GMNER: Grounded Multimodal Named Entity Recognition on Social Media

Giuseppe Frigeni

Computer Vision 2024/25

Outline

- Problem Statement
- Datasets and Metrics
- Experimental Setup
- Ablation Study
- Compression Techniques
- Conclusion and Future Work

GMNER

Text

Michael Jordan is now in Toronto, giving his distinguished lecture at the Fields Institute.

Image

- Michael Jordan, PER
- the Fields Institute, ORG
- Toronto, LOC

(a) Mulitmodal NER

Text

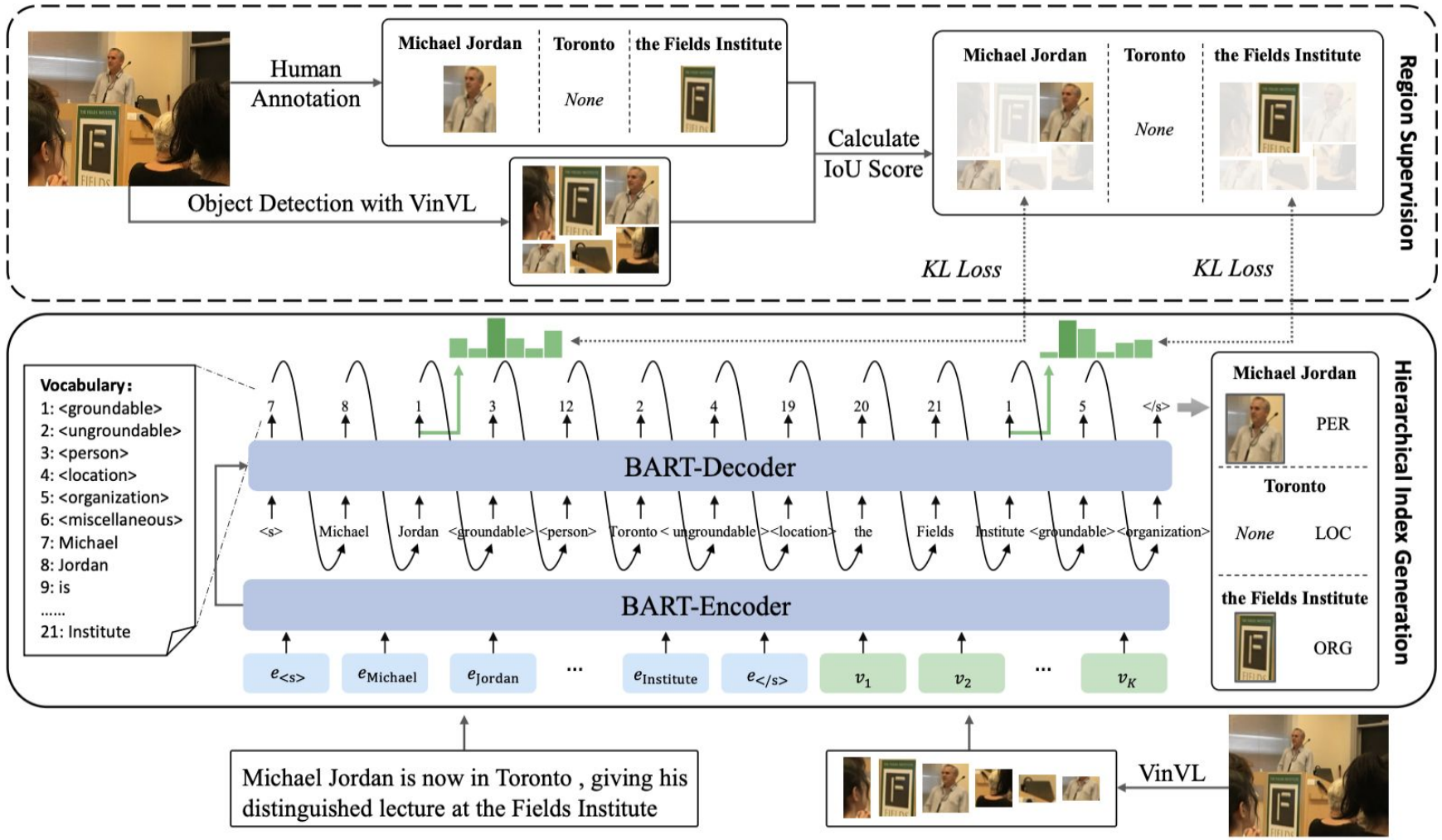
Michael Jordan is now in Toronto, giving his distinguished lecture at the Fields Institute.

