Direct Projection of Visual Features into Linguistic Embeddings for Domain-Adapted Multimodal Learning

Jordan Kralev

CLIB 2026 — 7–10 September 2026, Sofia, Bulgaria

Department of Systems and Control, Technical University of Sofia, Bulgaria
Institute for Bulgarian Language, Bulgarian Academy of Sciences

Outline

- Motivation & research goal
- Method overview
- Visual model & projection layer
- Language model & prompt injection
- Training strategy
- MIC21 dataset
- Setup, results & analysis
- Conclusions & future work

Motivation

- Large models still struggle to fuse domain-specific data across modalities (visual, audio, linguistic).
- Multimodal applications: medicine, autonomous systems, education, manufacturing.
- Core tasks: image captioning and visual question answering.
- Existing frameworks (e.g., CLIP) require large-scale joint pre-training of contrastive encoders.

Question

Can we bridge a strong pre-trained *visual* model and a pre-trained *language* model with minimal fine-tuning of a connecting layer?

Background

- **CLIP** — contrastive visual–language pre-training; two parallel encoders aligned in a shared embedding space.
  - Trained on WebImageText, $\sim$400M loosely correlated image–caption pairs.
  - Contrastive objective also shapes discriminability of the visual encoder.
- **Self-supervised visual learning** — non-contrastive objectives (BYOL, Barlow Twins, SwAV) avoid collapse without negative samples.
- **Gap:** region-level detector features (Detectron2 / YOLO) are incompatible with sequential, tokenised LLM embeddings.
  - Direct concatenation / shallow fusion $\Rightarrow$ poor cross-modal interaction.

Research Goal

Goal

A general, adaptable framework that enables object recognition to transfer across domains with custom categories through a learnable projection layer, unifying visual and linguistic embeddings for diverse downstream tasks.

- Connect robust pre-trained models with minimal fine-tuning of the connecting layers.
- Demonstrate domain adaptation on the Multilingual Image Corpus (MIC21).
- Support image captioning and visual question answering.

Method Overview



$$z_0 = (e_{\text{pre}},\ v,\ e_{\text{post}}) \in \mathbb{R}^{(256+T)\times d}, \qquad v = W_p\, F_{\text{drop}}(F_{\text{norm}}(x_f)) + b_p$$

Image → region-level features → one-layer projection into LLM space → frozen transformer decoder.

Visual Model: Detectron2 Backbone

- Mask R-CNN, `X-101-32x8d-FPN_3x` configuration (COCO instance segmentation).
- Backbone extracts, for RGB input $I \in \mathbb{R}^{H \times W \times 3}$, the tensor

$$x_i \in \mathbb{R}^{C \times H_F \times W_F}, \qquad C = 256,\ H_F = W_F = 16, \tag{1}$$

- Flattened into the feature matrix

$$x_f \in \mathbb{R}^{C \times (H_F W_F)}. \tag{2}$$

- Weights frozen — the backbone acts as a fixed region-level feature extractor.

Projection Layer

The feature matrix is normalized and projected into the LLM embedding space:

$$v = W_p\, F_{\text{drop}}\big(F_{\text{norm}}(x_f)\big) + b_p, \qquad W_p \in \mathbb{R}^{d\times C},\ b_p \in \mathbb{R}^{d}, \tag{3}$$

with $d = 2048$ (TinyLlama embedding dimension), so $v \in \mathbb{R}^{256\times 2048}$ behaves as 256 vision-derived tokens.

- Conditioning transform:

$$x_{fn} = \gamma \frac{x_f - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \qquad (\gamma, \beta \text{ trainable}), \tag{4}$$

- Dropout F_{drop} applied during training only.
- W_p and b_p are the only trainable parameters in the whole system.

- Token sequence $t = (t_1, \ldots, t_T)$ mapped to embeddings $e = (e_1, \ldots, e_T) \in \mathbb{R}^{T \times d}$.
- Visual projections injected into the prompt:

$$z_0 = (e_{\text{pre}},\ v,\ e_{\text{post}}) \in \mathbb{R}^{(256+T)\times d} \tag{5}$$

- Decoder layer (multi-head self-attention, L layers, causal mask M):

$$a_l = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right) V, \qquad M = \begin{pmatrix} 0_{256\times 256} & 0_{256\times T} \\ 0_{T\times 256} & U_{T\times T} \end{pmatrix} \tag{6}$$

- Autoregressive next-token generation: $y = \arg\max_i(z_i)$.

Generation Prompt

Visual features are injected at the end of the system message ($\rightarrow$ 256 vision-derived tokens per image):

```
<|system|>
Generate title and description for the provided image. The image features are:
<v1><v2>...<v256></s>
<|user|>
Generate a title:</s>
<|assistant|>
```

Config: max generation sequence 40 tokens; English-only text model (design extensible to multilingual).

