ITA: Image-Text Alignments for Multi-Modal Named Entity Recognition

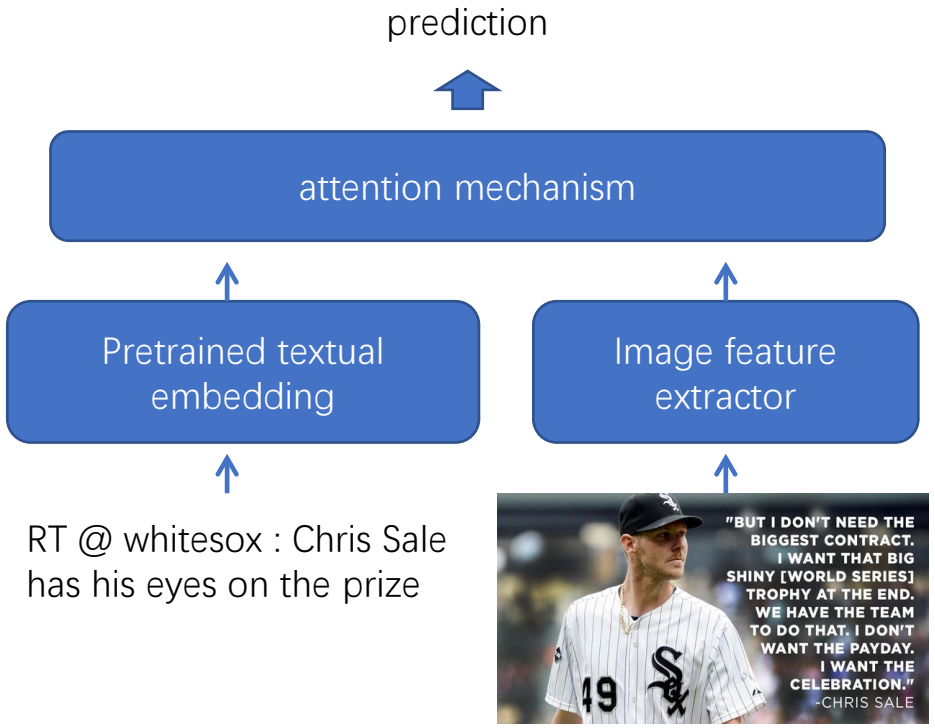
Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang,
Zhongqiang Huang, Fei Huang, Kewei Tu

上海科技大学
ShanghaiTech University

DAMO
ALIBABA DAMO ACADEMY

Background

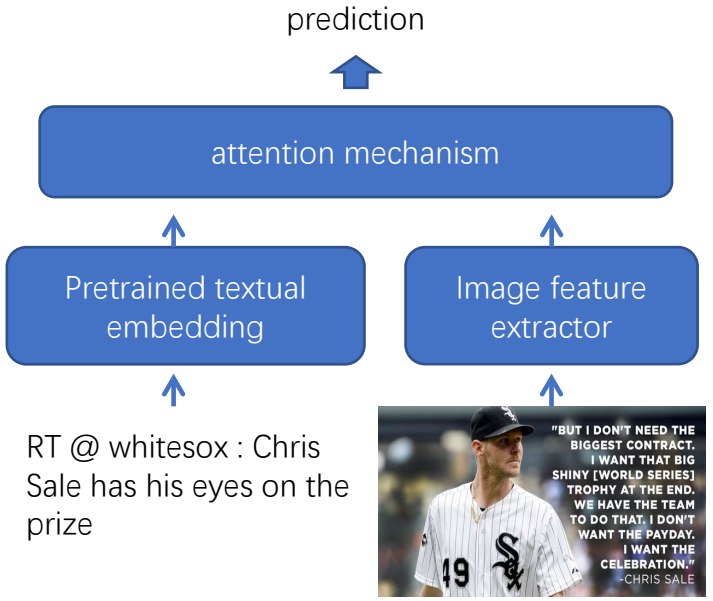
Named Entity Recognition (NER) has been applied to a lot of domains such as news , E-commerce , social media and bio-medicine . Several recent studies focus on improving the accuracy of NER models through utilizing image information (MNER) in tweets .

