, IEEE TIP, IEEE TCSVT, IEEE TSMCS, KDD, and ACM MM. He is the Area Chair, Senior PC member, or PC Member of many conferences such as NeurIPS and ICML. His research interests include machine learning, pattern recognition, and image processing.



**James Tin-Yau Kwok** received the Ph.D. degree in computer science from The Hong Kong University of Science and Technology, Hong Kong, in 1996. He is currently a Professor with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology. His current research interests include kernel methods, machine learning, pattern recognition, and artificial neural networks. He received the IEEE Outstanding Paper Award in 2004 and the Second Class Award in Natural Sciences from the Ministry of Education, China, in 2008. He has been a Program Co-Chair for a number of international conferences, and served as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS from 2006 to 2012. He is currently an Associate Editor of *Neurocomputing*.



**Jiayun Shen** received the M.S. degree from the School of Cyber Science and Engineering, Southeast University, in 2020. His research interests includes object detection, visual analysis, and cross modal analysis.



**Bo Liu** works as a professor and the doctoral advisor with Southeast University, China. She received her doctoral degree from Southeast University. She won the first class Science and Technology Progress Award of MoE in 2009, and she is currently working on two NSF projects. She has published more than 60 papers and most of them have been published in reputed journals and conferences including WWW, WWWJ, ToN and et al. Her current main research interests include spammer detection in social network, the evolution of social community, social influence, and social recommendation.