THE UNIVERSITY of ADELAIDE

类别标签的语义信息在视觉识别模型中的应用

刘凌峤
阿德莱德大学
（University of Adelaide）

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

基于 …

Peng Wang*, **Lingqiao Liu*,** Chunhua Shen, Zi Huang, Anton van den Hengel, Heng Tao Shen, Multi-attention Network for One Shot Learning**.** Computer Vision and Pattern Recognition (CVPR) 2017. (* Indicates equal contribution.)

Bohan Zhuang*, **Lingqiao Liu*,** Chunhua Shen, Ian Reid, Towards Context-aware Interaction Recognition**.** International Conference on Computer Vision (ICCV) 2017 (* Indicates equal contribution.)

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

重新审视类别标签

汽车

摩托车

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

重新审视类别标签

汽车 = class 1

摩托车 = class 2

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

重新审视类别标签

汽车 = class 1

语义信息?

摩托车 = class 2

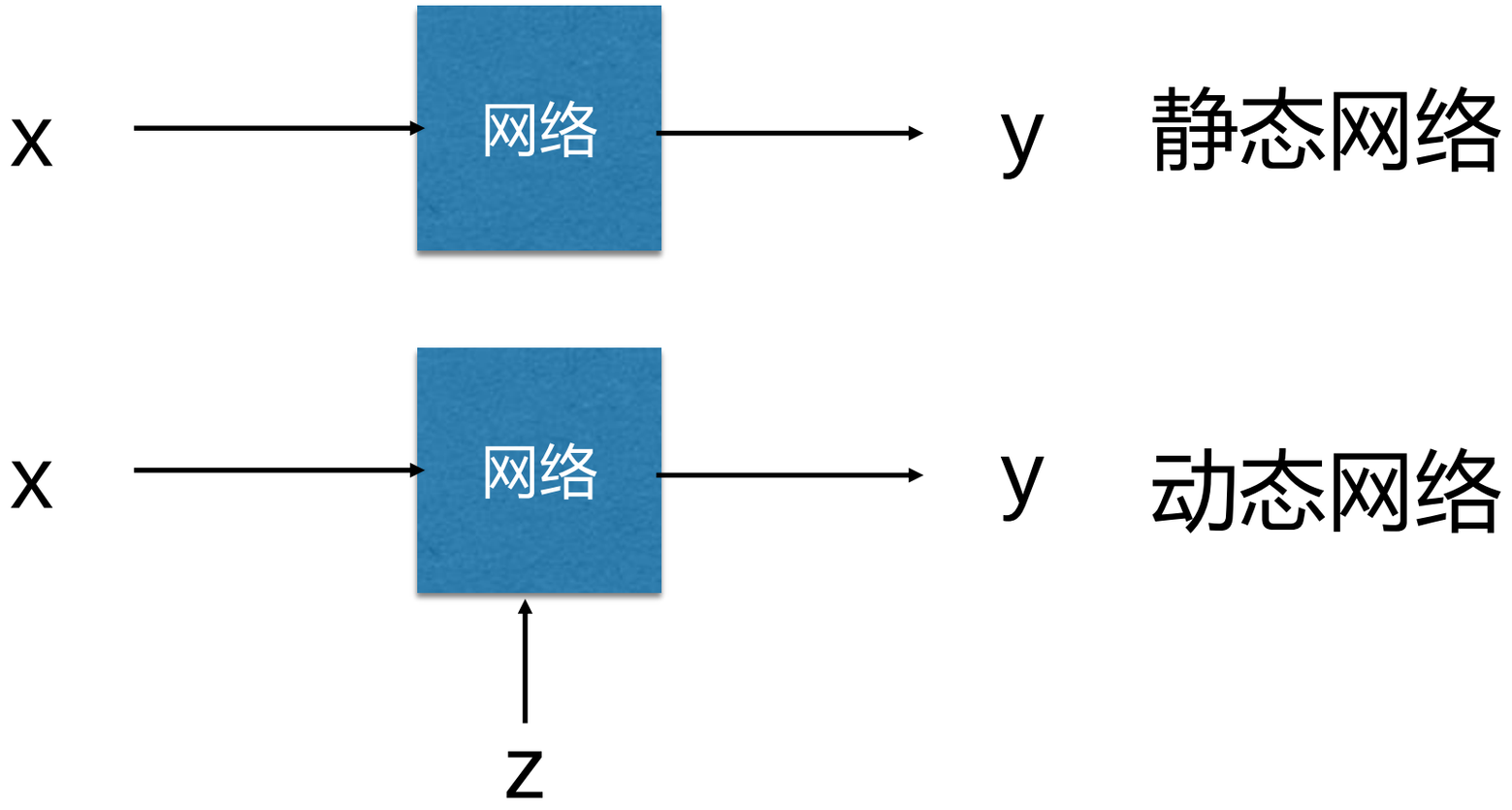
Australian Centre for Visual Technologies
Innovation and education in visual information processing.