Image

- Michael Jordan, PER,
- the Fields Institute, ORG,
- Toronto, LOC, None

(b) Grounded Mulitmodal NER

H-Index Framework



H-Index Framework

$$\mathbf{H}^e = [\mathbf{H}^e_{\mathrm{T}}; \mathbf{H}^e_{\mathrm{V}}] = \mathrm{Encoder}([\mathbf{T}; \mathbf{V}]),$$

$$\mathbf{h}_i = \mathrm{Decoder}(\mathbf{H}^e; \boldsymbol{y}_{<i}),$$
$$\bar{\mathbf{H}}^e_{\mathrm{T}} = \big(\mathbf{T} + \mathrm{MLP}(\mathbf{H}^e_{\mathrm{T}})\big)/2,$$
$$p(\boldsymbol{y}_i) = \mathrm{Softmax}([\mathbf{C}; \bar{\mathbf{H}}^e_{\mathrm{T}}] \cdot \mathbf{h}_i),$$

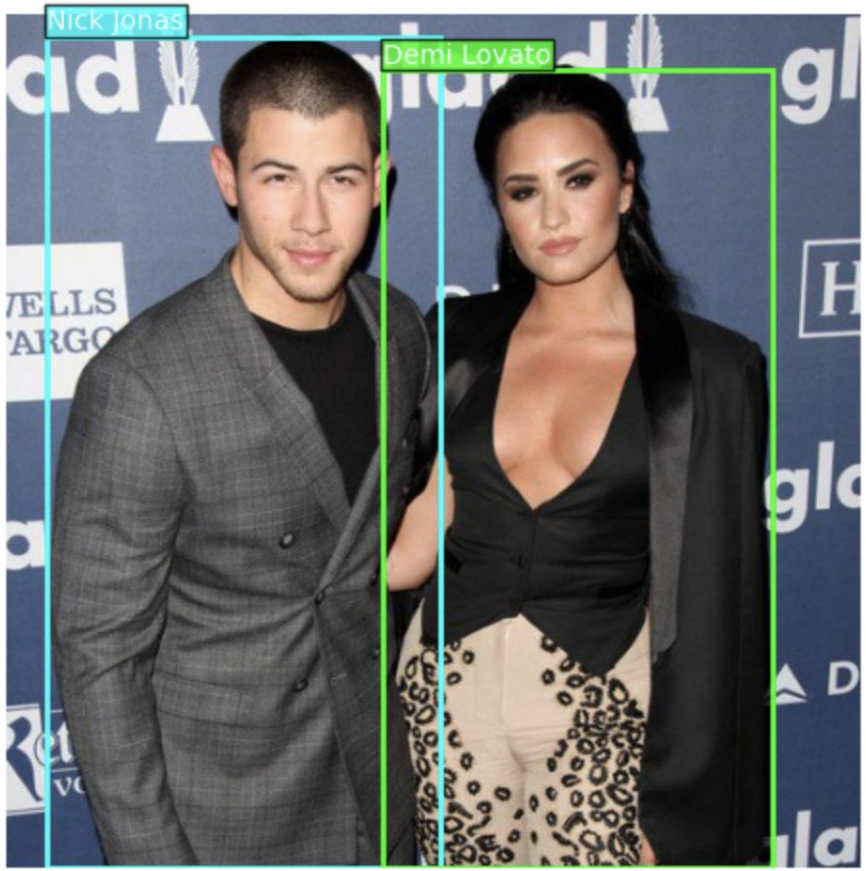
$$\mathcal{L}^T = -\frac{1}{NM}\sum_{j=1}^{N}\sum_{i=1}^{M}\log p(\boldsymbol{y}^j_i),$$

