Joint Visual-Textual Sentiment Analysis Based on Cross-modality Attention Mechanism

Xuelin Zhu, Biwei Cao, Jiuxin Cao et al.

Southeast University, Nanjing, China

The 25th International Conference on MultiMedia Modeling (MMM 2019)
January 8 - 11, 2019, Thessaloniki, Greece

OUTLINE

01

Introduction

01 Introduction

① Huge Volume of Image-Text Data

Statistics indicates that about 25% of tweets contains image information and 99% of image tweets contain textual information.

② Limitation of Single Modality

Due to the complexity and variability of user-generated content, the performance of sentiment analysis based on single modality (image or text) still lags behind of satisfaction.

01 Introduction

③ Challenge

Joint visual-textual sentiment analysis is challenging since image and text may deliver inconsistent sentiment.

(a) Woman enjoying a quiet time with a fresh cup of tea.

(b) Young people jumping on Mission Beach. San Diego, California, USA.

(c) My God, here is so crowded.

Visual information and textual information should differ in their contribution to sentiment analysis.

02

Related Work

02 Related Work

- Early Fusion and Late Fusion
- Attention for Multimodal Tasks

Early Fusion and Late Fusion

Early fusion employs feature fusion techniques to learn a joint visual-textual semantic representation for sentiment analysis, Late fusion treats image and text information separately by leveraging different domain-specific techniques, and subsequently utilize all modalities' sentiment label to obtain the ultimate results.

However, due to the **semantic gap** between visual and textual information, the performance of early fusion and late fusion is limited.

02 Related Works

Attention For Multimodal Tasks

Automatic image captioning and multimodal matching between image and sentence have shown the advance of deep neural networks in understanding and jointly modeling vision and text content, and inspired some ideas of joint feature learning, design of attention model, and so on.

[1] Vinyals O, Toshev A, Bengio S, et al. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge (TPAMI 2017).

[2] Xu K, Ba J, Kiros R, et al. Show, attend and tell: Neural image caption generation with visual attention (ICML 2015).

[3] Wang L, Li Y, Huang J, et al. Learning two-branch neural networks for image-text matching tasks (TPAMI 2018).

[4] Yang Z, Yang D, Dyer C, et al. Hierarchical attention networks for document classification (ACL 2016).

[5] You Q, Jin H, Luo J. Visual Sentiment Analysis by Attending on Local Image Regions (AAAI 2017).

[6] You Q, Cao L, Jin H, et al. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks (MM 2016).

02 Related Works

Summary on Related Work

- The performance of early fusion and late fusion is limited when image-text pairs carry inconsistent sentiment.
- So far, very few studies have considered that visual and textual information should differ in their contribution to sentiment analysis.

03

Model Description

03 Model Description

Intuition

- Not both text and image contribute equally to the sentiment classification.
- Visual information and several key emotional words in sequence mainly determine the semantic polarity.

Two Problems

- How to bridge the semantic gap between visual information and textual information?
- How to assign reasonable weights to visual information and textual information?

