Fusion of Object-Centric and Linguistic Features for Domain-Adapted Multimodal Learning

RANLP 2025

Jordan Kralev

Motivation and Problem Statement

- Need for Fine-Grained Multimodal Understanding
- Limitations of Generic Visual Recognition (e.g. COCO-trained models)
- Challenge of Domain Adaptation
- Goal of the Research
  - To bridge the gap between object-centric visual features and linguistic inference for fine-grained multimodal tasks

Related Work and Background

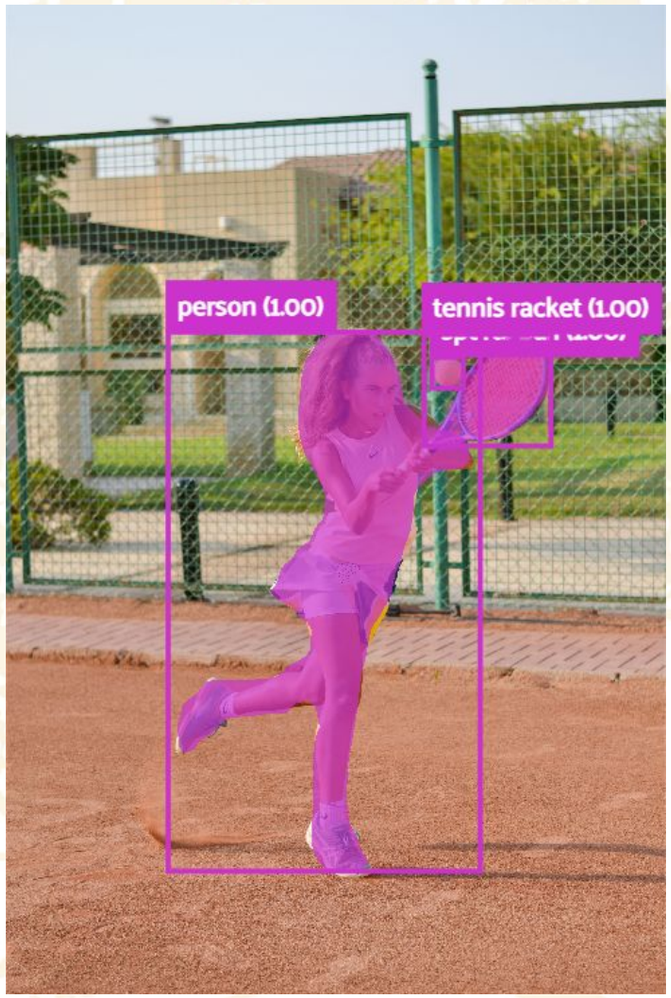
- Contrastive Language-Image Pre-training (CLIP) Framework
  - Aligns image and text in a shared embedding space
  - Employs separate image/text encoders with contrastive loss to keep matched pairs close and non-matched pairs apart.

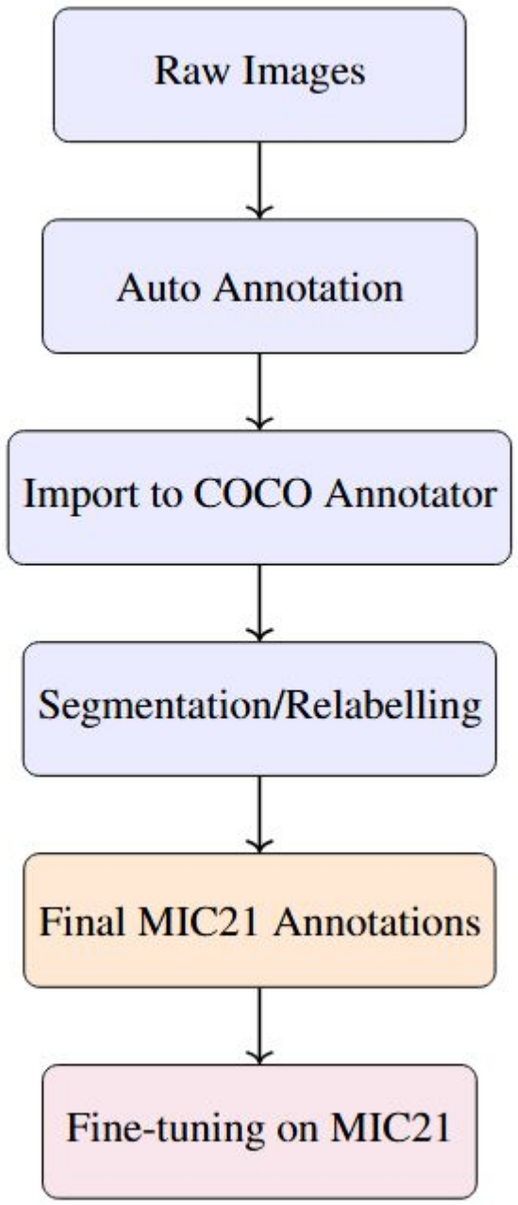
- Commonly Used Datasets and Models
  - Detectron2 and Yolact for object detection
  - COCO (objects in context), WebImageText, WebText