标签的语义信息

- 1. 标签的语义信息可以给我们提供什么?
- 作为连接不同类别的桥梁
- 对小样本学习或者大量类别识别提供帮助

思路概览

- 把标签的语义信息与深度神经网络结合



Australian Centre for Visual Technologies
Innovation and education in visual information processing.

报告大纲

- 工作1: 小样本下的图像识别
  - 用标签信息控制注意力模块实现图像信息的抽取
- 工作2: 视觉关系的检测
  - 用待识别标签的上下文信息 （详见相关页）
  - 用标签信息控制分类模块和注意力模块

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

概览

- 工作1: 小样本下的图像识别

- 工作2: 视觉关系的检测

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

研究动机

- 单（小）样本识别：一个类提供一个示例
- 常见传统思路：学习好的特征或者匹配函数
- siamese 网络
- triplet 网络
- 利用分类网络学习特征

THE UNIVERSITY of ADELAIDE

传统思路

CNN

CNN

Cls

-1/1

CNN

Cls

car
motorbike
…

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

语义不确定性

Tutu

找到同类的图像?

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

语义不确定性

Tutu: 一种服饰

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

语义不确定性

Tutu: 一种服饰

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

我们学到了什么？

- 标签的语义信息可以帮助减少语义不确定性
- 即便是弱的语义信息也会有帮助
- 具体过程：通过语义信息指导我们注意一部分图像内容（注意力机制）

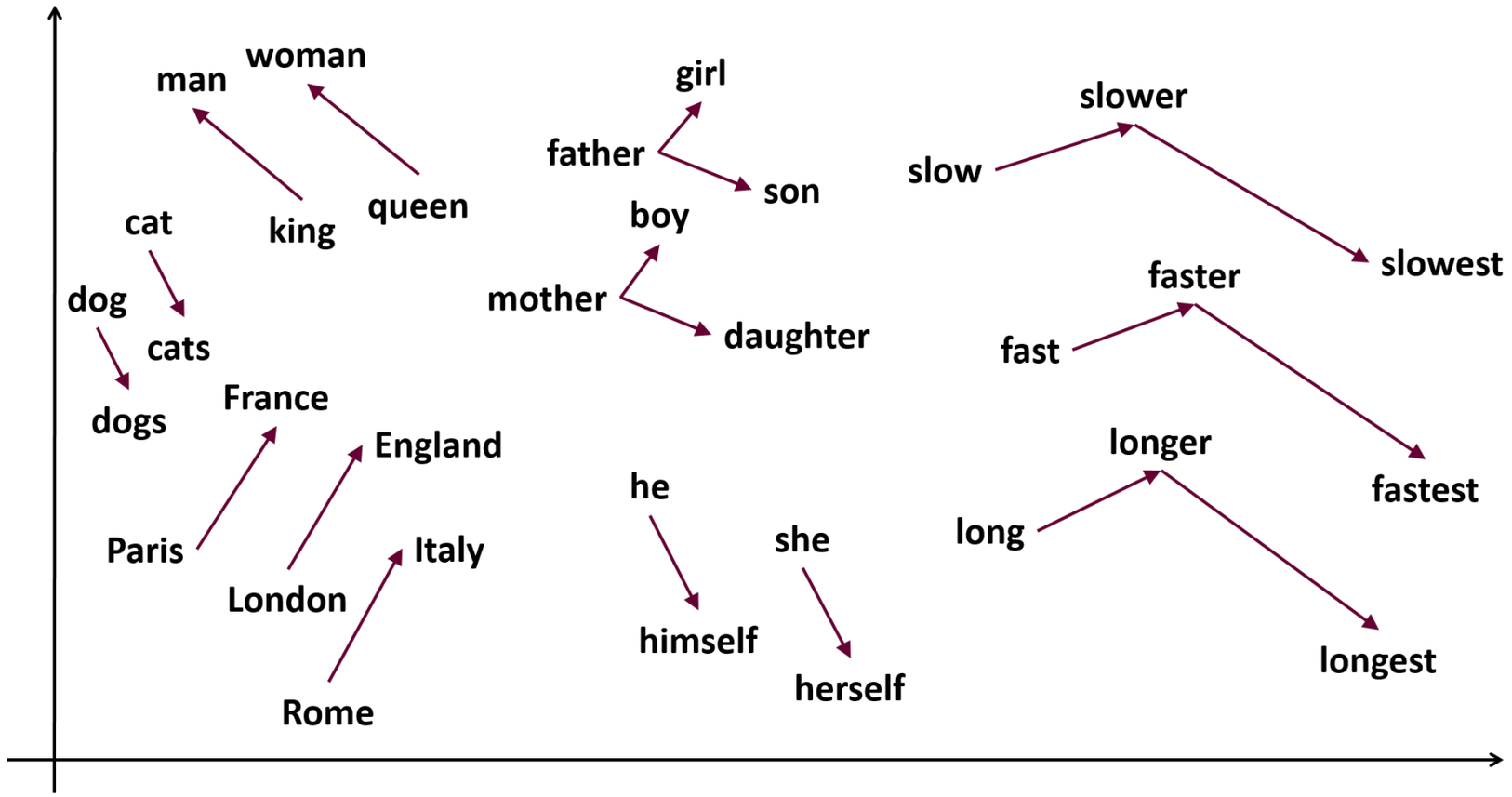
Australian Centre for Visual Technologies
Innovation and education in visual information processing.