$$\bar{\mathbf{H}}^e_{\mathrm{V}} = \big(\mathbf{V} + \mathrm{MLP}(\mathbf{H}^e_{\mathrm{V}})\big)/2,$$
$$p(\boldsymbol{z}_k) = \mathrm{Softmax}(\bar{\mathbf{H}}^e_{\mathrm{V}} \cdot \mathbf{h}_k).$$

$$\mathcal{L}^V = \frac{1}{NE}\sum_{j=1}^{N}\sum_{k=1}^{E} g(\boldsymbol{z}^j_k)\log\frac{g(\boldsymbol{z}^j_k)}{p(\boldsymbol{z}^j_k)},$$

Dataset

Split	#Tweet	#Entity	#Groundable Entity	#Box
Train	7,000	11,782	4,694	5,680
Dev	1,500	2,453	986	1,166
Test	1,500	2,543	1,036	1,244
Total	10,000	16,778	6,716	8,090



Example of tweet:
“Nick Jonas feared pal Demi Lovato would suffer drug death”

Evaluation metrics

- **Precision**: the ratio of the correct predictions over the total number of positives predicted

$$precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

- **Recall**: the ratio of correct predictions over the total number of real positives

$$recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

- **F1-score**: the Harmonic mean between Precision and Recall

$$F1\ Score = 2 \times \frac{recall \times precision}{recall + precision}$$

Experimental Setup

Before:

Requirement

- pytorch 1.7.1
- transformers 3.4.0
- fastnlp 0.6.0

```
from transformers.modeling_bart import *
# from transformers.models.bart.modeling_bart import *
```

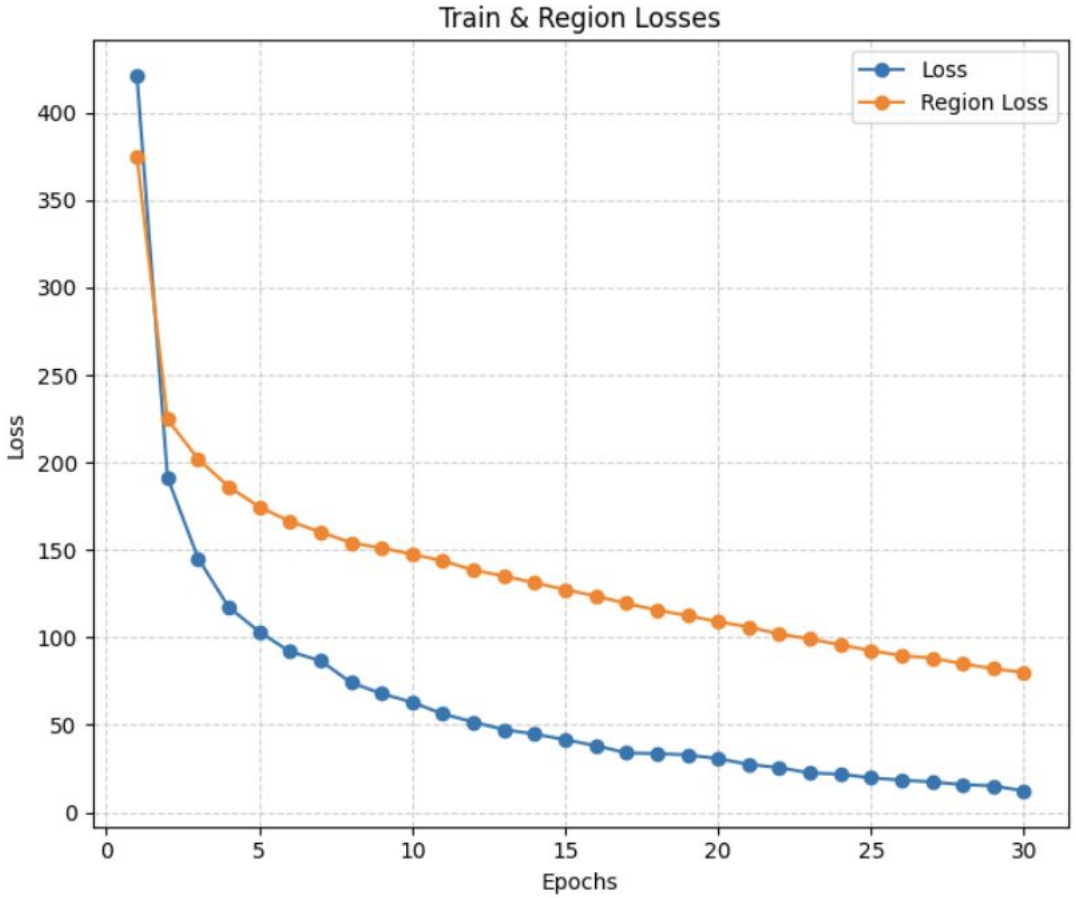
After:

