LREC-COLING 2024

MNER-MI: A Multi-image Dataset for Multimodal Named Entity Recognition in Social Media

Shizhou Huang[1], Bo Xu[2], Changqun Li[1], Jiabo Ye[1], and Xin Lin[1,3]

[1]School of Computer Science and Technology, East China Normal University

[2]School of Computer Science and Technology, Donghua University

[3]Shanghai Key Laboratory of Multidimensional Information Processing

Outline

LREC-COLING 2024

- **Background**
- Motivation of Our Work
- Our Datasets
- Our proposed Framework
- Experiments
- Conclusions

Background

LREC-COLING 2024

- Named Entity Recognition (NER)
  - Task Defination
    - Given a text as input, the task of NER is to detect entities from **text** and classify each entity into pre-defined types, e.g., Person (PER), Location (LOC).

Text: The things that Stephen Curry can do with a basketball should be studied. → The things that [$_{\text{PER}}$ Stephen Curry] can do with a basketball should be studied.

  - Limitation
    - The type of named entity in the text may be **ambiguous** when the text is short resulting in insufficient semantics. This is very common on social media platforms.

Text: Handsome Rob after a fish dinner. → Person or Others?

Background

LREC-COLING 2024

- Multimodal Named Entity Recognition (MNER)
  - MNER has become a popular research direction since images as additional input can be supplemented with semantics for disambiguation.
  - Task Defination
    - Given a **text** and its **associated image** as input, the task of MNER is to detect entities from **text** and each entity into pre-defined types.

Text: Handsome Rob after a fish dinner.

→ Handsome [$_{\text{MISC}}$ Rob] after a fish dinner.

Outline

LREC-COLING 2024

- Background
- **Motivation of Our Work**
- Our Datasets
- Our proposed Framework
- Experiments
- Conclusions

Motivation of Our Work

LREC-COLING 2024

- Limitations of Existing Work
  - Focusing on a single image
    - According to Zhang[1], over **42%** of tweets containing more than one image.
    - But the current MNER datasets and methods are predominantly based on text and a **single** accompanying image.

DEATH BATTLE [$_{PER}$ Kim Kardashian] VS [$_{MISC}$ Domo]

Only a single image is stored

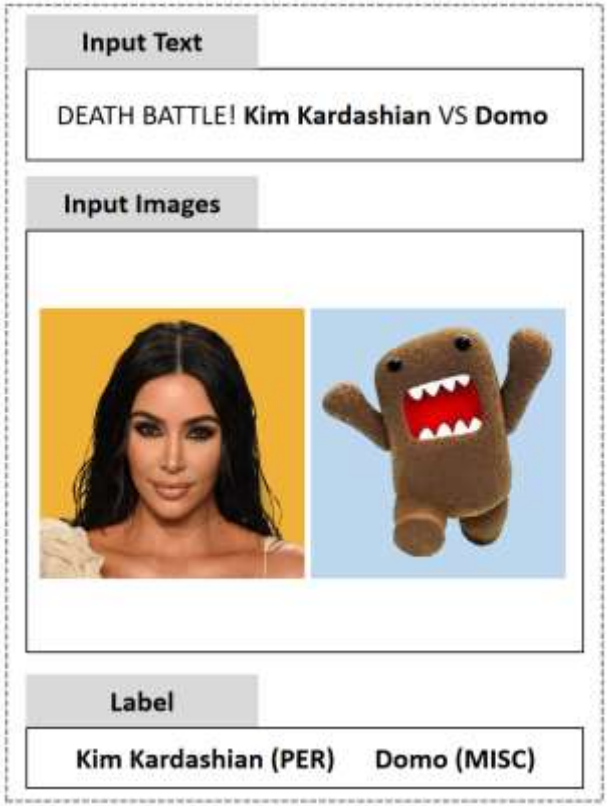
The current dataset will only save the first image (or a random one) for tweets with multiple images

[1] *Adaptive Co-attention Network for Named Entity Recognition in Tweets, 2018 AAAI*

Motivation of Our Work

LREC-COLING 2024

- Limitations of Existing Work
  - Overlooking the necessity of multiple images to understand multi-image posts
    - Posts with multiple images can help alleviate the ambiguity present in posts with only one image and identify more entities in the text.



