

# *Few-Shot Relation Extraction with Hybrid Visual Evidence*

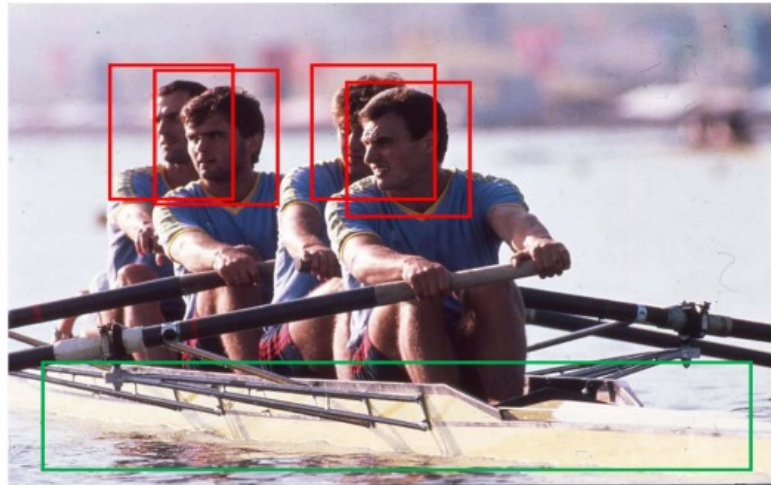
Jiaying Gong and Hoda Eldardiry\*

Virginia Tech, USA

LREC-COLING 2024

# Introduction

Relation extraction aims to predict the relation between two name entities in a sentence.



**Dimitrie Popescu** (born 10 September 1961 in Straja) is a retired Romanian **rower**.

Detected Objects:  
**person**, **boat**

Relation: <**Dimitrie Popescu**, **sport**, **rower**>



# *Related Works*



- Few-shot Relation Extraction:
  1. Only using plain text data
    - Prototypical Networks
    - Siamese Neural Networks
  2. Using external data sources
    - Relation Information
    - Concepts of Entities
    - Side Information
    - External Datasets
    - Graphs



# *Related Works*



- Few-shot Relation Extraction:

1. Only using plain text data

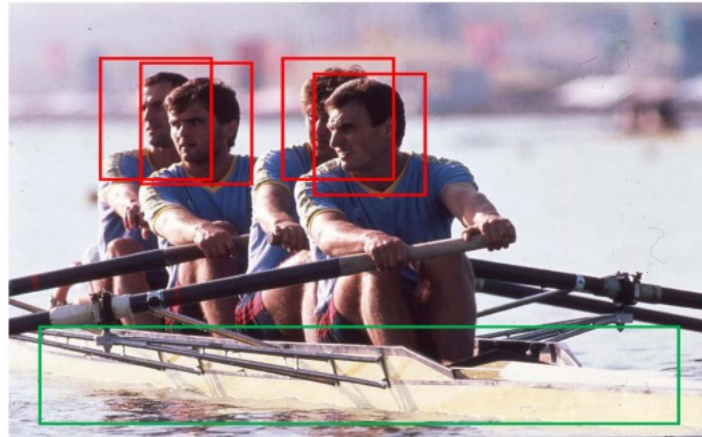
- Prototypical Networks
- Siamese Neural Networks

2. Using external data sources

- Relation Information
- Concepts of Entities
- Side Information
- External Datasets
- Graphs

**Single Modality**

# Motivation



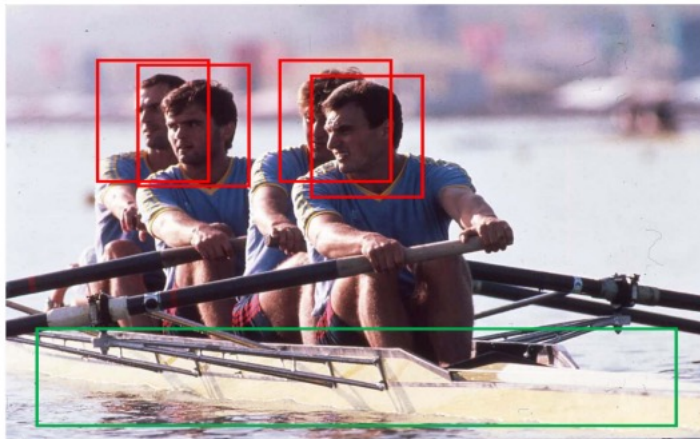
**Dimitrie Popescu** (born 10 September 1961 in Straja) is a retired Romanian **rower**.

