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

ShanghaiTech University & Alibaba Group

Paper:

Code:

Motivation

Multi-modal named-entity recognition (MNER) seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories with addictional visual information.

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

Solution: Pretrained textual embeddings can utilize contexts to improve the token representation of a sequence, so we propose:

- **ITA**: Convert the images into texts to utilize textual embeddings

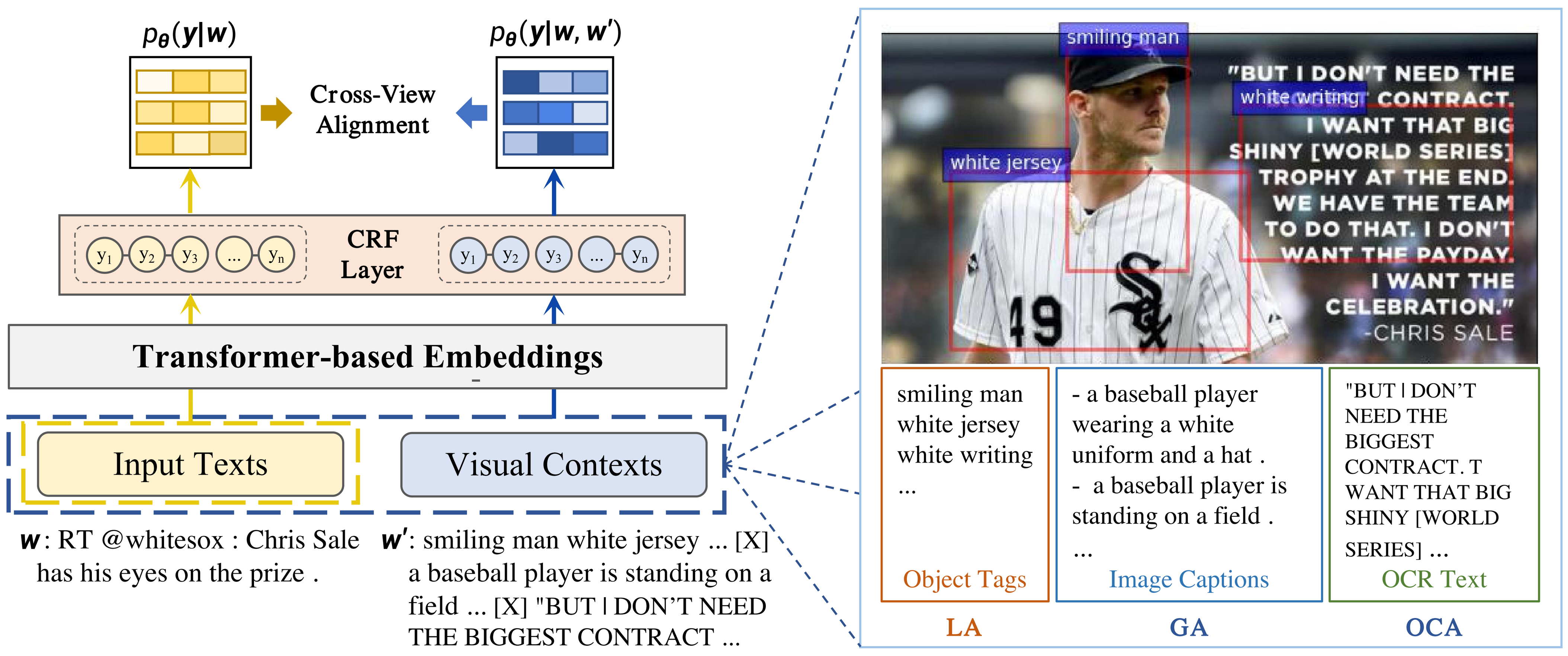

Figure 1. The architecture of ITA. ITA aligns an image into object tags, image captions and texts from OCR. ITA takes them as visual contexts and then feeds them together with the input texts into the transformer-based embeddings. In the cross-view alignment module, ITA minimizes the distance between the output distribution of cross-modal inputs and textual inputs.