怎样实现?

- 1. 怎样表示语义信息?
- 2. 怎样实现语义指导的注意模块，并将其用于图像特征提取

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

Distributed Word Embedding

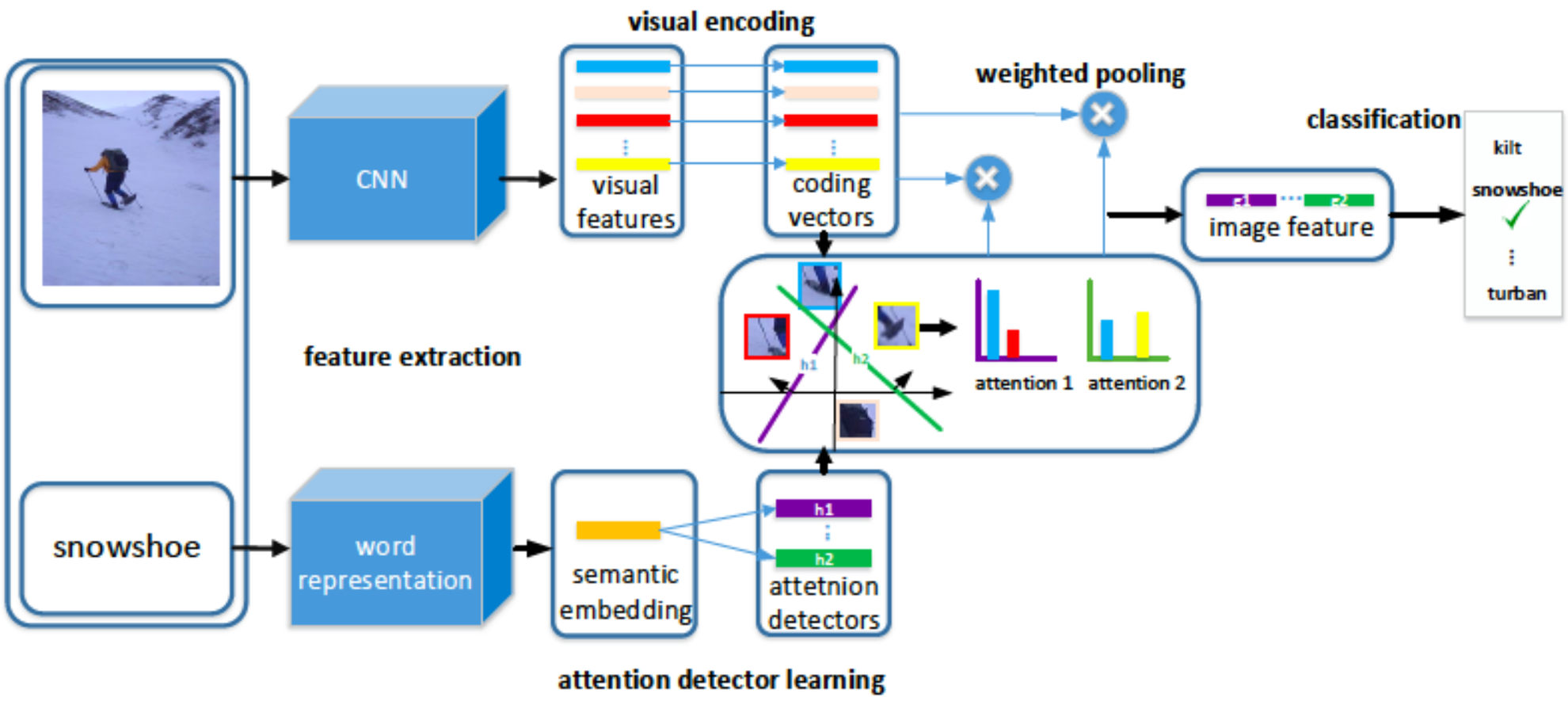


Distributed Word Embedding

- 语义接近的词，对应的词向量也接近
- 在自然语言处理领域被广泛应用，学习词向量一般不需要监督信号
- 噪声比较大，只包含较弱的语义信息。

THE UNIVERSITY of ADELAIDE

网络结构



THE UNIVERSITY of ADELAIDE

网络结构

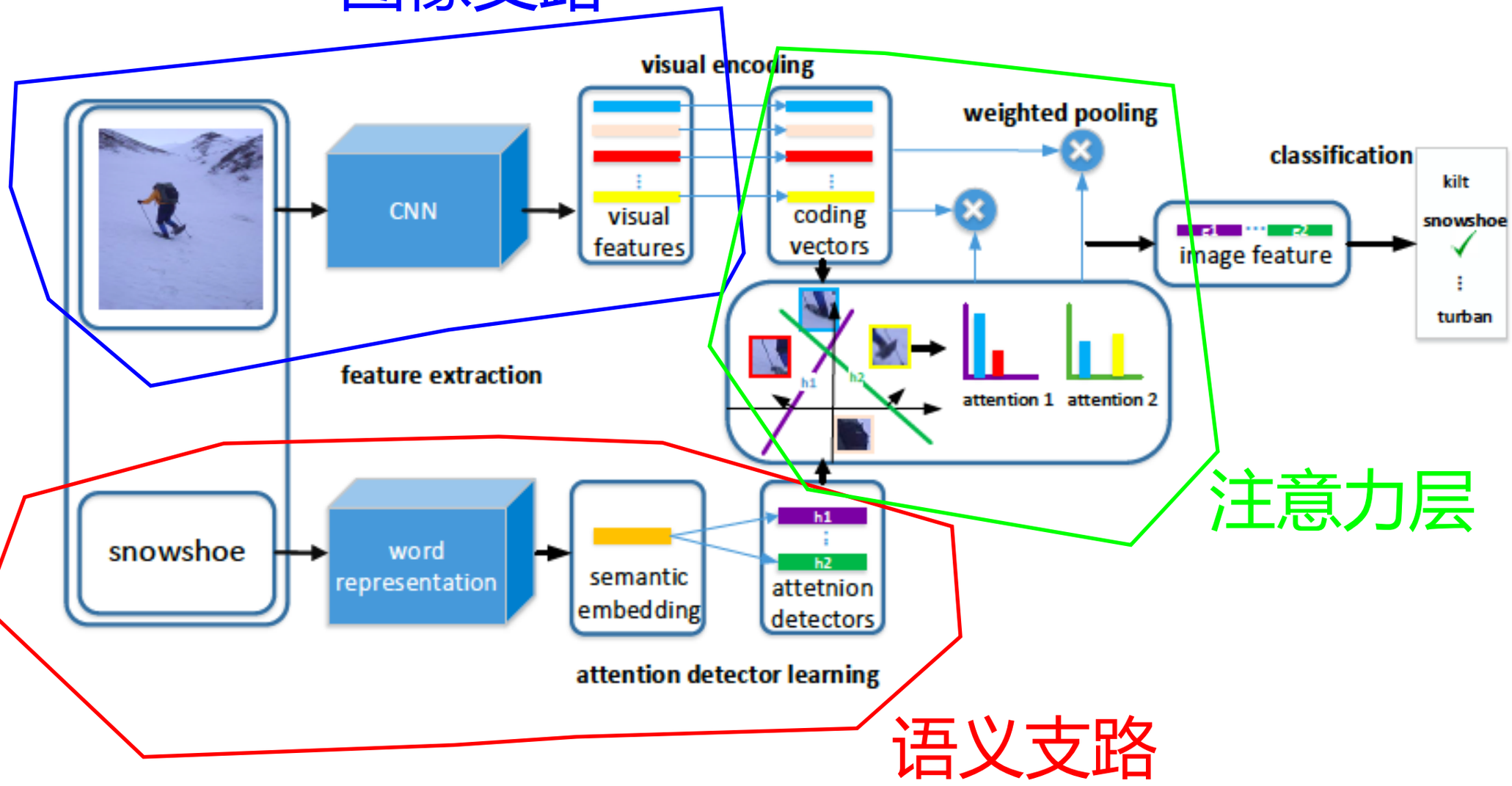
- 两条支路
  - 图像支路：提取图像特征
  - 语义支路：提取与图像相关的语义信息
- 注意力层
  - 作用于图像支路，受语义支路指导

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

网络结构

图像支路



具体实现

图像支路:

$$\mathbf{v}_i = f(\mathbf{W}_v \mathbf{x}_i + \mathbf{b}_v), \tag{1}$$

// 输入一个特征集合

where $f(a) = \max(0, a)$ is a rectified linear unit (ReLU), $\mathbf{x}_i$ is a local feature, $\mathbf{W}_v \in \mathbb{R}^{d \times d_v}$ and $\mathbf{b}_v \in \mathbb{R}^d$ are the model parameters.

$$\mathbf{g} = \sum_{i=1}^{|\mathcal{X}|} \mathbf{v}_i a_i, \tag{2}$$

// 加权平均池化

语义支路:

$$\mathbf{h} = \mathbf{W}_s \mathbf{c} + \mathbf{b}_s, \tag{3}$$

// 线性投影

where $\mathbf{c}$ is the semantic embedding of the class tag, $\mathbf{W}_s \in \mathbb{R}^{d \times d_w}$ and $\mathbf{b}_s \in \mathbb{R}^d$ are the model parameters to be learned.

