CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

Institute of Automation Chinese Academy of Sciences
Center for Research on Intelligent Perception and Computing (CRIPAC)

ICASSP 2024 KOREA

Interpretable Multimodal Out-of-context Detection with Soft Logic Regularization

Huanhuan Ma, Jinghao Zhang, Qiang Liu, Shu Wu, Liang Wang

Background

- **Rapid Information Spread:** Mobile devices and media platforms facilitate the fast dissemination of news, increasing the exposure to false or deceptive content.
- **Misinformation Challenges:** Misinformation, particularly out-of-context news (where images or information are shared in misleading ways), poses serious societal risks.
- **Current Detection Limitations:**
  - Existing methods to identify misleading information often lack transparency.
  - Many current technologies offer limited explanations for their findings, complicating efforts to build trust and understanding.
- **Need for Improved Methods:**
  - There is a crucial need for methods that not only detect misinformation effectively but also provide clear, interpretable reasons for their assessments.
  - Enhancing interpretability can help in educating the public and aiding analysts in combating false information.

The Task

- Image repurposing, also known as out-of-context photos are a powerful low-tech form of misinformation

Brazillian and Colombian boxers take apart a joint training session

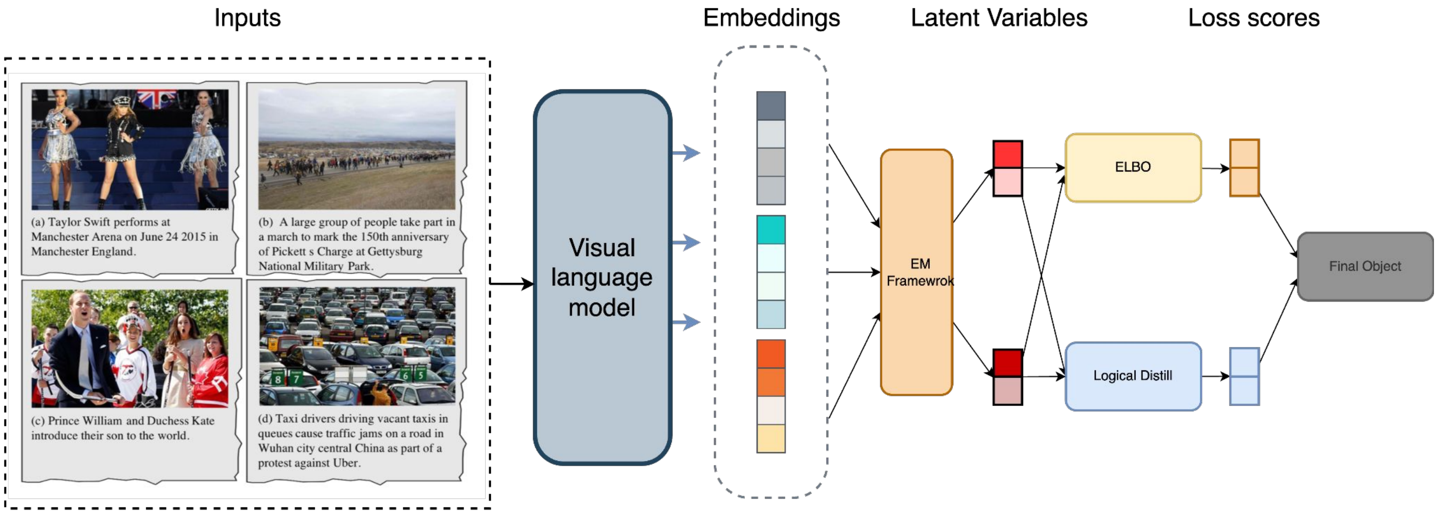
Brazillian and Colombian boxers take apart a joint training session

LOGic **R**egularization for out-of-context **AN**alysis (**LOGRAN**)

- **Caption Detection** Given a caption sentence c and its image I, our goal is to model the probability distribution p(y|c, I), where y ∈ {Pristine, Falsified} is a two-valued variable indicating the veracity of the caption's image.