Image-text Alignments (ITA)

Object Tags as Local Alignment (LA)

The image is localized into objects with an object detector (OD) and the tags of each region textually describe the local information in the image.

$$\boldsymbol{a}, \boldsymbol{o} = \mathrm{OD}(\boldsymbol{I}), \text{ where}$$

$$\boldsymbol{a} = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_l\} \text{ and } \boldsymbol{o} = \{\boldsymbol{o}_1, \boldsymbol{o}_2, \cdots, \boldsymbol{o}_l\};$$

$$\boldsymbol{w}^{\mathrm{LA}} = \{\boldsymbol{a}_1, \boldsymbol{o}_1, \boldsymbol{a}_2, \boldsymbol{o}_2, \cdots, \boldsymbol{a}_l, \boldsymbol{o}_l\}$$

Image Captions as Global Alignment (GA)

The global information of the image is presented by captions, which is predicted by an image captioning model (IC).

$$\{\boldsymbol{w}^1, \boldsymbol{w}^2, \cdots, \boldsymbol{w}^k\} = \mathrm{IC}(\boldsymbol{I})$$

$$\boldsymbol{w}^{\mathrm{GA}} = [\boldsymbol{w}^1, [\mathrm{X}], \boldsymbol{w}^2, [\mathrm{X}], \cdots, [\mathrm{X}], \boldsymbol{w}^k]$$

Optical Character Alignment (OCA)

The texts in the images are extracted by OCR model to better utilize the enriched semantic information conveyed by the images.

$$\boldsymbol{w}^{\mathrm{OCA}} = \mathrm{OCR}(\boldsymbol{I})$$

Cross-View Alignment (CVA)

Text-image (**I+T**) input view is denoted as one of $\boldsymbol{w}^{\mathrm{LA}}$, $\boldsymbol{w}^{\mathrm{GA}}$, $\boldsymbol{w}^{\mathrm{OCA}}$ or the concatenation of all (**All**). CVA minimizes the KL divergence over the probability distribution of **I+T** and text (**T**) input views to overcome noises from the image information.

$$\mathcal{L}_{\mathrm{CVA}}(\theta) = \mathrm{KL}(p_\theta(\boldsymbol{y}|\hat{\boldsymbol{w}}) || p_\theta(\boldsymbol{y}|\boldsymbol{w}))$$

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

Experiment Results

Train Modal	Approach	Twitter-15 Eval Modal T	Twitter-15 Eval Modal I+T	Twitter-17 Eval Modal T	Twitter-17 Eval Modal I+T	SNAP Eval Modal T	SNAP Eval Modal I+T
	BERT-CRF						
T	**BERT-CRF**	74.79	-	85.18	-	85.98	-
I+T	**ITA-LA**	-	75.18	-	85.67	-	86.26
	ITA-GA	-	75.17	-	85.75	-	86.72
	ITA-OCA	-	75.01	-	85.64	-	86.52
	ITA-All	-	75.15	-	85.78	-	86.79
	ITA-LA$_{+CVA}$	75.26	75.20	85.72	85.62	86.51	86.41
	ITA-GA$_{+CVA}$	75.45	75.52	85.96	**85.85**	86.42	86.39
	ITA-OCA$_{+CVA}$	75.26	75.30	85.73	85.79	86.64	86.59
	ITA-All$_{+CVA}$	**75.67**	**75.60**	**85.98**	85.72	**86.83**	**86.75**
	XLMR-CRF						
T	**XLMR-CRF**	77.37	-	88.73	-	89.39	-
I+T	**ITA-LA**	-	77.64	-	89.29	-	89.68
	ITA-GA	-	77.78	-	89.32	-	89.78
	ITA-OCA	-	77.94	-	89.31	-	89.64
	ITA-All	-	77.81	-	89.62	-	90.10
	ITA-LA$_{+CVA}$	77.87	77.93	89.45	**89.90**	89.85	89.91
	ITA-GA$_{+CVA}$	78.03	78.02	89.41	89.62	89.85	90.09
	ITA-OCA$_{+CVA}$	77.57	77.59	89.32	89.55	89.90	89.84
	ITA-All$_{+CVA}$	**78.25**	**78.03**	**89.47**	89.75	**90.02**	**90.15**

Table 1. A comparison of ITA and our baseline.