Detected Objects:  
**person**, **boat**

Relation: <**Dimitrie Popescu**, **sport**, **rower**>

Figure 1: An example of multi-modal relation extraction based on visual information.

# Challenges



**Dimitrie Popescu** (born 10 September 1961 in Straja) is a retired Romanian **rower**.

Detected Objects:  
**person**, **boat**

Relation: <**Dimitrie Popescu**, **sport**, **rower**>

Figure 1: An example of multi-modal relation extraction based on visual information.

## Challenges

1. Simply concatenating textual and visual features without considering semantic information may even have a negative impact on the performance.
2. Existing multi-modal models mainly focus on fusing global visual features with text without considering the semantic information of visual objects in images.

**Can visual information be a good external source to supplement the missing contexts in textual sentences for few-shot relation extraction?**

# Problem Definition

- Few-shot Learning: The N-way-K-shot setting means N classes with K examples of each. Typically K is no more than 10. There is no overlap between the classes in training data and testing data.

- Input:  $(x_i, h_i, t_i, y_i, r_i)$   
sentence ↑  $x_i$  ↑ tail entity ↑  $t_i$  ↑ relation ↑  $r_i$   
 $h_i$  ↓ head entity ↓  $y_i$  ↓ image ↓

# Model Overview

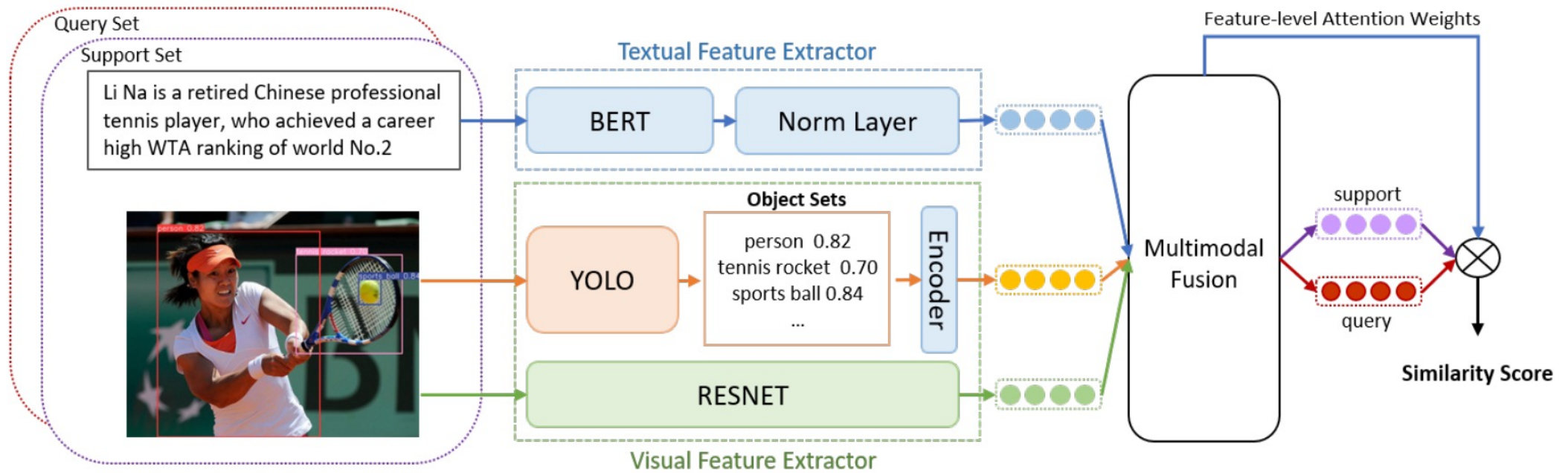


Figure 2: The overview of MFS-HVE. Details of multi-modal fusion is introduced in Sec. 3.3 and Figure 3

