Session ARS-18
Machine Learning for Recognition in Images and Videos II

Visual Relationship Classification with Negative-Sample Mining

Roberto de Moura Estevão Filho

robertomest@poli.ufrj.br

Federal University of Rio de Janeiro

José Gabriel Rodríguez Carneiro Gomes

Federal University of Rio de Janeiro

Leonardo de Oliveira Nunes

Microsoft Advanced Technology Labs Brazil

Contents

- Problem Definition
- Dataset
- Network Architecture
- Training Procedure
- Experiments
- Conclusions

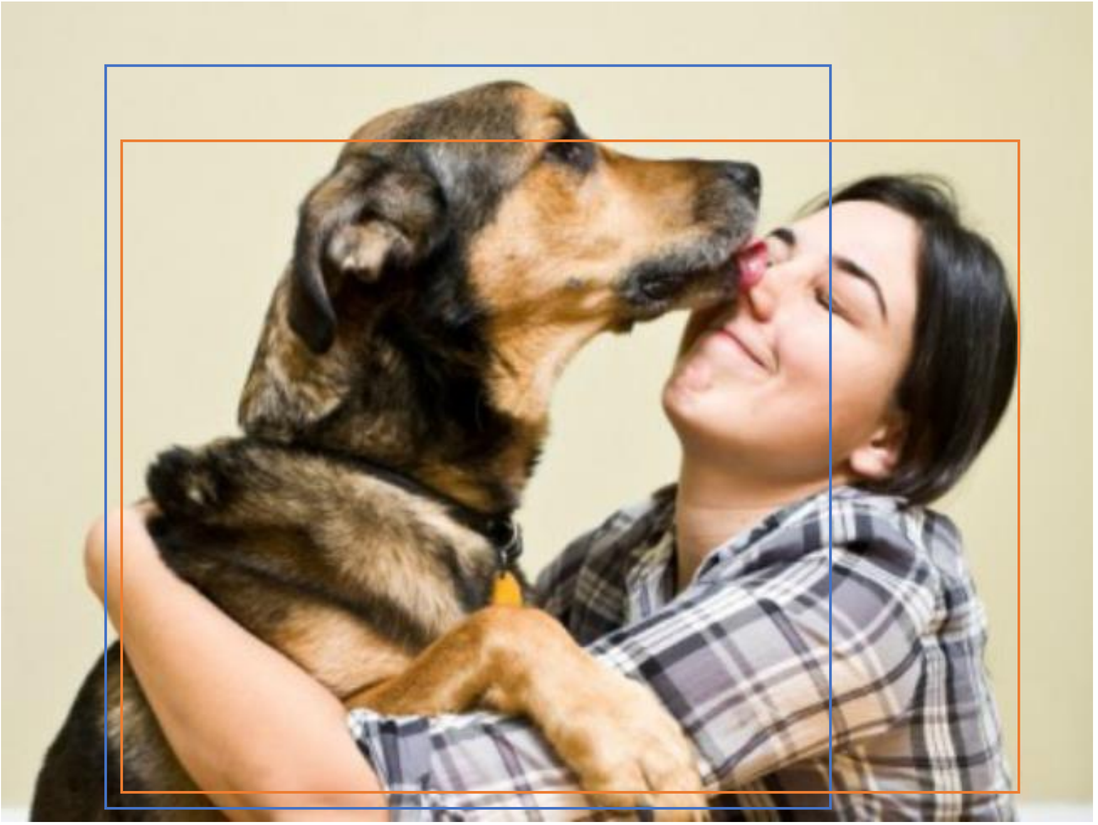
Contents

- **Problem Definition**
- Dataset
- Network Architecture
- Training Procedure
- Experiments
- Conclusions

Problem Definition

- Relationships are **triplets** `<subject − predicate − object>`
- Visual relationship classification is assigning predicates to object pairs
- An object pair is an **ordered pair** with a `subject` and an `object`
- The **predicate indicates** the **relationship** between the objects

Problem Definition



dog1 – lick – person1

person1 – hold – dog1

Problem Definition

- Most work focuses on using relationships as **content descriptors**
- Datasets are usually not exhaustively annotated
- Evaluation based on **recall**
- We focus on **event detection**
- Important to avoid false positives. Needs high **precision**
- We adopt **mAP as our evaluation metric**