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**

Table 2. A comparison of our approaches and state-of-the-art approaches.

Analysis and Discussion

Approach	Twitter-15 Eval Modal T	Twitter-15 Eval Modal I+T	Twitter-17 Eval Modal T	Twitter-17 Eval Modal I+T	SNAP Eval Modal T	SNAP Eval Modal I+T
ITA-Random	-	74.67	-	84.98	-	85.82
ITA-GA$_{BU}$	-	75.10	-	85.77	-	86.51
ITA-LA$_{BU}$	-	75.18	-	85.59	-	86.57
ITA-OCA$_{Paddle}$	-	75.12	-	85.87	-	86.66
BERT-CRF$_{+ImgFeat}$	-	74.70	-	84.99	-	85.90
VinVL-CRF	-	60.58	-	75.55	-	74.53
BERT+VinVL-CRF	-	74.89	-	85.19	-	86.14
ITA-Joint	74.88	75.22	85.31	85.60	86.06	86.34
References						
RpBERT w/o Rp	-	72.60	-	-	-	86.20
ITA-All$_{+CVA}$	**75.50**	**75.41**	**85.89**	**85.84**	**86.83**	**86.75**

Table 3. A comparison of other variants of MNER models.

	LOC P	LOC R	LOC F1	ORG P	ORG R	ORG F1	PER P	PER R	PER F1	OTHER P	OTHER R	OTHER F1
Twitter-15												
BERT-CRF	80.0	83.8	81.8	65.9	61.0	63.3	84.2	86.8	85.4	44.2	44.2	44.1
ITA-All$_{+CVA}$	81.1	84.2	82.6	68.8	60.6	64.4	84.0	87.2	85.6	44.9	44.6	44.8
Δ	1.1	0.4	0.8	2.8	-0.4	1.1	-0.2	0.4	0.1	0.8	0.5	0.6
Twitter-17												
BERT-CRF	85.5	84.4	84.9	83.5	83.8	83.7	90.7	90.8	90.7	68.9	65.1	66.9
ITA-All$_{+CVA}$	86.0	83.7	84.8	83.9	84.2	84.0	91.9	90.9	91.4	73.7	64.3	68.6
Δ	0.5	-0.7	-0.1	0.3	0.4	0.4	1.2	0.1	0.7	4.8	-0.8	1.7
SNAP												
BERT-CRF	82.1	82.8	82.5	87.8	86.9	87.3	91.0	91.5	91.2	72.3	75.1	73.7
ITA-All$_{+CVA}$	80.3	81.7	81.0	87.8	86.5	87.1	90.1	91.2	90.6	70.1	73.2	71.6
Δ	1.9	1.1	1.5	0.6	0.5	0.5	0.9	0.3	0.6	2.2	1.9	2.1

Table 4. A comparison between our ITA and the baseline in precision (P), recall (R) and F1. **Δ**: the relevant improvement of ITA over the Baseline.