Dataset Creation and Ontology

- MIC21 Multilingual Image Corpus
  - 21,000 images and 200,000 annotations collected from diverse public sources
  - 700+ categories into 130 thematic subdomains
  - Linked to WordNet and supporting 25 languages
- Annotation Workflow
  - Automated image segmentation followed by manual refinement
  - Automated caption and description generation followed by manual refinement

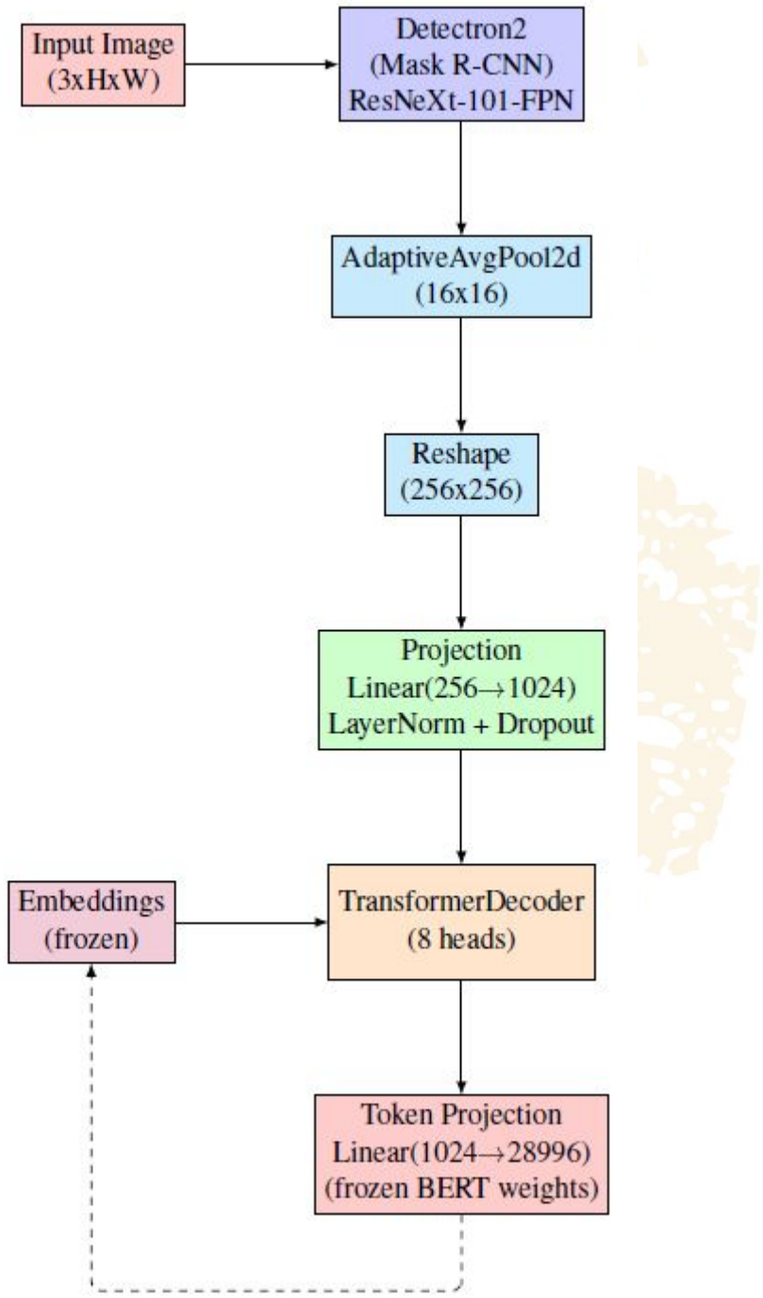
Dataset Creation and Ontology





Model Architecture Overview

- Visual Feature Extraction
- Trainable Projection Layer
- Transformer Decoder Stack
- Output Layer and Token Prediction