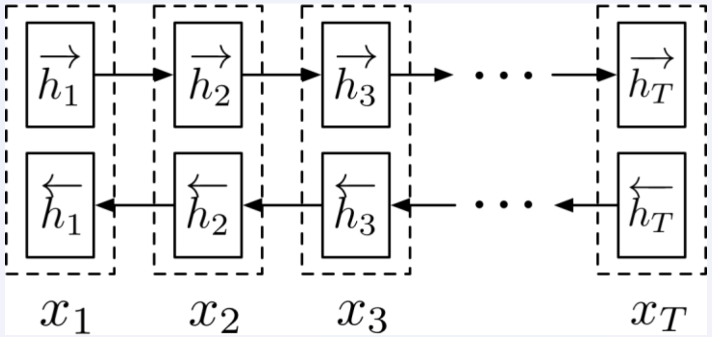
03 Model Description

Bidirectional RNN For Semantic Embedding

Given the input words sequence: $\{x_1, x_2, ..., x_T\}$

Forward RNN

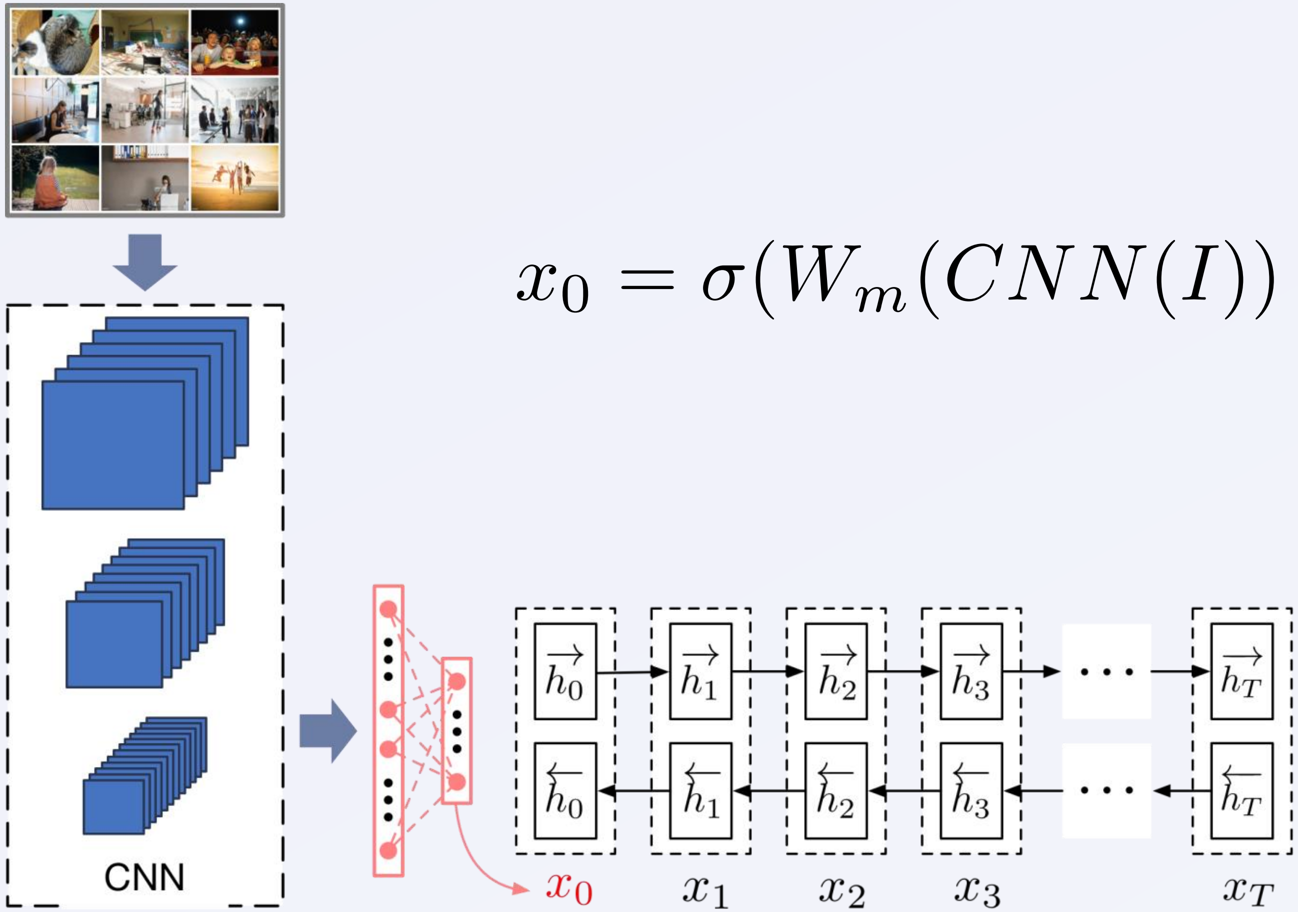
Backward RNN



Hidden State: $h_j = \left[\overrightarrow{h}_j^T; \overleftarrow{h}_j^T\right]^T$