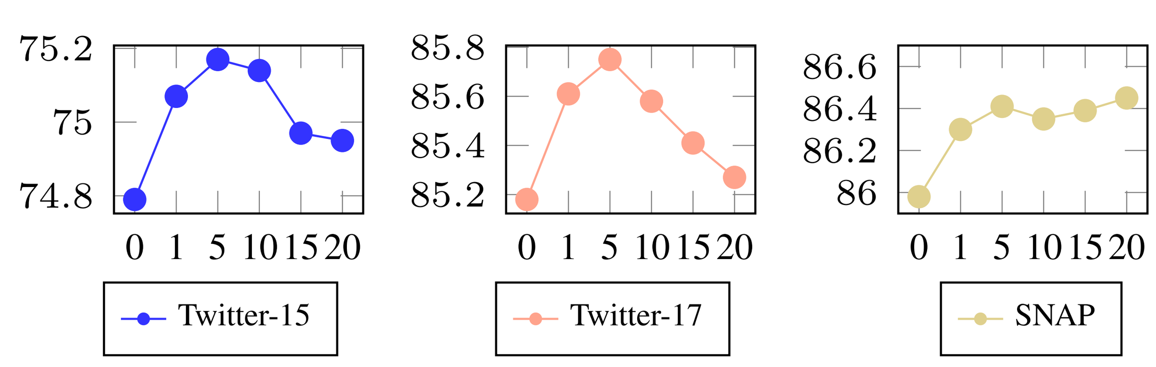

Figure 2. A relation between the number of captions input to the MNER model and model accuracy. The x-axis is the number of captions. The y-axis is the averaged F1 score on the test set.

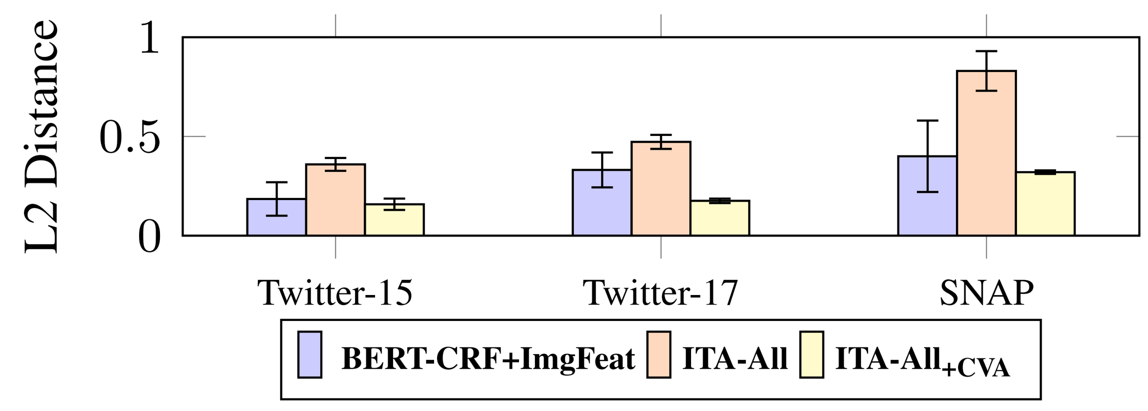

Figure 3. Averaged L2 distance between the token representations without image input (r_i) and with image input (r_i'). The error bars mean the standard deviation over **5** runs.

Case Study

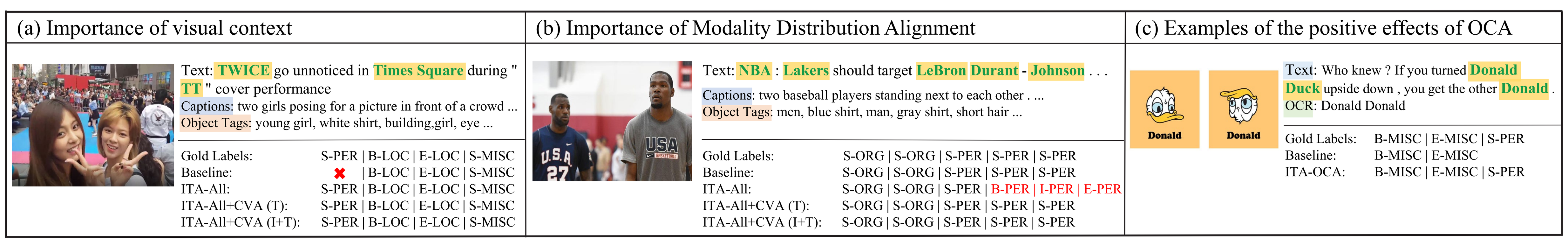

Figure 4. Three case studies to show the effectiveness of ITA.