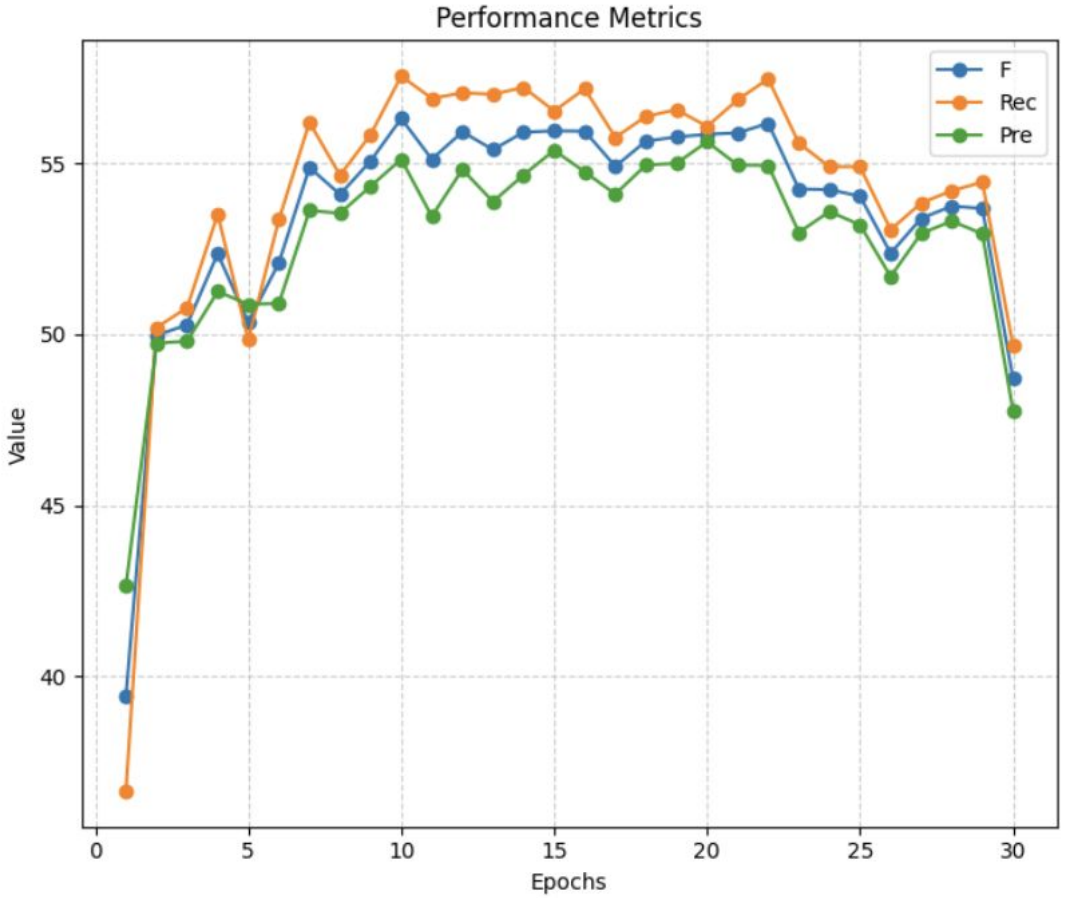
Requrirements

- pytorch 2.5.1
- transformers 4.51.1

```
from transformers.activations import ACT2FN
from transformers import PreTrainedModel, BartConfig
```