03 Model Description

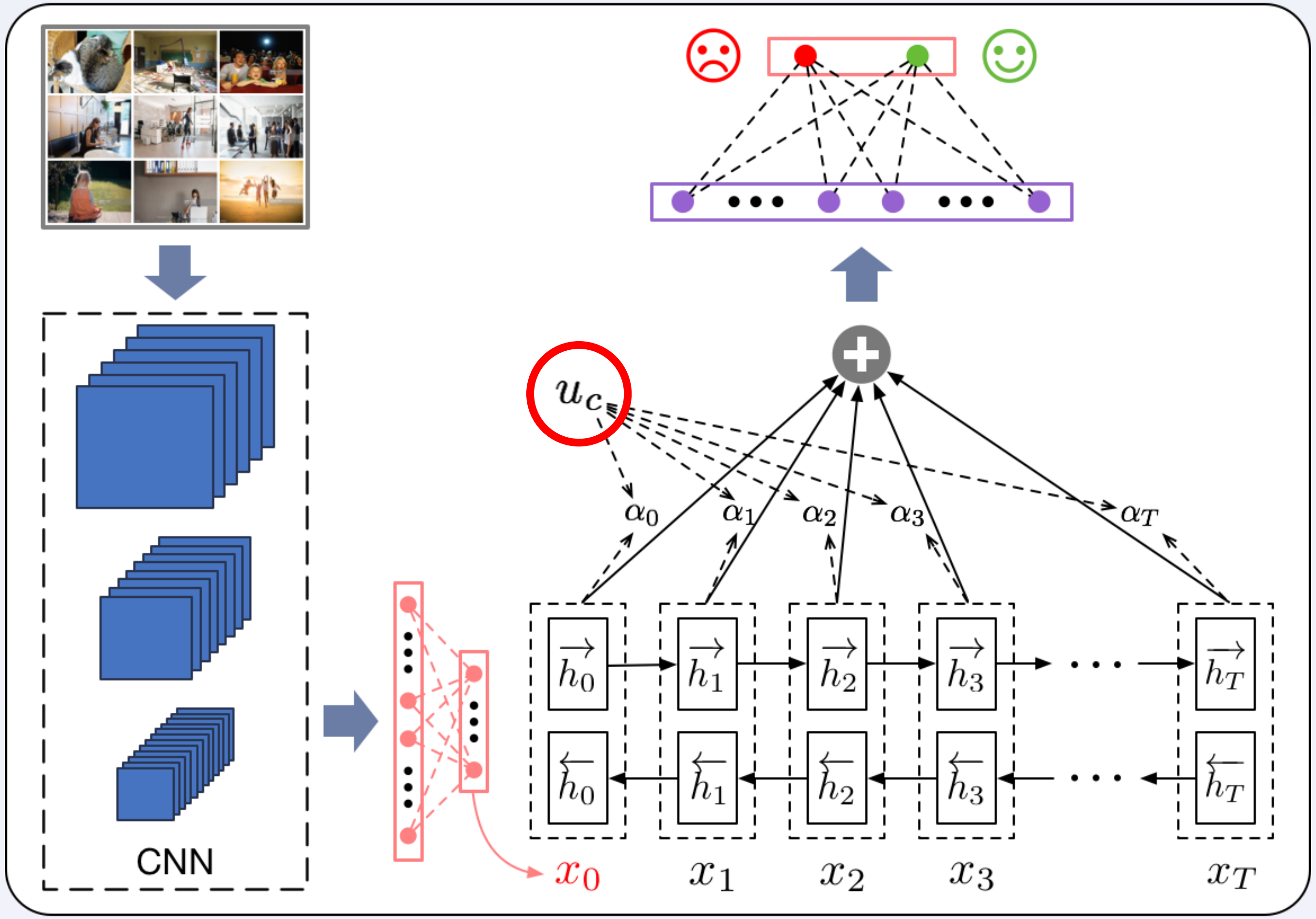
Bidirectional RNN For Semantic Embedding



$$x_0 = \sigma(W_m(CNN(I)) + b_m)$$

03 Model Description

Cross-modality Attention Mechanism



$$For \quad i = 0, 1, ..., T$$

$$u_i = \tanh(W_w h_i + b_w)$$

$$\alpha_i = \frac{\exp(u_i^T u_c)}{\sum_{i=0}^{T} \exp(u_i^T u_c)}$$

$$s = \sum_{i=0}^{T} \alpha_i h_i$$

$$logit = W_s(\sigma(W_h s) + b_h) + b_s$$

04

Experiments

04 Experiments

4-1 Datasets

4-2 Comparison Methods

4-3 Results & Analysis

4-1 Datasets

Table I. Statistics of two datasets.