Outline

LREC-COLING 2024

- Background
- Motivation of Our Work
- **Our Datasets**
- Our proposed Framework
- Experiments
- Conclusions

Our Datasets

LREC-COLING 2024

- MNER-MI
  - Dataset Construction
    - To bridge the critical research gap in multiple images at MNER, we annotate a novel dataset.
    - **URL** can help subsequent works to find our original data source and expand on it.

Text: DEATH BATTLE [$_{\text{PER}}$ Kim Kardashian] VS [$_{\text{MISC}}$ Domo]

Images:

URL:https://twitter.com/i/web/status/1554138525414526977

Our Datasets

LREC-COLING 2024

- Statistics of MNER-MI and MNER-MI-Plus
  - MNER-MI
    - Multi-image dataset constructed by us, containing at least 2 images and at most 4 images (Twitter limit).
  - MNER-MI-Plus
    - Obtained by merging MNER-MI and Twitter-2017 for evaluating the model in both single-image scenarios and multi-image scenarios.

Type	MNER-MI			MNER-MI-Plus		
	Train	Dev	Test	Train	Dev	Test
Person	4,529	573	439	7,472	1,199	1,060
Location	1,878	210	156	2,609	383	334
Organization	1,273	165	92	2,947	540	487
Miscellaneous	2,054	260	233	2,755	410	390
Total	9,734	1,208	920	1,5783	2,532	2,271
# One Image	0	0	0	3,373	723	723
# Two Images	3,711	446	455	3,711	446	455
# Three Images	814	110	135	814	110	135
# Four Images	2,331	304	270	2,331	304	270
# Images per Tweet	2.799	2.835	2.785	2.206	1.997	1.970
# Tweets	6,856	860	860	10,229	1,583	1,583