Training Strategy

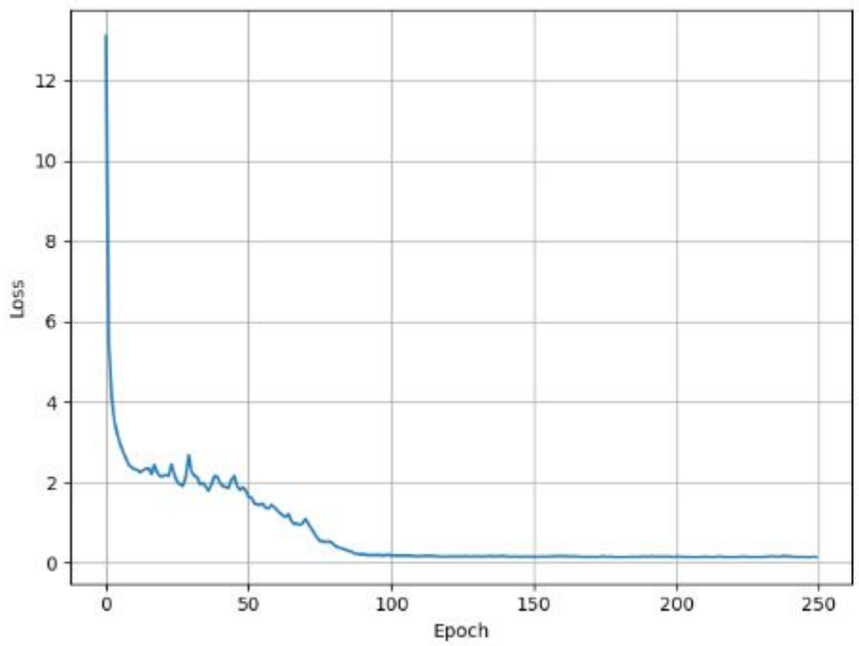
- Frozen Components
  - Visual processing backbone
  - Textual embedding layer.
- Trainable Components
  - Projection layer
  - Transformer decoder

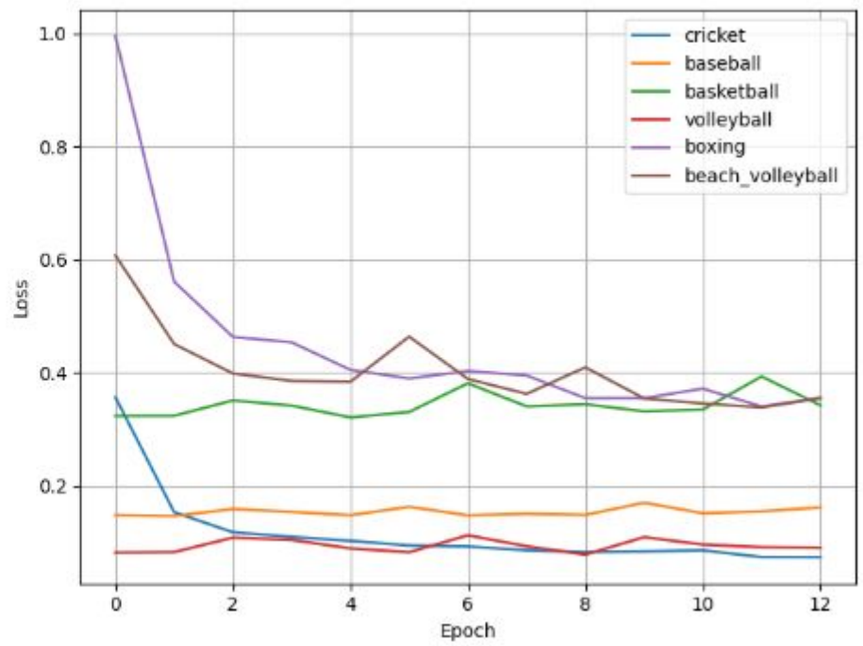
Epoch 3,4:
essessess Ch Ch Ch
Chusedessessess Ch Ch Focused
…
Epoch 10:
Ch Conessessess Contemplating
…
Epoch 90:
Intense Chess Concentration: A Player's Deep Thought

- Objective Function
  - Negative log-likelihood of the target token sequence conditioned on the image and preceding tokens.

$$\mathcal{L} = -\log p(w, I, \theta) = -\sum_{t=1}^{T-1} \log p(w_{t+1} | w_1, \ldots, w_t, I, \theta)$$

Experimental Results





Test	MIC21	Gemma 3
BLEU-1	0.71	0.79
BLEU-2	0.53	0.64
BLEU-3	0.47	0.53
BLEU-4	0.39	0.43

Conclusion

- Future Work
  - Extend the MIC21 ontology to include dynamic interactions between objects to capture richer scene understanding.
  - Integration with pre-trained language models
  - Compare against state-of-the-art models in generation and tagging tasks
  - Apply to practical, domain-critical applications

Acknowledgments

- The work is part of the project *Infrastructure for Fine-tuning Pre-trained Large Language Models*, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria

- http://ifgpt.dcl.bas.bg