# Multimodal Fusion

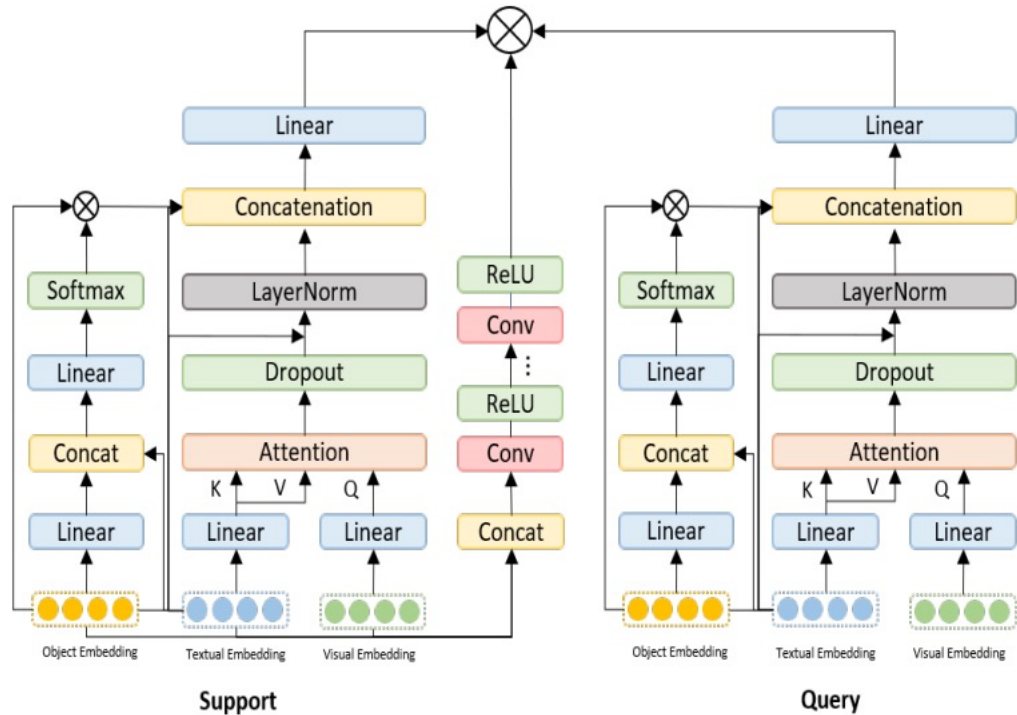


Image-Guided Attention  
Object-Guided Attention  
Hybrid Feature Attention

Figure 3: Detailed structure of multi-modal fusion.

# Model Training

Multi-modal Representation:  $L_{multi} = \tanh(W_{multi} \cdot (r_t \oplus \hat{r}_o \oplus \hat{r}_i) + b_{multi})$

Multi-modal Prototype: 
$$P_{multi}(S) = \frac{1}{K} \sum_{i=1}^K L_{multi}$$

Probability of query instance: 
$$Pr(y = r_i | q) = \frac{\exp(-d((L_{multi}), P_m(S)))}{\sum_{i=1}^{|R|} \exp(-d((L_{multi}), P_i(S)))}$$

Distance Function: 
$$d(s_1, s_2) = \alpha_i \cdot \cosh^{-1} \left( 1 + 2 \frac{\|s_1 - s_2\|^2}{(1 - \|s_1\|^2)(1 - \|s_2\|^2)} \right)$$



# Datasets



Table 1: The statistics of each dataset.

|                         | #instances | #relations | avg. len. |
|-------------------------|------------|------------|-----------|
| MNRE                    | 15,484     | 23         | 16.67     |
| FewRel                  | 56,000     | 80         | 24.95     |
| FewRel <sub>small</sub> | 3,703      | 80         | 23.90     |

# Main Results

Table 3: Results of Accuracy Comparison Among Models (%) on MNRE and FewRel<sub>small</sub> Datasets.