- **Phrase Detection** We decompose the caption into phrases and predict the out-of-context label z_i for each caption phrase $w_i \in W_c$ using the probability $p(z_i| c, w_i, I)$, where z_i is treated as a binary latent variable $z_i \in$ {Pristine, Falsified}

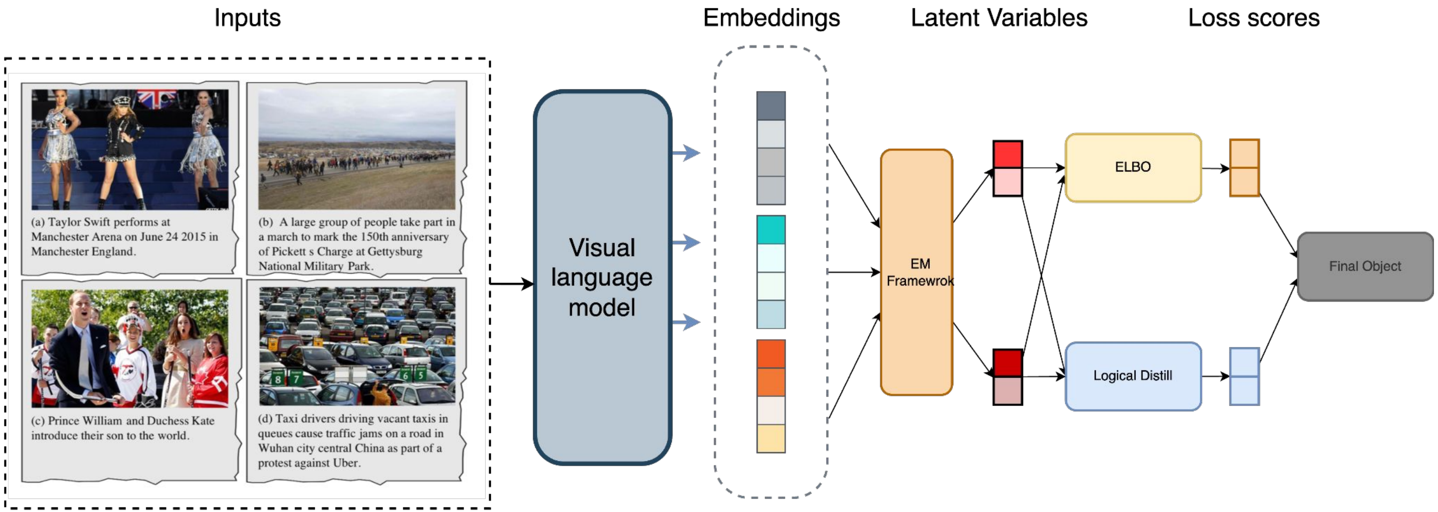
The Framework



Follow the EM framework to model the latent variables

$$p_t(\boldsymbol{y}|x) = \sum_{\boldsymbol{z}} p_t(\boldsymbol{y}|\boldsymbol{z}, x)p(\boldsymbol{z}|x)$$