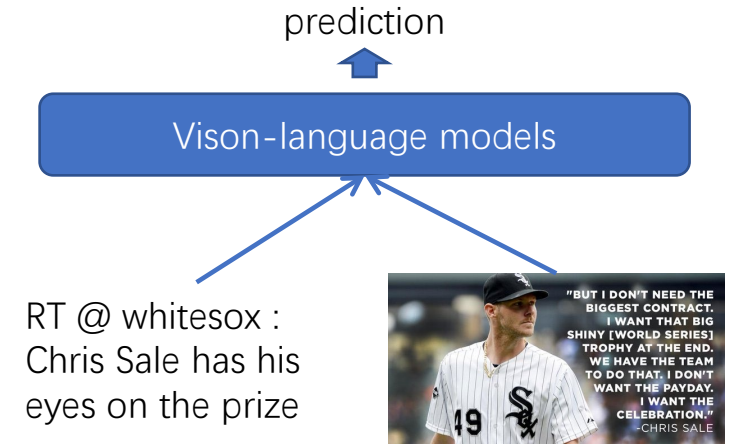


Most approches to MNER use the attention mechanism to model the interaction between image and text representations which are pretrained based on mono-modal data separately.

Problems

- Image and text representations are trained separately and not aligned
- Pretrained vision-language (V+L) models do not work well on MNER
  - The models are trained with common nouns instead of named entities
  - The image modality only plays an auxiliary role in MNER





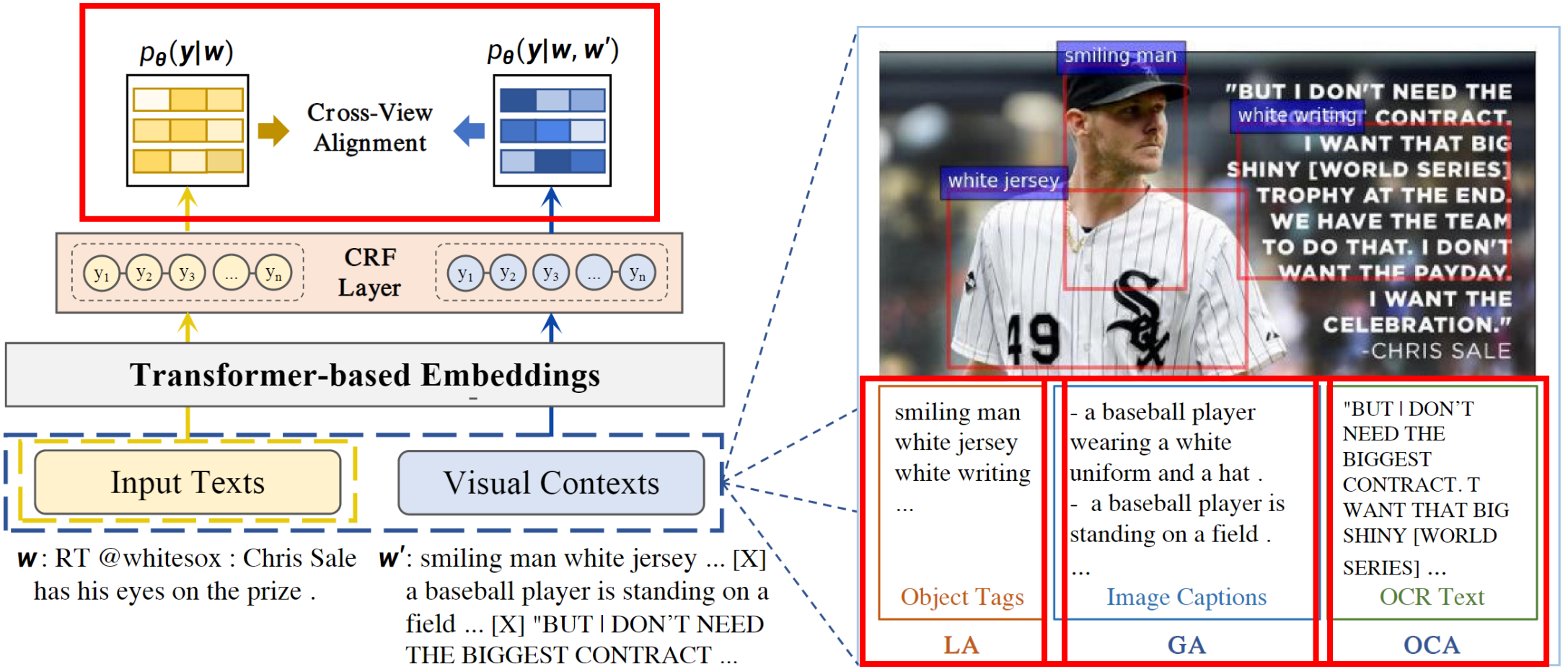
MNER based on VLP models

Motivation

Pretrained textual embeddings can utilize contexts to improve the token representation of a sequence, maybe the images in MNER can be converted to texts as contexts?