具体实现

注意力层

$$a'_i = \mathbf{h}^\top \mathbf{v}_i. \tag{4}$$

// 衡量语义与图像局部特征的匹配程度

$$a_i = \frac{b(a'_i)}{\sum_i b(a'_i)}, \tag{5}$$

// 归一化

图像特征

$$\mathbf{g} = \sum_{i=1}^{|\mathcal{X}|} \mathbf{v}_i a_i, \tag{2}$$

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

训练过程

- 在一个辅助数据库上训练：该数据库上所使用的类与测试时的类没有重叠
- 损失函数: 普通的分类损失函数，比如交叉熵
- 区别
  - 图像 - 标签 vs（图像 对应的标签的词向量）- 标签

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

测试过程

- 测试过程: 每一类提供类别名字和一个示例

- 去除分类层，只使用分类层前的特征层对图像提取图像特征

- 使用该特征搜索相似（同一类）的图像，或使用最近邻法分类图像

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

注意力层的效果

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

多重注意力

- 使用多重注意力
  - 提取多个图像区域
  - 避免漏掉重要区域
- 串联多次加权池化的结果

$$\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, \cdots, \mathbf{g}_t].$$

实验结果

- 训练: 40 类
- 测试: 10 类
- Close-world
  - 只测试测试类
- Open-world
  - 测试训练类和测试类

Table 1. Comparison of the attention network to alternative solutions on the Animal Dataset.

close-world	Global FC	68.9%
	SE	67.7%
	Zero shot learning	55.2%
	Attention (attribute)	**72.4%**
	Attention (word vector)	**74.0%**
	Multi-Attention (2)	**82.4%**
	Multi-Attention (5)	**77.7%**
open-world	Global FC	42.8%
	SE	42.1%
	Zero shot learning	15.3%
	Attention (attribute)	**47.4%**
	Attention (word vector)	**49.0%**
	Multi-Attention (2)	**55.7%**
	Multi-Attention (5)	**56.8%**

Experiments

- 训练： 1000 类
- 测试： 100 类

Table 2. Comparison of the attention network to alternative solutions on the Artifact Dataset.

close-world	SE (256D)	27.8%
	SE (512D)	28.2%
	SE + Joint Bayesian	31.5%
	Attention	**34.5%**
	Attention + Joint Bayesian	**36.8%**
	Multi-Attention (2)	**50.5%**
open-world	SE (256D)	11.6%
	SE (512D)	12.2%
	SE + Joint Bayesian	11.4%
	Attention	**15.2%**
	Multi-Attention (2)	**28.2%**

THE UNIVERSITY of ADELAIDE

概览

- 工作1: 小样本下的图像识别
- 工作2: 视觉关系的检测

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

研究动机

- 背景: 视觉关系的识别
- 例如 “人骑车”，“人在车后”
- 主体 - 关系- 客体
- (主体, 客体) = 关系的上下文

研究动机

- 关系识别: 最困难的模块
- 视觉关系并一定与唯一的图像模式相关联
- 比如. “play” boy play football, girl play piano