Outline

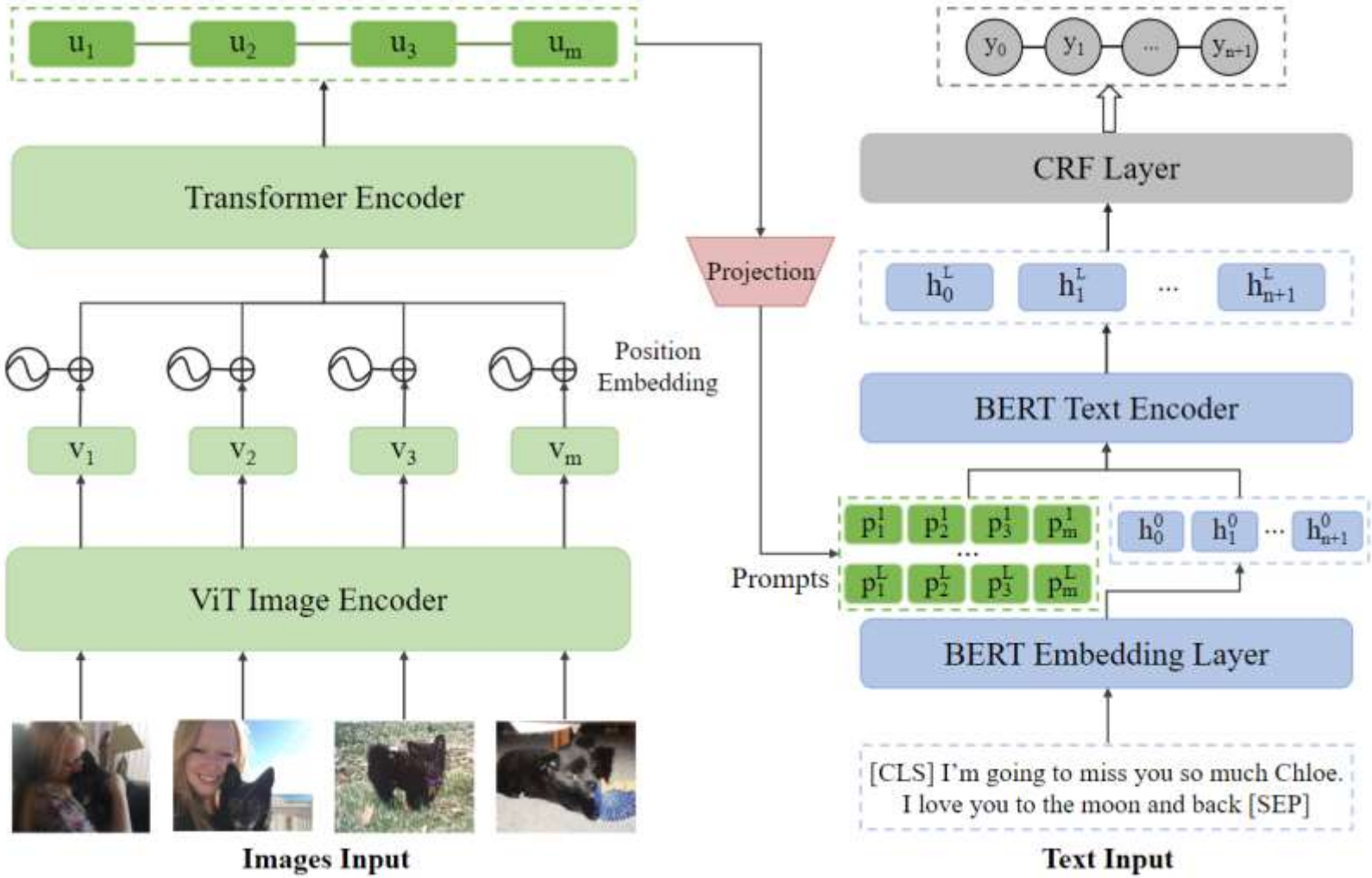
LREC-COLING 2024

- Background
- Motivation of Our Work
- Our Datasets
- **Our proposed Framework**
- Experiments
- Conclusions