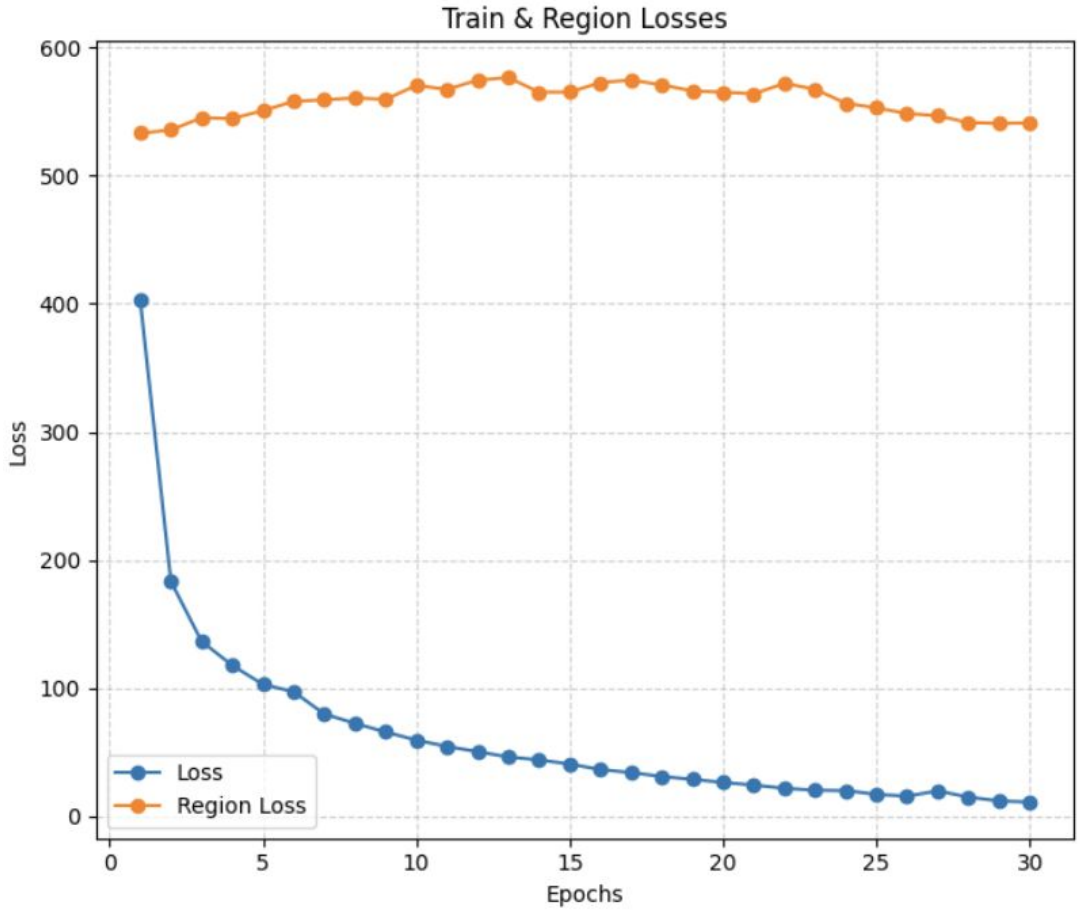
Results

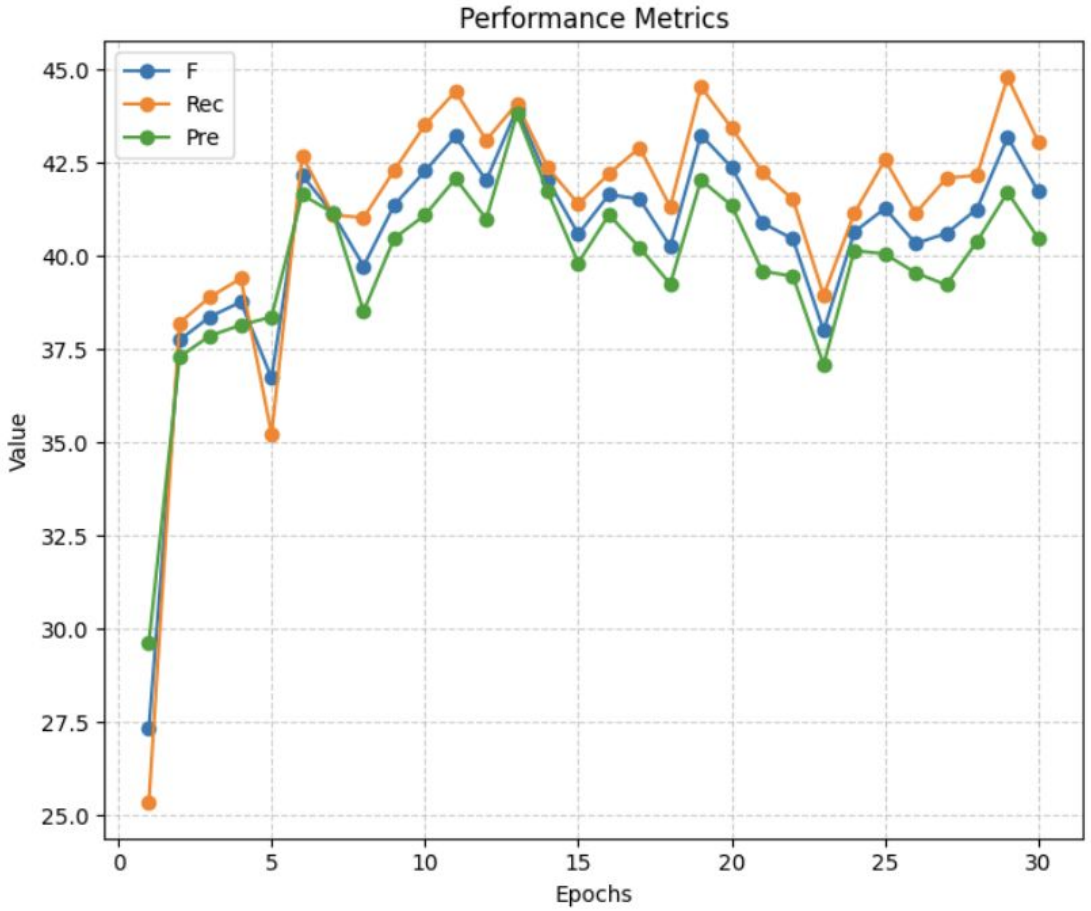