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

视觉关系识别

- 两种途径识别视觉关系 （定义视觉关系的类别）
  - 方式 1: 一种关系为一类，不管上下文信息
  - 方式 2: 一种关系和主体客体的组合为一类

THE UNIVERSITY of ADELAIDE

视觉关系识别

- 两种方法各有局限性
- 方法1: 较大的类内差异
- 方法2: 类的个数太多，不能泛化到新的组合

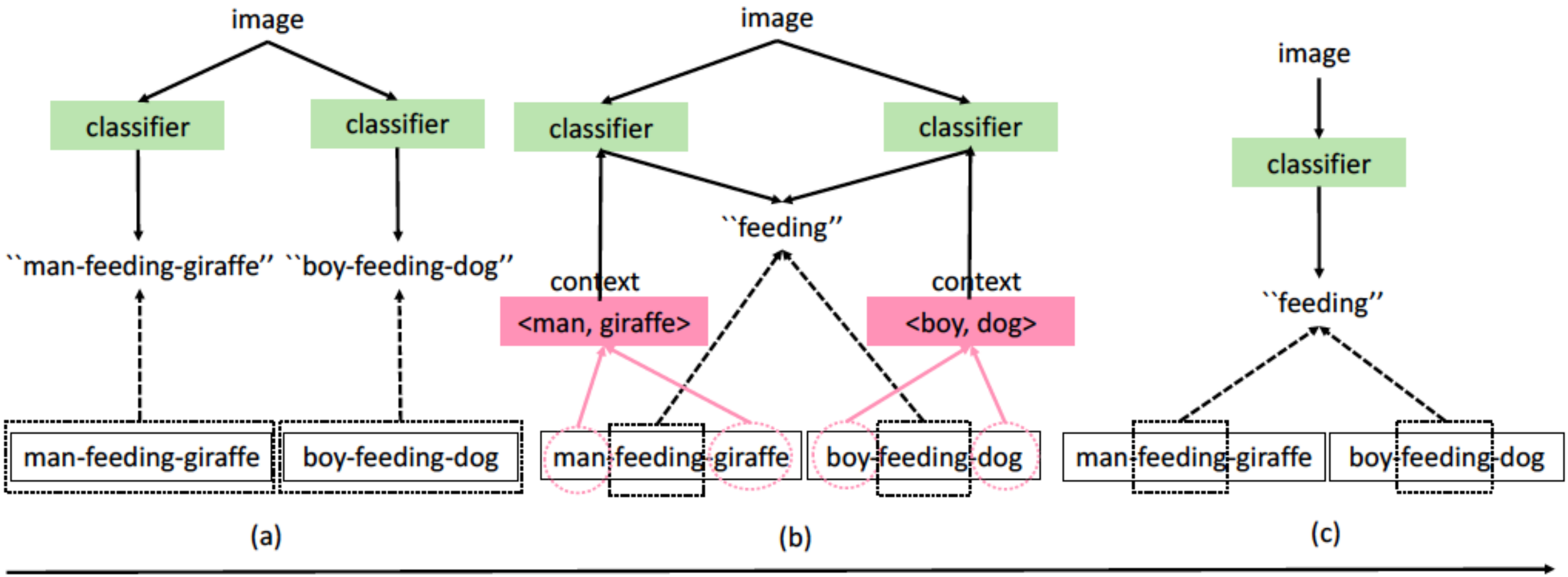
Australian Centre for Visual Technologies
Innovation and education in visual information processing.

我们的解决方案

- 我们的解决方案:
  - 定义每一种关系为一类，并分配一个分类器
  - 但是让该分类器受控于关系类的上下文信息： 动态分类器
- 主要区别:
  - 传统分类器: 静态的向量 w
  - 动态分类器：一个函数 w(z)