Our proposed Framework

LREC-COLING 2024

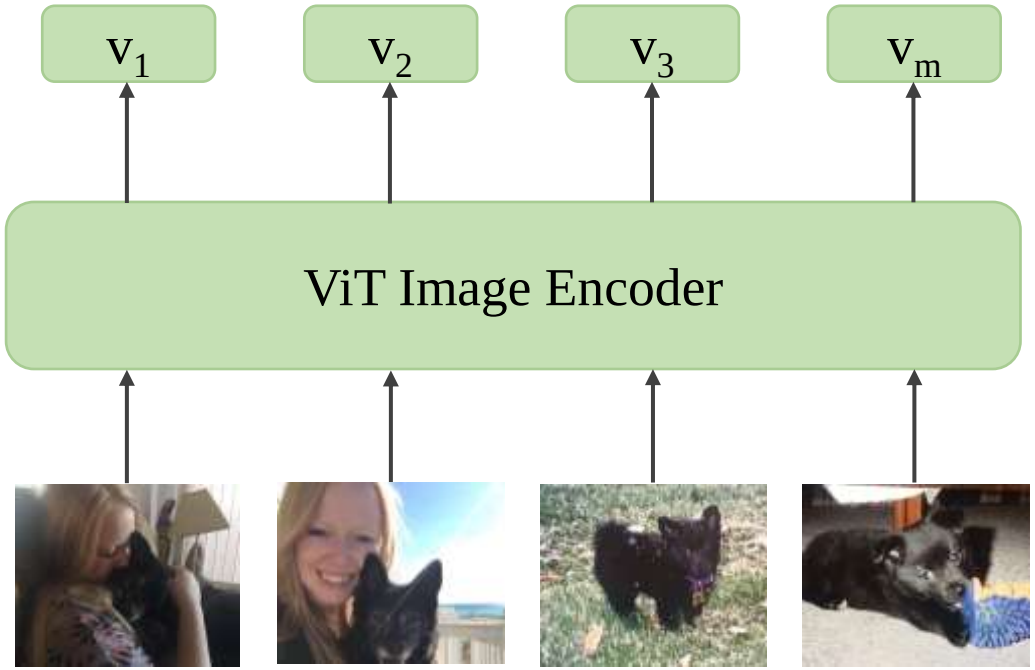
- Overall Architecture of TPM-MI



Our proposed Framework

LREC-COLING 2024

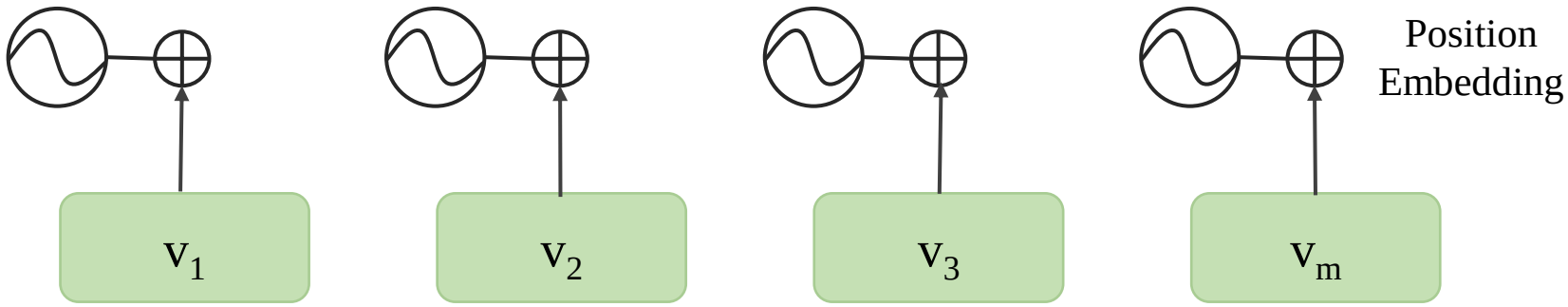
- TPM-MI
  - Multi-Image Representation
    - We use ViT as the image encoder for obtaining the representation of each image in m input images $\boldsymbol{V} \in \mathbb{R}^{d_v \times m}$.