The Framework



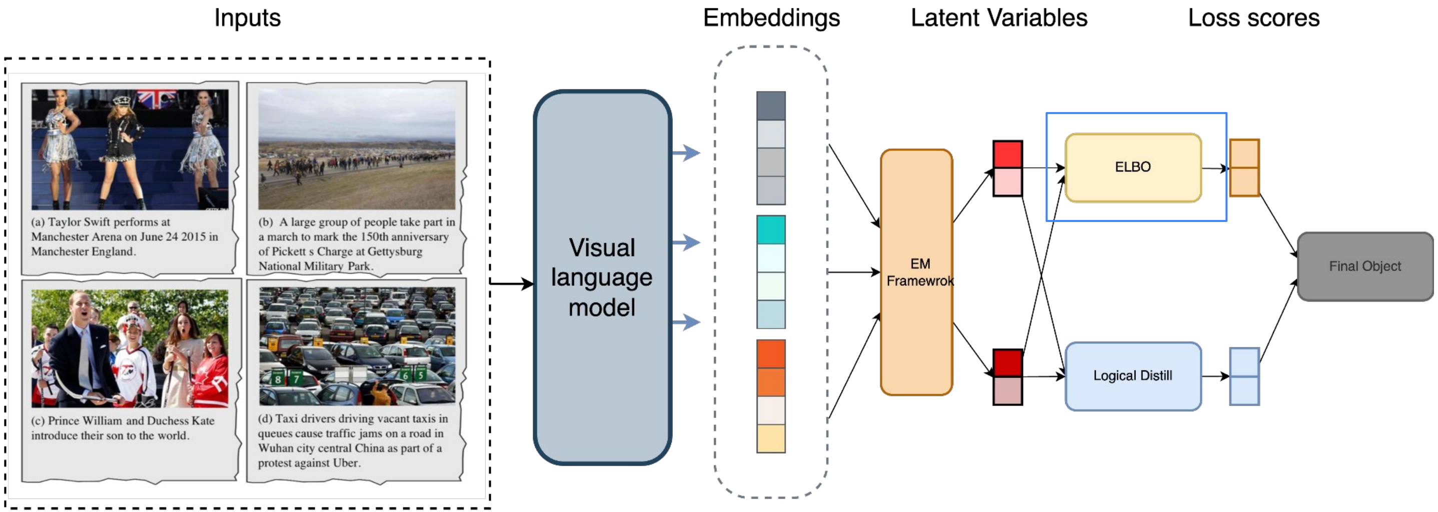
Follow the EM framework to model the latent variables

$$p_t(\boldsymbol{y}|x) = \sum_{\boldsymbol{z}} p_t(\boldsymbol{y}|\boldsymbol{z}, x)p(\boldsymbol{z}|x)$$

Weak supervise learning:

- **ELBO**
- **Logical regularization**

The Framework

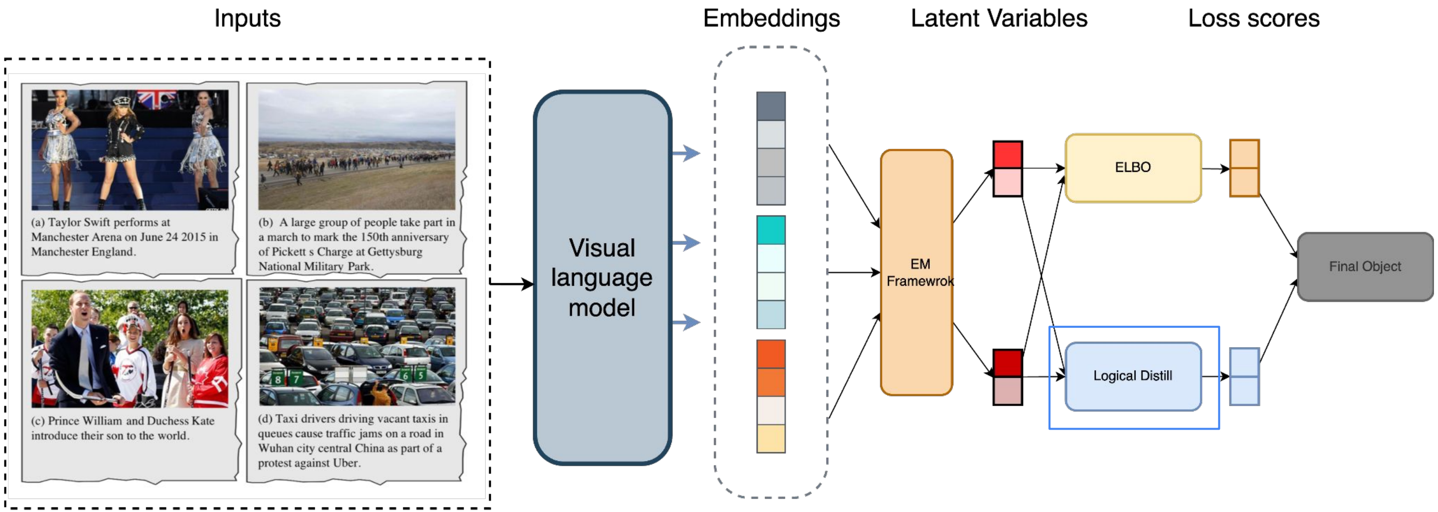


- **ELBO loss:**

$$\mathcal{L}_{\text{var}}(t, l):\ \ -\mathbb{E}_{q_l}[\log p_t(y^*|\boldsymbol{z}, x))] + D_{\text{KL}}(q_l(\boldsymbol{z}|\boldsymbol{y}, x) \parallel p(\boldsymbol{z}|x))$$