THE UNIVERSITY of ADELAIDE

Our solution



Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

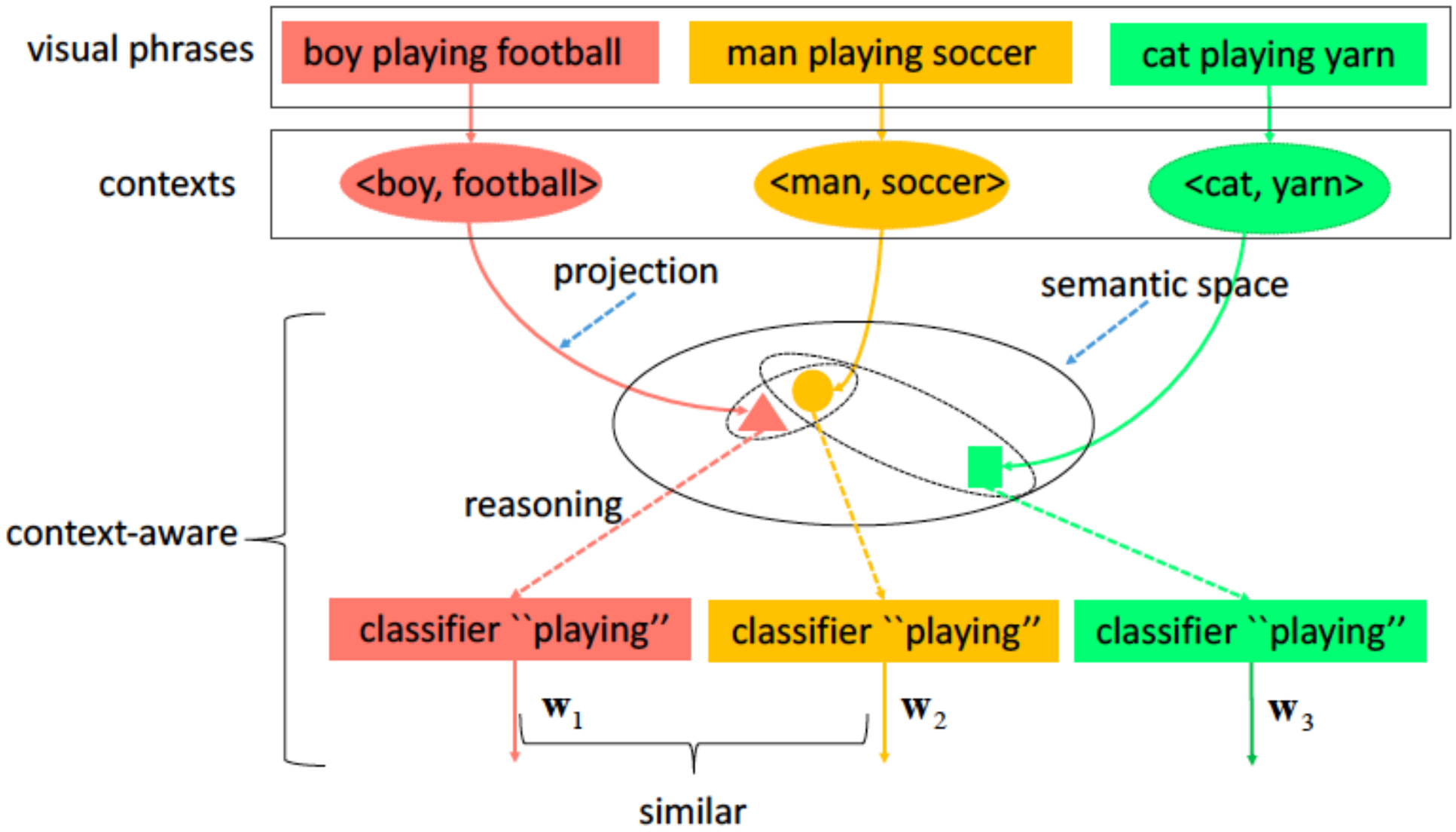
动态分类器

- 动态分类器 w(z) 的特性
- 相似的上下文, 相似的w(z) 输出
- 假设：上下文相似，视觉关系对应的图像外观也相似

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

动态分类器



Australian Centre for Visual Technologies
Innovation and education in visual information processing.

动态分类器

动态分类器

静态部分：不依赖于上下文

$$\mathbf{w}_p(O1, O2) = \bar{\mathbf{w}}_p + r_p(O1, O2),$$

上下文信息

动态部分：依赖于上下文

动态分类器

动态分类器

所有类共享

$$r_p(O1, O2) = \mathbf{V}_p f(\mathbf{Q} E(O1, O2)),$$

两层网络

每类不同

特征提取

- 适合识别视觉关系的图像特征
  - 输入:两个检测框（bounding box) 以及检测到的主体客体类别
  - 基于图片内容: 两个框并集所框选的图像内容，使用CNN来提取特征 （可以端到端训练）
  - 基于框的几何关系: 人为定义的几何关系特征