Our proposed Framework

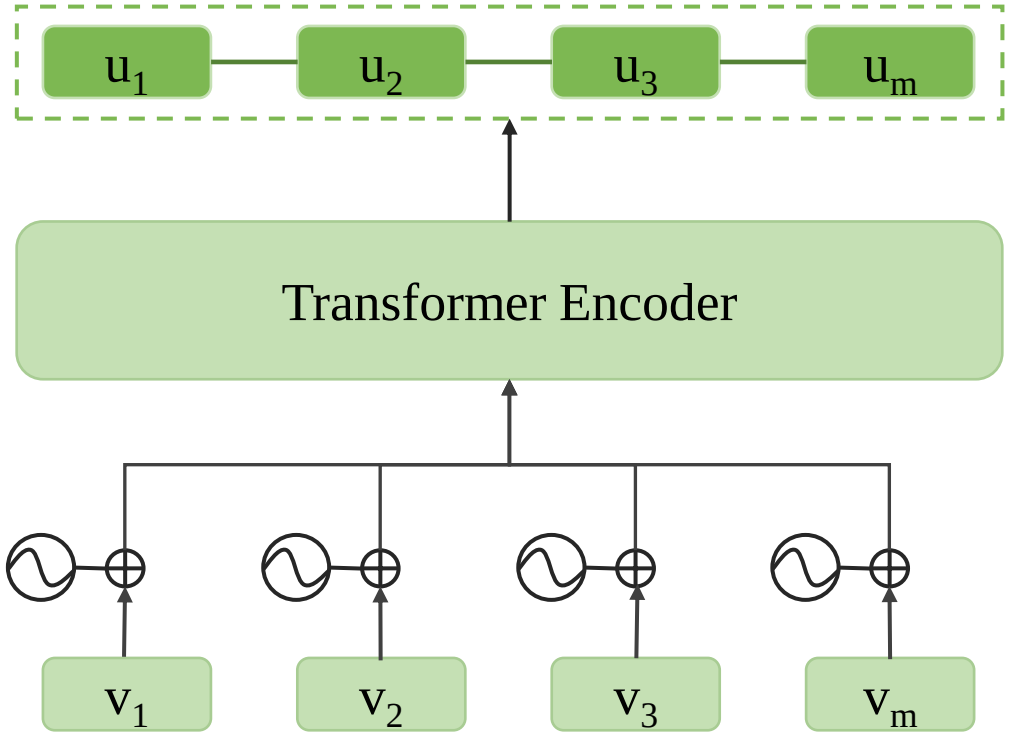
LREC-COLING 2024

- TPM-MI
  - Multi-Image Representation
    - We add the learnable temporal positional encoding $\boldsymbol{T} = (t_1, t_2, \ldots, t_m)$ *onto* $\mathbf{V}$: $\boldsymbol{C} = \boldsymbol{V} + \boldsymbol{T}$, where $t_i \in V \in \mathbb{R}^{d_v}$.
    - This is similar to positional encoding in text encoders such as BERT, which can indicate the position of an image (e.g. whether it is the first or second image)



Our proposed Framework

- TPM-MI
  - Multi-Image Representation