Datasets	Positive	Negative	Total
Getty[1]	188,028	181,008	369,036
VSO[2]	118,869	87,139	206,008

1. https://www.gettyimages.co.uk/
2. http://www.ee.columbia.edu/ln/dvmm/vso/download/flickr_dataset.html

4-2 Comparison Methods

◆ Early Fusion、 Later Fusion、 T-LSTM Embedding

You Q, Cao L, Jin H, et al. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks (ACMMM 2016).

◆ CCR

You Q, Luo J, Jin H, et al. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia (WSDM 2016).

◆ Deep Fusion

Chen X, Wang Y, Liu Q. Visual and textual sentiment analysis using deep fusion convolutional neural networks (ICIP 2017)

4-2 Comparison Methods

- **RNN Embedding**

  Learn the BiRNN with semantic embedding.

- **RNN-CA**

  Learn the BiRNN with cross-modality attention mechanism.

- **RNN-CA Embedding**

  Learn the BiRNN with cross-modality attention mechanism and semantic embedding simultaneously.

4-3 Results & Analysis

I. Results on the Getty testing dataset

Models	Prec.	Rec.	F1	Acc.
Early Fusion	0.684	0.706	0.695	0.684
Later Fusion	0.717	0.745	0.731	0.720
CCR	0.811	0.746	0.777	0.782
T-LSTM Embedding	0.889	0.903	0.896	0.892
Deep Fusion	0.895	0.919	0.907	0.905
RNN Embedding	0.881	0.902	0.891	0.888
RNN-CA	0.877	0.896	0.886	0.884
RNN-CA Embedding	**0.909**	**0.923**	**0.916**	**0.913**

4-3 Results & Analysis

II. Results on the VSO testing dataset

Models	Prec.	Rec.	F1	Acc.
Early Fusion	0.636	0.800	0.709	0.620
Later Fusion	0.645	0.885	0.746	0.652
CCR	0.653	0.661	0.657	0.668
T-LSTM Embedding	0.823	0.834	0.828	0.829
Deep Fusion	0.827	0.849	0.838	0.842
RNN Embedding	0.813	0.831	0.822	0.827
RNN-CA	0.806	0.823	0.814	0.815
RNN-CA Embedding	**0.838**	**0.856**	**0.847**	**0.851**

4-3 Results & Analysis

III. Results on the image-text pairs with opposite sentiments

How？ RNTN[1]、Fine-tuned CaffeNet[2]

Datasets	Early Fusion	Later Fusion	CCR	T-LSTM Embedding	Deep Fusion	RNN-CA Embedding
Getty	0.650	0.700	0.753	0.856	0.873	**0.911**
VSO	0.583	0.631	0.649	0.795	0.801	**0.849**

[1] Socher, Richard, et al. “Recursive deep models for semantic compositionality over a sentiment treebank.” Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.

[2] Campos, Victor, et al. "Diving deep into sentiment: Understanding fine-tuned cnns for visual sentiment prediction." Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia. ACM, 2015.

4-3 Results & Analysis

IV. Qualitative attention analysis

1) [image] Mother and daughter having fun time in bed room.
2) [image] Shot of a happy senior woman spending quality time with her daughter outdoors.
3) [image] Portrait of an attractive young woman enjoying a boat ride on the lake.

(a) Top RNN-CA Embedding positive examples.

1) [image] Breakup of a couple with bad girl and sad boyfriend.
2) [image] A powerful EF-5 tornado rips through Greensburg, destroying most of the town.
3) [image] Office worker stressed and upset in office.

(b) Top RNN-CA Embedding negative examples.

4-3 Results & Analysis

IV. Qualitative attention analysis

1) [image] Little girl sleeping on her Father on the train.
2) [image] Two men are busy working in office.
3) [image] Young couple hugging in front of cars.

(c) Image dominating sentiment classification examples.

1) [image] Portrait of a woman against rocket launch.
2) [image] Sad girl sitting with head down.
3) [image] My God, here is too crowded.

(d) Text dominating sentiment classification examples.

05

Conclusion

05 Conclusion

BiRNN For Semantic Embedding

BiRNN is capable of semantic embedding learning and bridging semantic gap between image information and text information.

Cross-modality Attention Mechanism

The cross-modality attention model is qualified for automatically assigning weights to visual and textual information.

Extensive Experiments

Extensive Experiments validate the superiority of the proposed model, especially when images and texts carry opposite sentiments.

THANK YOU!