Contents

- Problem Definition
- **Dataset**
- Network Architecture
- Training Procedure
- Experiments
- Conclusions

Dataset

- **Open Images** dataset
- **9 predicate classes**
- Relationships are **exhaustively annotated**
- Almost **60k images** and over **180k relationships**
- **97% of possible pairings have no relationship**
- **Class imbalance**
  - **Most** common predicate has over **100k samples**
  - **Least** common has 34 **samples**

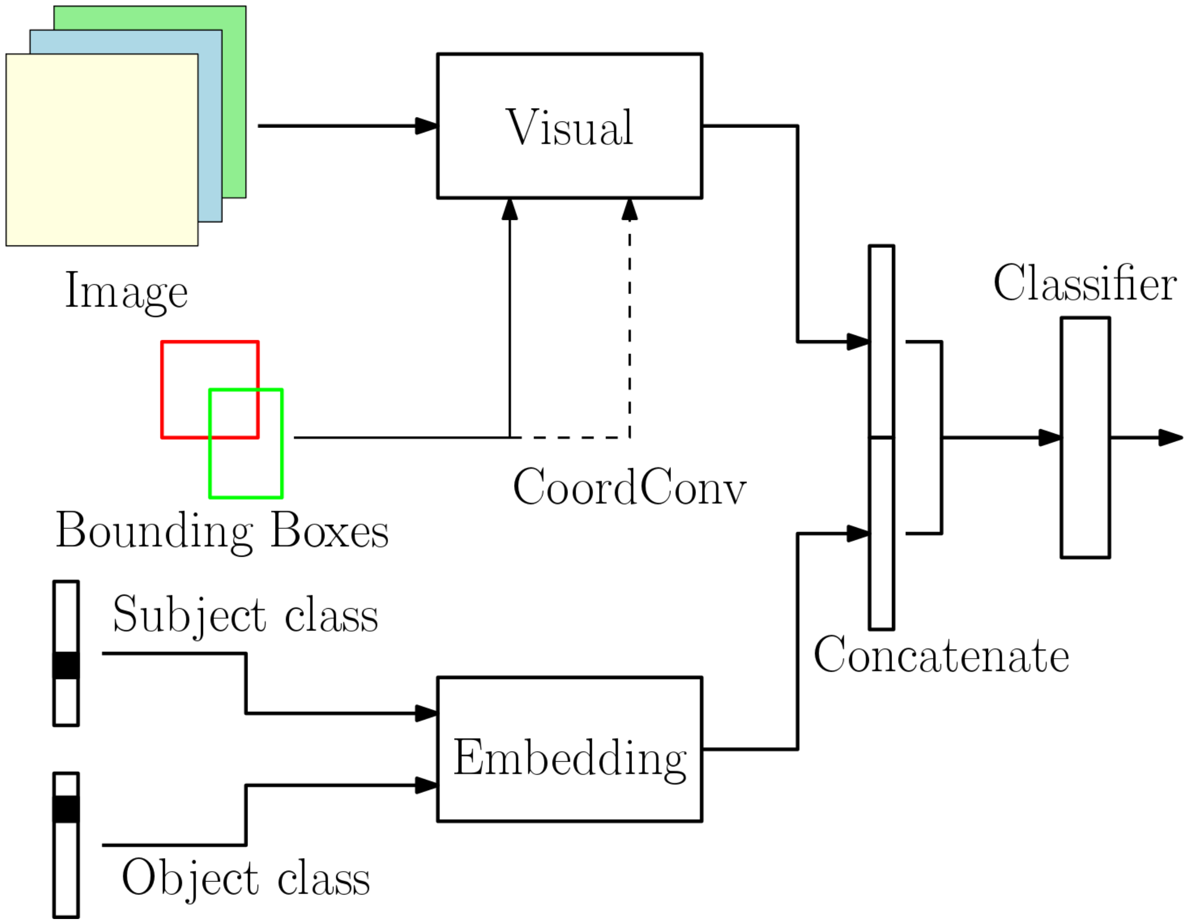
Contents

- Problem Definition
- Dataset
- **Network Architecture**
- Training Procedure
- Experiments
- Conclusions