    - To establish relationships between the multi images for a more global image representation, we feed $\boldsymbol{C}$ into a Transformer Encoder to obtain the multi-images representation $\boldsymbol{U} = (u_1, u_2, \ldots, u_m) \in \mathbb{R}^{d_v \times m}$.



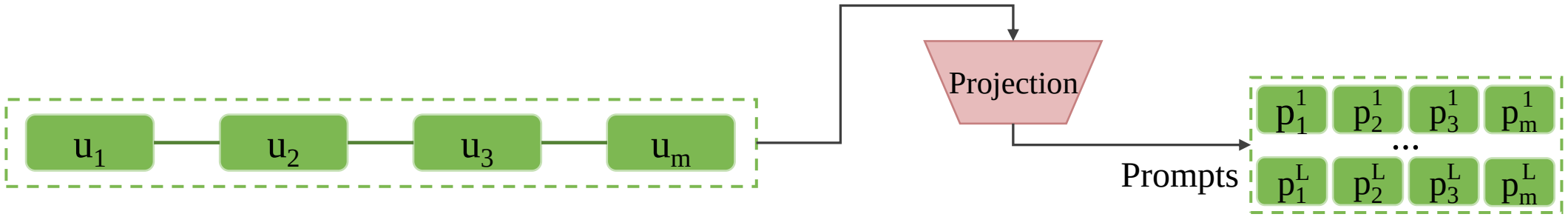
Our proposed Framework

LREC-COLING 2024

- TPM-MI
  - Projection
    - To interact the image with the text, we project the image as prompts (Some pseudo-words in the text):

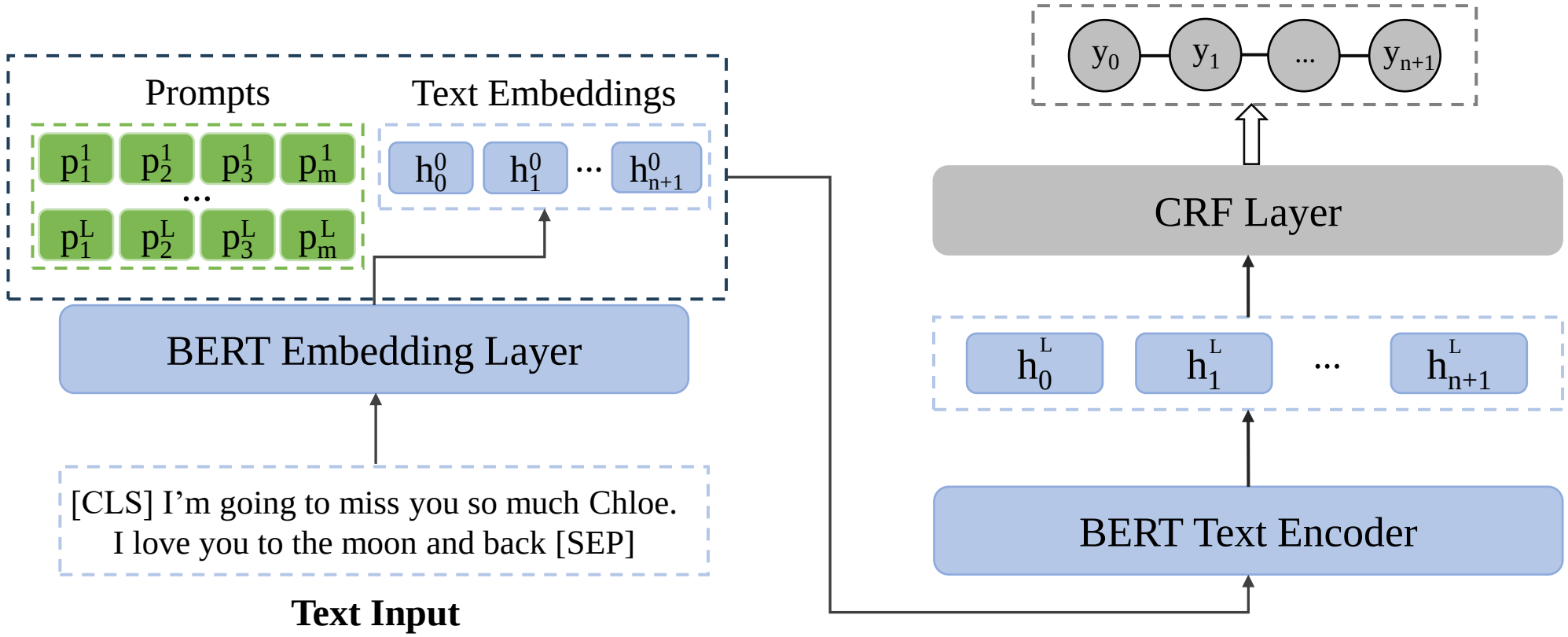
$$\boldsymbol{P^l} = \boldsymbol{W_p^l U}, 1 \leq l \leq L$$