Training Strategy

- Visual and linguistic components frozen; learning focuses on the projection + normalisation layers.
- Objective: minimise negative log-likelihood of generated captions

$$\mathcal{L} = -\sum_{t=1}^{T-1} \log p(w_{t+1} \mid w_{1:t}, I, \theta), \qquad \theta = \{W_p, b_p, \gamma, \beta\} \tag{7}$$

- Autoregressive generation loss with softmax probabilities $p(w_{t+1} = i \mid \cdot) = \frac{\exp(z_{t,i})}{\sum_j \exp(z_{t,j})}$.
- Backpropagation only through the projection layer:

$$\frac{\partial \mathcal{L}}{\partial W_p} = \sum_{t=1}^{T} \left(\frac{\partial \mathcal{L}_t}{\partial z_t} W_e \right) \frac{\partial y_t}{\partial v} \frac{\partial v}{\partial W_p} \tag{8}$$

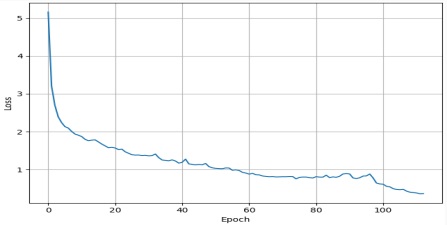
- Avoids vanishing/exploding gradients of unrolled multi-layer fine-tuning.

Dataset: Multilingual Image Corpus (MIC21)

- Expands the 80 COCO categories into an extensive ontology:
  - $>$ 700 classes in 130 thematic subdomains (sport, transport, arts, security, ...)
- Corpus scale: $\approx$ 21,000 images with $>$ 200,000 object detections.
- Ontology linked to WordNet, translated into 25 languages (multilingual captions/descriptions).
- Evaluation domain here: "cricket" (sport).
- Design extensible to multilingual generation.

Experimental Setup

- **Visual:** Detectron2 Mask R-CNN (X_101_32x8d_FPN_3x), frozen.
- **Language:** TinyLlama-1.1B-Chat-v1.0, frozen.
- **Hardware:** 2× GTX 1080 + 1× RTX 3070.
- 115 epochs on the cricket domain; batch size 7; max generation length 40.
- Adam optimizer, LR $= 10^{-4}$.



Average training loss; caption quality appears after ≈ epoch 100.

Example: Detections from Detectron2



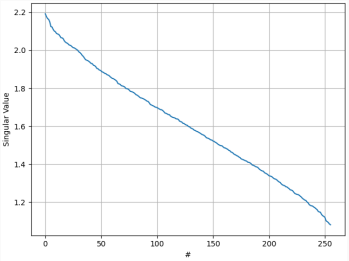
Diverse cricket-domain scenes; detections traced through consecutive Detectron2 layers (not projected — future work).

Results: Generation Metrics (cricket domain)

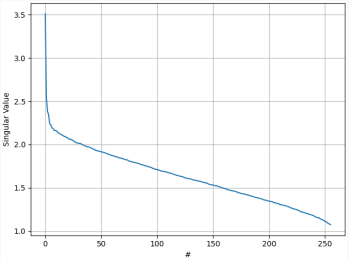
Metric	Epoch 115	Epoch 0
BLEU	**0.70**	0.0
ROUGE 1	**0.84**	0.005
ROUGE 2	**0.74**	0.0
ROUGE L	**0.83**	0.005
ROUGE L S	**0.83**	0.005
METEOR	**0.81**	0.004
Precision	**0.93**	0.46
Recall	**0.92**	0.52
F1 Score	**0.92**	0.49
WER	**0.23**	1.35

- Untrained model (epoch 0): near-zero BLEU / ROUGE / METEOR, WER $= 1.35$.
- After 115 epochs: large gains in all metrics.
- Precision / Recall / F1 > 0.9; WER drops to 0.23.
- Strong fidelity to reference captions and robust classification/retrieval.

What Changed Inside the Projection Layer? — SVD



Before training: well-conditioned, flat spectrum.



After training: energy concentrates in the leading mode (> 3.5).

Interpretation

Training polarises the projection directions — a few dominant modes carry most cross-modal information; the remainder compress (emergent sparsity), prioritising task-relevant structure and improving robustness/generalisation.

Conclusions

- A trainable projection layer directly maps high-dimensional region-level visual features into the TinyLlama embedding space with minimal fine-tuning.
- Frozen Detectron2 + TinyLlama backbone; only projection / normalisation parameters are learned — cheap, scalable domain adaptation.
- On MIC21 ($>$ 21k images, $>$ 700 categories): substantial gains in BLEU, ROUGE, METEOR, precision, recall and WER.
- SVD analysis confirms the layer learns task-relevant cross-modal directions.

Future Work

Domain-specific Detectron2 fine-tuning; multi-domain / multilingual generation; integrative captioning & VQA benchmarks beyond cricket.

Thank you!

Questions & discussion

Jordan Kralev
jkralev@tu-sofia.bg · jordan@dcl.bas.bg