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

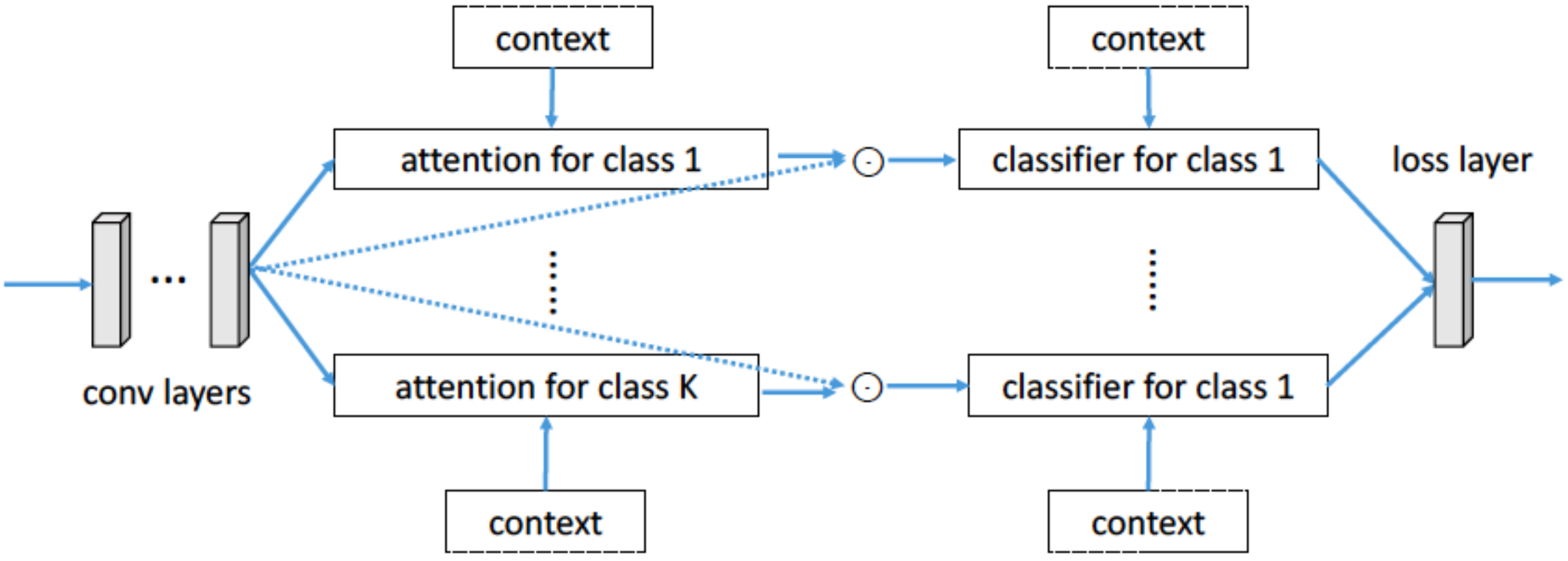
动态注意力模块

- 在关系识别中引入注意力机制
- 图像的少量区域提供了识别视觉关系的信息
- 传统注意力模块: 由局部特征决定
- 动态注意力模块: 由局部特征和上下文决定（与工作1类似）

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

网络结构



Australian Centre for Visual Technologies
Innovation and education in visual information processing.

实验结果

- 任务
  - 视觉交互关系 (predicate) 识别
    - 基于一个已有方法（文献【17】）
    - 只替代视觉关系（predicate）识别模块

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

实验结果

Method	Predicate Det.		Phrase Det.		Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Visual Phrase [30]	1.91	0.97	0.07	0.04	-	-
Language Priors [17]	47.87	47.87	17.03	16.17	14.70	13.86
Baseline1-app	18.13	18.13	6.02	5.42	5.54	5.01
Baseline1-spatial	17.77	17.77	5.24	4.77	4.54	4.19
Baseline2-app	27.23	27.23	9.30	7.91	8.34	7.03
Baseline2-spatial	13.85	13.85	4.15	3.06	3.63	2.63
Spatial+C	51.17	51.17	17.61	15.46	15.43	13.51
AP+C	52.36	52.36	18.69	16.91	16.46	14.88
AP+C+AT	53.12	53.12	19.08	17.30	16.89	15.40
AP+C+CAT	**53.59**	**53.59**	**19.24**	**17.60**	**17.39**	**15.63**

Table 1 – Evaluation of different methods on the visual relationship benchmark dataset. The results reported include visual phase detection (Phrase Det.), visual relationship detection (Relationship Det.) and predicate detection (Predicate Det.) measured by Top-100 recall (R@100) and Top-50 recall (R@50).

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

实验结果

Method	**Predicate Det.**		Phrase Det.		Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50
Language Priors [17]	8.45	8.45	3.75	3.36	3.52	3.13
Baseline1-app	7.44	7.44	3.08	2.82	2.91	2.74
Baseline1-spatial	7.27	7.27	2.14	2.14	2.14	2.14
Baseline2-app	7.36	7.36	2.22	1.71	2.05	1.54
Baseline2-spatial	0.43	0.43	0.09	0.09	0.09	0.09
Spatial+C	**16.42**	**16.42**	6.24	5.82	5.65	5.30
AP+C	15.06	15.06	5.82	5.05	5.22	4.62
AP+C+AT	15.00	15.00	5.62	5.02	5.36	4.76
AP+C+CAT	16.37	16.37	**6.59**	**5.99**	**5.99**	**5.47**

Table 3 – Results for zero-shot visual relationship detection on the visual relationship benchmark dataset.

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

实验结果

Comparing with the state-of-the-art methods

* using better detectors as in the recent works

Method	Phrase Det.		Relationship Det.		Zero-Shot Phrase Det.		Zero-Shot Relationship Det.	
	R@100	R@50	R@100	R@50	R@100	R@50	R@100	R@50
CLC (CCA+Size+Position) [24]	20.70	16.89	18.37	15.08	**15.23**	**10.86**	**13.43**	**9.67**
VTransE [39]	22.42	19.42	15.20	14.07	3.51	2.65	2.14	1.71
Vip-CNN [14]	**27.91**	22.78	20.01	17.32	-	-	-	-
VRL [15]	22.60	21.37	20.79	18.19	10.31	9.17	8.52	7.94
Faster-RCNN + (AP+C+CAT)	25.26	23.88	23.39	20.14	11.28	10.73	10.17	9.57
Faster-RCNN + (AP+C+CAT) + Language Priors	25.56	**24.04**	**23.52**	**20.35**	11.30	10.78	10.26	9.54

Australian Centre for Visual Technologies

Innovation and education in visual information processing.

这两份工作的启发

- 重新看待类别标签

- 1. 类别标签里的语义信息可以帮助我们更好的理解图像内容

- 2. 类别标签中的语义信息也可以帮助我们联系不同类别，充分使用这种联系可以帮助我们更好的设计分类器。

Australian Centre for Visual Technologies
Innovation and education in visual information processing.

THE UNIVERSITY of ADELAIDE

Thanks & Questions?

Australian Centre for Visual Technologies
Innovation and education in visual information processing.