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

The Framework



The Logical Rule:

A caption is considered: 1) Falsified if there is inconsistency in at least one caption phrase; 2) Pristine if all caption phrases are consistent

Constructing Teacher module:

Projecting the variational distribution

$$q_l(\boldsymbol{z}|\boldsymbol{y}, x) \Rightarrow q_l^{\mathrm{T}}(\boldsymbol{y}_z|\boldsymbol{y}, x)$$

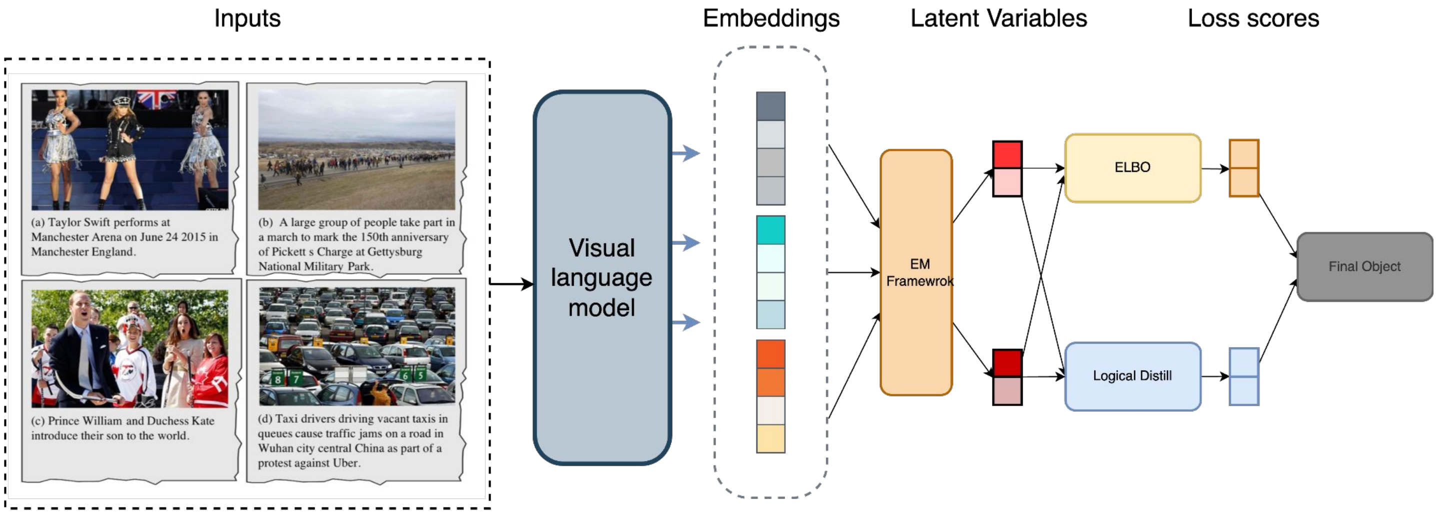
By simulating the outputs of teacher module, transfer logical knowledge into student model $p_t(\boldsymbol{y}|\boldsymbol{z}, x)$

Logical distill loss:

$$\mathcal{L}_{\text{logic}}(t, l) = D_{\text{KL}}\left(p_t(\boldsymbol{y}|\boldsymbol{z}, x) \parallel q_l^{\mathrm{T}}(\boldsymbol{y}_z|\boldsymbol{y}, x)\right)$$

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

The Framework



The final loss function:

$$\mathcal{L}_{\text{final}}(t,l) = (1-\lambda)\mathcal{L}_{\text{var}}(t,l) + \lambda\mathcal{L}_{\text{logic}}(t,l)$$

Experiments

Dataset:

- **NewsCLIPpings** comprising both pristine and falsified images. It employs automation to match captions and images from the VisualNews corpus, offering various subsets based on matching methods.