| Modality    | Model   | MNRE            |                  | FewRel <sub>small</sub> |                 |                  |                  |
|-------------|---|-----------------|------------------|-------------------------|-----------------|------------------|------------------|
|             |   | 5-Way<br>1-Shot | 10-Way<br>1-Shot | 5-Way<br>1-Shot         | 5-Way<br>5-Shot | 10-Way<br>1-Shot | 10-Way<br>5-Shot |
| Only Text   | GNN (Satorras and Estrach, 2018)              | 29.08           | 22.53            | 46.38                   | 70.45           | 28.74            | 62.07            |
|             | Snail (Mishra et al., 2018)                   | 30.90           | 19.43            | 40.16                   | 60.07           | 21.19            | 47.56            |
|             | Siamese (Koch et al., 2015)                   | 36.08           | 26.50            | 62.74                   | 73.92           | 42.17            | 65.05            |
|             | MLMAN (Ye and Ling, 2019)                     | 35.08           | 29.06            | 63.47                   | 74.47           | 61.86            | 72.58            |
|             | Proto_BERT (Snell et al., 2017)               | 49.75           | 33.57            | 75.64                   | 84.64           | 64.17            | 75.27            |
|             | MTB (Baldini Soares et al., 2019)             | 46.02           | 32.35            | 76.38                   | 86.27           | 65.27            | 73.81            |
| Text+Others | ZSLRC (Gong and Eldardiry, 2021)              | 45.65           | 32.23            | 71.82                   | 81.74           | 64.88            | 71.81            |
|             | ConceptFERE (Yang et al., 2021)               | -               | -                | 75.86                   | 83.38           | 68.38            | 76.06            |
|             | REGRAB (Qu et al., 2020)                      | -               | -                | 78.53                   | 84.96           | 70.65            | 78.00            |
|             | HCRP (Han et al., 2021)                       | 31.10           | 10.45            | 78.04                   | 84.68           | 69.54            | 77.91            |
|             | MapRE (Dong et al., 2021)                     | 51.92           | 35.20            | 79.44                   | 85.60           | 70.71            | 78.84            |
|             | GM_GEN (Li and Qian, 2022)                    | 52.58           | 35.82            | 60.04                   | 73.74           | 42.22            | 59.23            |
|             | FAEA (Dou et al., 2022)                       | 52.14           | 33.37            | 80.80                   | 87.94           | 71.30            | 79.29            |
|             | SimpleFSRE (Liu et al., 2022b)                | 50.32           | 35.05            | 80.84                   | 87.46           | <b>71.67</b>     | 80.14            |
| Text+Image  | Concat (Wan et al., 2021a)                    | 40.17           | 29.83            | 74.10                   | 84.69           | 66.08            | 75.95            |
|             | CirculantFusion (Gong et al., 2023)           | 38.39           | 29.19            | 73.21                   | 83.58           | 65.11            | 76.29            |
|             | DeepFusion (Wang et al., 2020a)               | 48.27           | 33.28            | 78.38                   | 86.76           | 66.36            | 76.08            |
|             | Proto <sub>multimodal</sub> (Ni et al., 2022) | 50.84           | 34.10            | 77.18                   | 86.28           | 68.19            | 78.29            |
|             | Dual Co-Att (Liu et al., 2021)                | 52.52           | 35.62            | 77.60                   | 87.24           | 68.69            | 78.54            |
|             | <b>MFS-HVE</b>                                | <b>54.88</b>    | <b>36.62</b>     | <b>81.32</b>            | <b>89.65</b>    | 69.52            | <b>80.55</b>     |

# Robustness

Table 5: Results of performance decrease in Accuracy(%) from FewRel to FewRel<sub>small</sub>.

| Model   | 5-Way<br>1-Shot | 5-Way<br>5-Shot | 10-Way<br>1-Shot | 10-Way<br>5-Shot |
|---|-----------------|-----------------|------------------|------------------|
| GNN (Satorras and Estrach, 2018)              | 12.32           | 10.90           | 12.18            | 6.53             |
| Snail (Mishra et al., 2018)                   | 9.88            | 12.12           | 11.19            | 11.58            |
| Siamese (Koch et al., 2015)                   | 5.61            | 8.36            | 14.24            | 4.25             |
| MLMAN (Ye and Ling, 2019)                     | 3.30            | 1.97            | 2.70             | 2.25             |
| Proto_BERT (Snell et al., 2017)               | 2.20            | 4.31            | 3.11             | 7.35             |
| MTB (Baldini Soares et al., 2019)             | 3.14            | 1.00            | 3.54             | 3.66             |
| ZSLRC (Gong and Eldardiry, 2021)              | 4.01            | 6.10            | 2.34             | 5.83             |
| ConceptFERE (Yang et al., 2021)               | 3.56            | 2.96            | 3.34             | 3.76             |
| REGRAB (Qu et al., 2020)                      | 4.32            | 4.88            | 3.44             | 4.07             |
| HCRP (Han et al., 2021)                       | 4.36            | 3.00            | 2.76             | 4.24             |
| MapRE (Dong et al., 2021)                     | 6.29            | 7.24            | 8.47             | 8.80             |
| FAEA (Dou et al., 2022)                       | 7.97            | 6.78            | 4.55             | 5.59             |
| SimpleFSRE (Liu et al., 2022b)                | 5.45            | 7.45            | 5.79             | 7.54             |
| Concat (Wan et al., 2021a)                    | 3.08            | 1.32            | 2.58             | 0.83             |
| DeepFusion (Wang et al., 2020a)               | 2.14            | 4.72            | 0.38             | <b>0.39</b>      |
| CirculantFusion (Gong et al., 2023)           | 3.99            | 2.60            | 5.75             | 2.24             |
| Dual Co-Att (Liu et al., 2021)                | 2.60            | 1.58            | 3.67             | 2.02             |
| Proto <sub>multimodal</sub> (Ni et al., 2022) | 2.01            | 3.08            | 3.75             | 2.99             |
| <b>MFS-HVE</b>                                | <b>1.95</b>     | <b>0.83</b>     | <b>0.27</b>      | 1.32             |