where L is the number of layers of text encoder, $P^l \in \mathbb{R}^{d_t \times m}$ it the prompts corresponding the l-th layer, $\boldsymbol{W_p^l} \in \mathbb{R}^{d_t \times d_v}$ is the weight metric, d_t is the dimension of the text representation.



Our proposed Framework

- TPM-MI
  - Text Representation
    - We use BERT as text encoder and feed prompts and text into it to obtain the final text representation and use CRF as a decoder to predict the result:



Outline

LREC-COLING 2024

- Background
- Motivation of Our Work
- Our Datasets
- Our proposed Framework
- **Experiments**
- Conclusions

Experiments

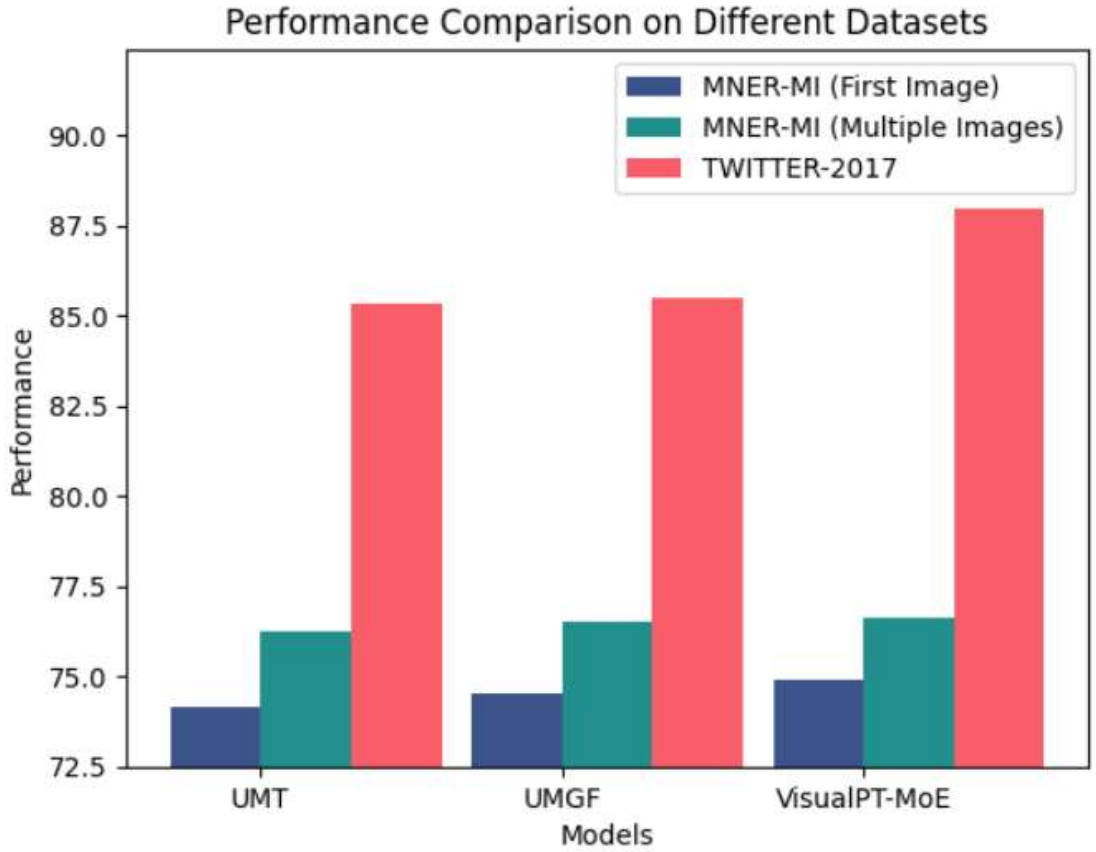
LREC-COLING 2024

- Main Result

Modality	Model	MNER-MI			MNER-MI-Plus		
		P	R	F1	P	R	F1
Text Only	BiLSTM-CRF	64.03	65.91	64.96	73.65	70.74	72.17
	CNN-BiLSTM-CRF	64.89	66.89	65.87	73.71	71.97	72.83
	GPT4	64.28	67.91	66.05	63.76	69.12	66.33
	HBiLSTM-CRF	64.51	68.55	66.47	72.19	74.34	73.25
	BERT	69.04	73.54	71.22	77.35	79.19	78.26
	BERT-CRF	70.78	75.05	72.85	80.15	78.52	79.33
Text + Single Image	MiniGPT4	59.87	62.37	61.09	62.22	64.27	63.23
	GVATT-HBiLSTM-CRF	67.83	67.19	67.51	76.31	73.11	74.68
	AdaCAN-CNN-BiLSTM-CRF	67.89	68.24	68.06	75.67	73.85	74.75
	OCSGA	75.75	72.04	73.85	81.44	79.13	80.27
	UMT	74.23	74.03	74.13	81.71	79.50	80.59
	MAF	74.91	73.60	74.25	80.17	81.29	80.73
	UMGF	73.74	75.30	74.51	82.31	79.65	80.96
	ITA	74.95	74.21	74.58	79.64	81.46	80.54
	promptMNER	75.80	73.46	74.61	81.13	81.39	81.26
	VisualPT-MoE	74.77	75.01	74.89	82.72	80.64	81.67
	HVPNeT	74.93	75.28	75.10	81.88	80.94	81.41
Text + Multiple Images	UMT-MI	76.56	75.90	76.23	82.26	82.96	82.61
	UMGF-MI	75.88	77.14	76.50	82.55	82.25	82.40
	VisualPT-MoE-MI	76.87	76.38	76.62	82.61	82.79	82.70
	TPM-MI	77.45	77.19	77.32†	83.66	83.18	83.42†

Experiments

- Performance Comparison on Different Datasets
  - Models that perform well on single-image datasets perform poorly on multi-image datasets



Outline

LREC-COLING 2024

- Background
- Motivation of Our Work
- Our Datasets
- Our proposed Framework
- Experiments
- **Conclusions**

Conclusions

LREC-COLING 2024

- Conclusions
  - We propose a multi-image MNER dataset MNER-MI to address the research gaps in MNER as well as to expand the scope of MNER for real-world applications.
  - We establish a comprehensive set of representative baseline model for the challenges of MNER with multiple images.
  - We have conducted extensive experiments to demonstrate that multiple images can provide more information to better help MNER compared to a single image and the effectiveness of our method.

Conclusions

LREC-COLING 2024

- Future Work
  - Although we model multiple images as frames in a video in this paper, we recognize the need for more efficient representations to fully capture the unique characteristics of multiple images.
  - We are aware of the limitations of our approach: we treat each image equally, while in reality, different images have different importance in understanding the post, and we plan to explicitly establish the weight of each image in the future.