Network Architecture

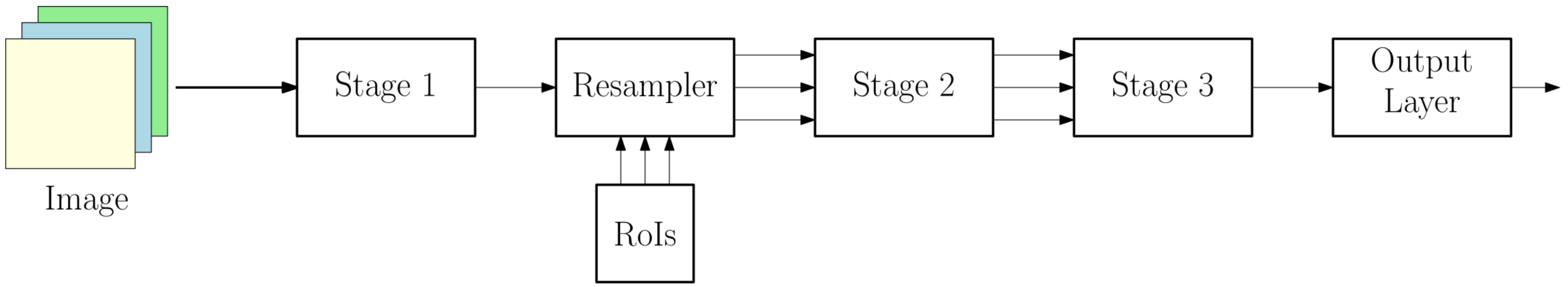
- Neural network that incorporates **three types of information**
- **Visual**: **CNN**
- **Spatial**: **CoordConv**-like explicit positional encoding feature maps
- **Object class**: learned class **embeddings**

Network Architecture



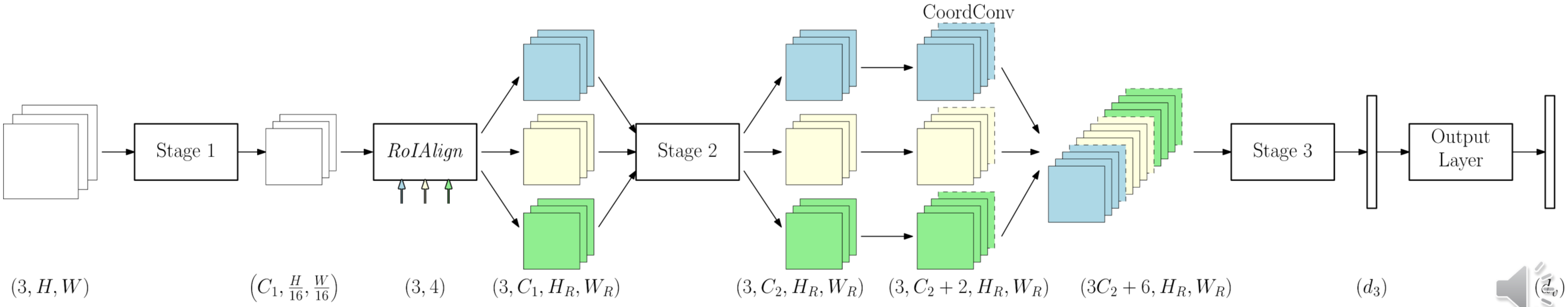
Network Architecture – Visual Module

- Divided in three main stages
  - **Stage 1**: extracts features from the **whole image**
  - **Resampler:** obtains features for **subject**, **object**, and **union RoIs**
  - **Stage 2**: **RoIs are processed independently**
  - **Stage 3**: extracts features by **combining information from all** three **RoI feature maps**
  - **Output layer: FC layer** that maps features into an output visual feature vector