Backbone model:

- **CLIP** utilizes distinct encoders for processing images and text, which are trained to produce comparable representations for associated concepts.
- **VisualBERT** is another multimodal model that integrates visual and textual information. It includes a sequence of Transformer layers that use self-attention to automatically align components of a given text input with specific regions in a corresponding image input.

Results and Discussion

Table 1. Classification accuracy on the test set for the following models: (I) VisualBERT, (II) VisualBERT with LOGRAN, (III) Multimodal CLIP, and (IV) CLIP with LOGRAN. The underlined portions represent improvements from LOGRAN

	VisualBERT	VisualBERT-LOGRAN	CLIP	CLIP-LOGRAN
(a) Semantics/CLIP Text-Image	55.12	56.88	58.59	59.03
(b) Semantics/CLIP Text-Text	53.47	55.62	68.36	70.81
(c) Person/SBERT-WK Text-Text	63.32	65.27	66.57	71.42
(d) Scene/ResNet Place	60.72	62.41	69.64	73.14
Merged/Balanced	61.32	63.18	67.27	70.51

Improvement observed in both backbone models, as well as across all sub-test sets.

Results and Discussion

Table 1. Classification accuracy on the test set for the following models: (I) VisualBERT, (II) VisualBERT with LOGRAN, (III) Multimodal CLIP, and (IV) CLIP with LOGRAN. The underlined portions represent improvements from LOGRAN

	VisualBERT	VisualBERT-LOGRAN	CLIP	CLIP-LOGRAN
(a) Semantics/CLIP Text-Image	55.12	<u>56.88</u>	58.59	<u>59.03</u>
(b) Semantics/CLIP Text-Text	53.47	<u>55.62</u>	68.36	<u>70.81</u>
(c) Person/SBERT-WK Text-Text	63.32	<u>65.27</u>	66.57	<u>71.42</u>
(d) Scene/ResNet Place	60.72	<u>62.41</u>	69.64	<u>73.14</u>
Merged/Balanced	61.32	<u>63.18</u>	67.27	<u>70.51</u>

Improvement observed in both backbone models, as well as across all sub-test sets.

- **VisualBERT vs VisualBERT-LOGRAN**
  The average improvement is 2%

Results and Discussion

Table 1. Classification accuracy on the test set for the following models: (I) VisualBERT, (II) VisualBERT with LOGRAN, (III) Multimodal CLIP, and (IV) CLIP with LOGRAN. The underlined portions represent improvements from LOGRAN

	VisualBERT	VisualBERT-LOGRAN	CLIP	CLIP-LOGRAN
(a) Semantics/CLIP Text-Image	55.12	<u>56.88</u>	58.59	<u>59.03</u>
(b) Semantics/CLIP Text-Text	53.47	<u>55.62</u>	68.36	<u>70.81</u>
(c) Person/SBERT-WK Text-Text	63.32	<u>65.27</u>	66.57	<u>71.42</u>
(d) Scene/ResNet Place	60.72	<u>62.41</u>	69.64	<u>73.14</u>
Merged/Balanced	61.32	<u>63.18</u>	67.27	<u>70.51</u>

Improvement observed in both backbone models, as well as across all sub-test sets.

- **VisualBERT vs VisualBERT-LOGRAN**
  The average improvement is 2%
- **CLIP vs CLIP-LOGRAN**
  The average improvement is 3%

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

Case study

We can easily identify the ‘Culprit’ in each case:

- New York Paris Tokyo
- Brazillian and Colombian boxers

which provides some level of interpretability

CRIPAC
智能感知与计算研究中心
Center for Research on Intelligent Perception and Computing

Conclusion

- We proposed a novel frame work for out-of-context detection named **LOG**ic **R**egularization for out-of-context **AN**alysis (**LOGRAN**)
- Decomposes detection task from caption level to phrase level. Utilizes **latent variables** within an EM framework to predict out-of-context label for each phrase
- Implements two weak supervision methods: **ELBO loss** and **logical rule regularization**
- Conducted experiments on **NewsCLIPpings** dataset using **VisualBERT** and **CLIP** backbone models. Achieved **overall performance improvement**. Provides **phrase-level predictions** for **enhanced interpretability**

Thank you!