# Ablation Study

Table 4: Ablation study over MFS-HVE components (%) on MNRE and FewRel<sub>small</sub> datasets.

| Model Component        | MNRE            |                  | FewRel <sub>small</sub> |                 |                  |                  |
|------------------------|-----------------|------------------|-------------------------|-----------------|------------------|------------------|
|                        | 5-Way<br>1-Shot | 10-Way<br>1-Shot | 5-Way<br>1-Shot         | 5-Way<br>5-Shot | 10-Way<br>1-Shot | 10-Way<br>5-Shot |
| Only Text              | 49.39           | 31.95            | 76.66                   | 85.82           | 63.54            | 76.73            |
| Image Attention        | 50.43           | 32.40            | 78.37                   | 86.75           | 66.28            | 77.18            |
| Object Attention       | 50.57           | 33.63            | 78.85                   | 86.24           | 66.96            | 77.96            |
| Image&Object Attention | 52.26           | 35.38            | 80.50                   | 88.72           | 69.49            | 79.17            |
| <b>MFS-HVE</b>         | <b>54.88</b>    | <b>36.62</b>     | <b>81.32</b>            | <b>89.65</b>    | <b>69.52</b>     | <b>80.55</b>     |

# Case Study



(a)

Kit Harington (**Jon Snow**) and **Rose Leslie** are getting **married**.

Detected Objects: **person**, **person**  
Ground Truth: <Jon Snow, couple, Rose Leslie>  
Text-based Model: < Jon Snow, couple, Rose Leslie > ✓  
Our MFS-HVE Model: < Jon Snow, couple, Rose Leslie > ✓



(b)

Congratulations to **Angela** and **Mark Salmons**!

Detected Objects: **person**, **person**  
Ground Truth: <Angela, couple, Mark Salmons>  
Text-based Model: <Angela, peer, Mark Salmons> ✗  
Our MFS-HVE Model: <Angela, couple, Mark Salmons> ✓



(c)

**Rabin**, Arafat and Israeli Foreign Minister Shimon Peres were **awarded** the 1994 **Nobel Peace Prize**.

Ground Truth: <Rabin, winner, Nobel Peace Prize>  
Text-based Model: <Rabin, winner, Nobel Peace Prize> ✓  
Our MFS-HVE Model: <Rabin, winner, Nobel Peace Prize> ✓



(d)

She is the younger sister of **biathlete** and cross-country skier **Lars Berger**.

Detected Objects: **skis**, **person**  
Ground Truth: <biathlete, sports, Lars Berger>  
Text-based Model: <Biathlete, sibling, Lars Berger> ✗  
Our MFS-HVE Model: <Biathlete, sports, Lars Berger> ✓

Figure 4: The examples of our proposed model MFS-HVE comparing to a text-based model on both the MNRE and FewRel datasets. We present the relation extraction results with the detected objects from the relevant image in the right column. The head entities are highlighted in green, whereas the tail entities are highlighted in red.

# Conclusion and Future Work

- We propose MFS-HVE, a multi-modal few-shot relation extraction approach leveraging semantic visual information to supplement the missing contexts in textual sentences. Our multimodal fusion module consists of image-guided attention, object-guided attention, and hybrid feature attention that integrates information from different modalities.
- Future Work:
  1. We will implement other powerful SOTA image encoders such as ViT to generate feature-level image embeddings.
  2. We will explore utilizing the semantic visual information as an external source in zero-shot learning.