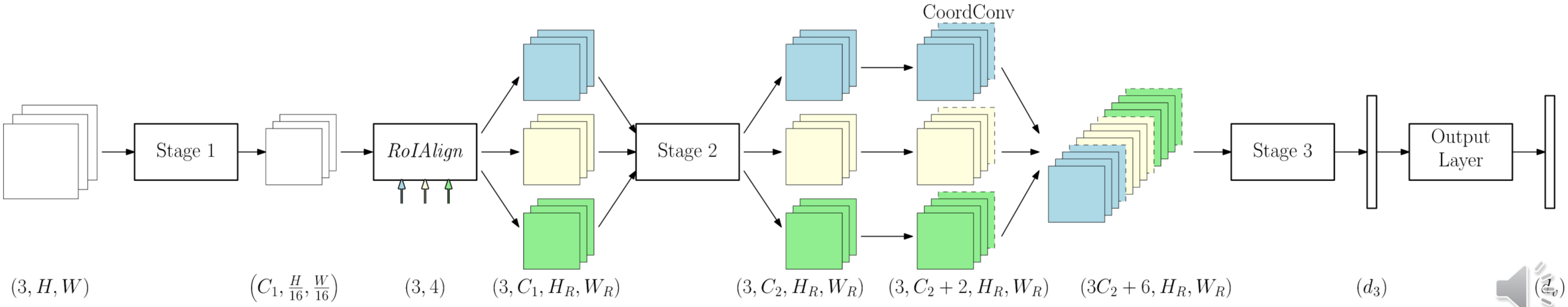
Network Architecture – Visual Module

- Stages 1 and 2: layers from **ResNet-18 pretrained on ImageNet**
- Features are resampled using **RoIAlign**
- Stage 3: **two convolutional layers** with 512 filters followed by **GAP**



Network Architecture – CoordConv

- Explicit **positional encoding** feature maps added to the **output of stage 2**
- Encode interpolated pixel **positions relative to** the **union RoI**
- Relative **position** and **scale** of subject and object RoIs
- **2 feature maps per RoI**



Network Architecture – Embedding

- **Linear** embedding functions
- Two different embeddings matrices
- Embedding **output vector** is the **concatenation** of both embedding vectors
- Used as **extra features in** the **classifier**

Contents

Batch Sampling and Balancing

- Batch is constructed by sampling 40 images from one of two bins (**wide** vs **tall** images)
- All images are resized:
  - Smaller size to 480 pixels
  - Larger size not larger than 960 pixels
- Images are packed in a tensor with **zero padding**

Batch Sampling and Balancing

- All **annotated relationships** from the images are **positive samples**
- All **unannotated object pairs** are used as **negative samples**
- **Batches are balanced** by sampling with following criteria
  - Maximum of 400 relationships per batch
  - **At least 25%** of **positive** relationships
  - **At most 75%** of **positive** relationships

Contents

Metrics

- **Metrics based on** mean average precision (**mAP**)
- **Two variants**:
  - * indicates ignoring predicate class ‘under’
  - The underscript $_{FG}$ indicates only considering positive samples
- Total of **four metrics**: mAP, mAP*, mAP_{FG}, mAP*_{FG}

Experiments

- Our **negative mining** sampling method **vs** using **only annotated relationships**.
- Averages of three runs (standard deviation in parenthesis)

Method	mAP	mAP*	mAP_{FG}	mAP^*_{FG}
GT only	34.6 (1.2)	38.8 (1.3)	91.0 (2.1)	96.7 (0.3)
Ours	78.2 (0.7)	83.3 (0.5)	88.7 (1.5)	94.4 (0.4)

Ablation Experiments

- Comparisons based on **mAP***
- **Baseline** uses only **visual information**
- **Spatial information improves** performance **slightly**
- **Class information** is **more important**

Model	mAP	mAP*		mAP_{FG}	mAP^*_{FG}
Baseline	78.2 (0.3)	**80.8 (0.3)**		88.8 (1.3)	92.1 (0.1)
+ CoordConv	78.1 (2.2)	**81.6 (0.8)**	**+0.8** **+2.3**	90.1 (0.8)	92.7 (0.9)
+ Embedding	79.0 (1.4)	**83.1 (0.8)**		89.5 (2.1)	94.3 (0.3)
All	78.2 (0.7)	**83.3 (0.5)**	**+0.2**	88.7 (1.5)	94.4 (0.4)

Contents

- Problem Definition
- Dataset
- Network Architecture
- Training Procedure
- Experiments
- **Conclusions**

Conclusions

- Use of a **visual relationship classifier for event detection**
- **Focus on precision**
- Training scheme **improves rejection of unrelated pairs**
- **Small penalty to classification** between predicates
- **CoordConv** provides **small performance improvement** at **minor computational cost**
- **Extension to other datasets**
  - **How to measure precision in non-exhaustively annotated datasets?**

This work was supported by
Microsoft ATL Brazil
CNPq
CAPES