- By converting the image to texts:
  - the image representations can be aligned to the space of text representations
  - the attention module of the pretrained textual embeddings can easily model the interactions between aligned image and text representations, without introducing a new attention module.

Model Architecture



Cross-View Alignment

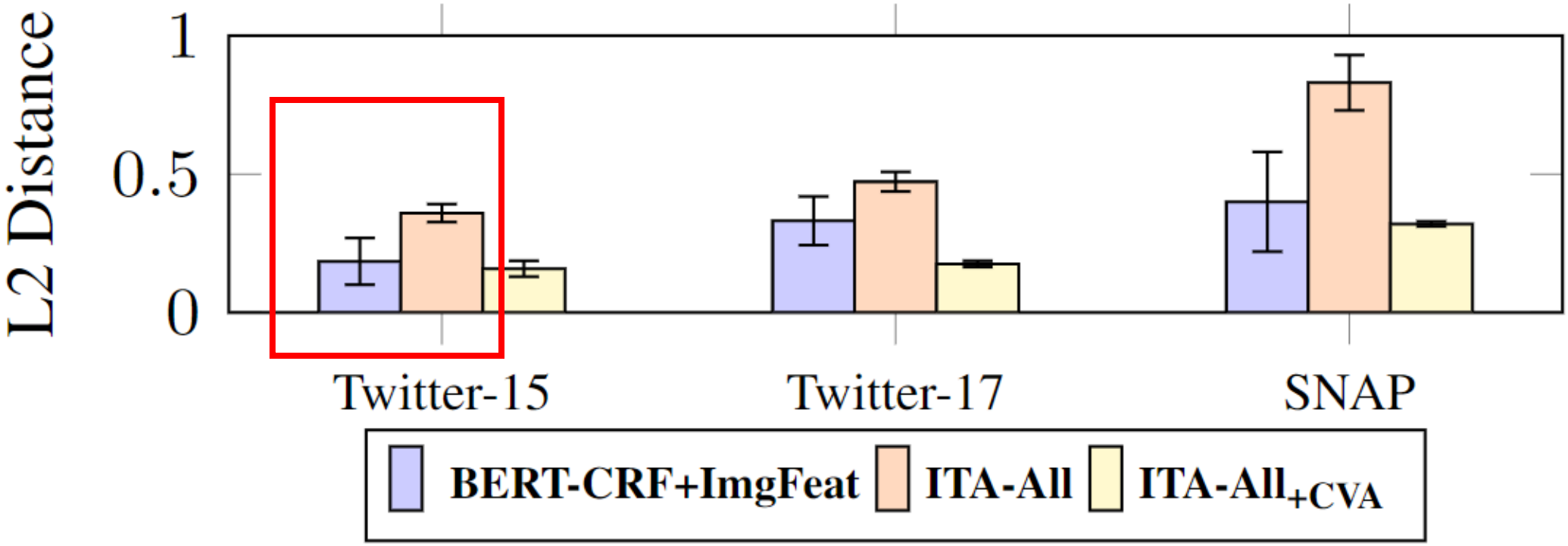
- Limitations
  - What if images are not available?
  - What about time-critical scenarios (aligning images to texts requires several steps in pre-processing)?
  - Noises in the image can mislead the MNER to make wrong predictions
- Solution
  - Cross-View Alignment (CVA) between the multi-modal input view and text-only input view

$$\mathcal{L}_{\mathrm{CVA}}(\theta) = \sum_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} p_\theta(\boldsymbol{y}|\hat{\boldsymbol{w}}) \log p_\theta(\boldsymbol{y}|\boldsymbol{w})$$

Results

Approach	Twitter-15	Twitter-17	SNAP
REPORTED F1 OF PREVIOUS APPROACHES			
BERT-CRF†	71.81	83.44	-
OCSGA♣	72.92	-	-
UMT†	73.41	85.31	-
RIVA‡	73.80	-	86.80
RpBERT$_{base}$♠	74.40	-	87.40
UMGF◇	74.85	85.51	-
OUR REPRODUCTIONS			
BERT-CRF	74.79	85.18	85.98
UMT	72.83	84.88	-
UMGF	74.42	85.27	-
RpBERT$_{base}$	67.21	-	62.14
Ours: **ITA-All$_{+CVA}$**	**76.01**	**86.45**	**87.44**

How ITA Eases the Cross-Modal Alignments



Conclusions

- ITA converts images into object labels, captions and OCR texts to align the image representations into textual space
- CVA let the MNER models better utilize the text information in the input
- We show that ITA significantly outperforms previous state-of-the-art approaches on MNER datasets
- We further analyze how ITA eases the cross-modal alignments and how the images affect the NER prediction