best validation metrics: {'f1': 56.31, 'rec': 57.55, 'pre': 55.12}

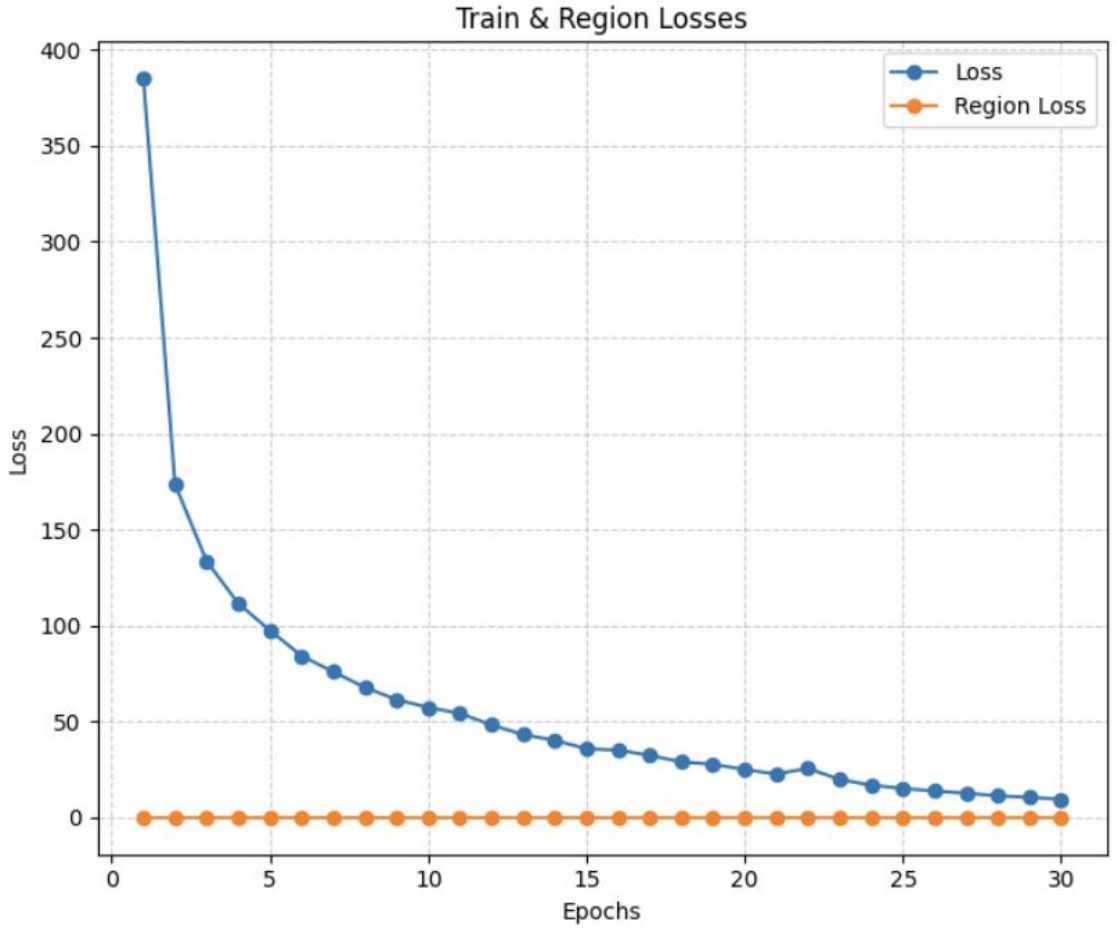
Ablation Study: zeroing the image feature

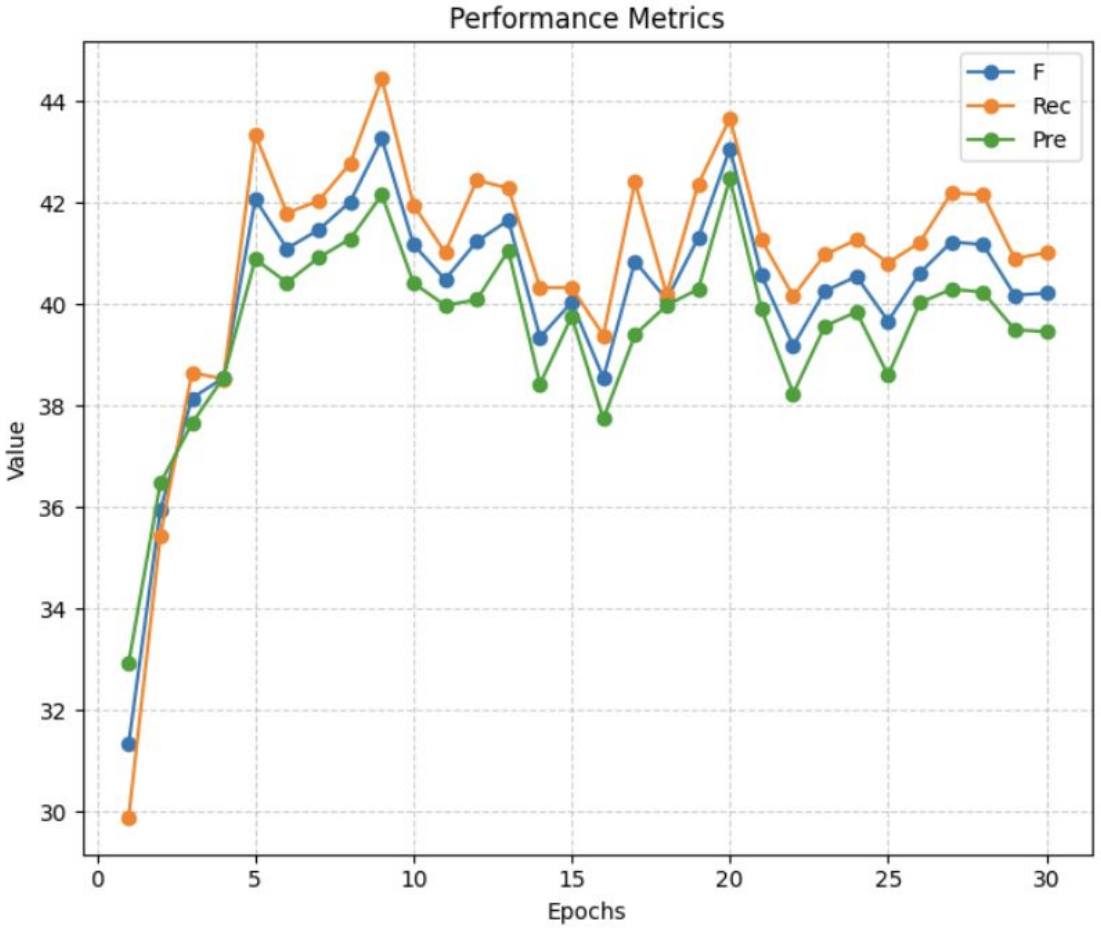




best validation metrics: {'f1': 43.94, 'rec': 44.08, 'pre': 43.8}

Ablation Study: disabling multimodality





best validation metrics: {'f1': 43.27, 'rec': 44.45, 'pre': 42.16}

Compression Techniques: Quantization Aware Training

Quantization-aware training (QAT) is a technique used during the training of machine learning models, to simulate the effects of quantization (converting model weights and activations to lower precision formats like INT8)

```
--- Size Comparison (.pth files) ---
Original size: 545.81 MB
Quantized size: 254.51 MB
Size reduction: 291.30 MB
Size reduction percentage: 53.37%
Size ratio (Original / Quantized): 2.14x
```

Metrics during training: {'f1': 54.0, 'rec': 54.61, 'pre': 53.39}

Metrics after converting: {'f1': 6.51, 'rec': 4.65, 'pre': 10.84}

Compression Techniques: Dynamic Quantization

Dynamic quantization is a method of model compression where model weights are quantized offline (before runtime) and activations are quantized dynamically during inference

```
--- Size Comparison (.pth files) ---
Original size: 545.79 MB
Quantized size: 252.78 MB
Size reduction: 293.02 MB
Size reduction percentage: 53.69%
Size ratio (Original / Quantized): 2.16x
```

Validation Metrics {'f1': 53.36, 'rec': 51.35, 'pre': 55.54}

Time per task (CPU): Original 116.4 ms Quantized 83.3 ms

Conclusions and Future work

- Implemented the model getting rid of outdated requirements
- Conducted ablation study of only text vs text and image in two different ways
- Implemented two ways to compress the model
- As future work could be interesting to see how a modern Multimodal LLM does on this task.

References:

- Yu, Jianfei, et al. "Grounded multimodal named entity recognition on social media." Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2023.