Vision-language Models

Internship Presentation

10 May 2021 – 9 Jul 2021

Akkapaka Saikiran

CSE, IIT Bombay

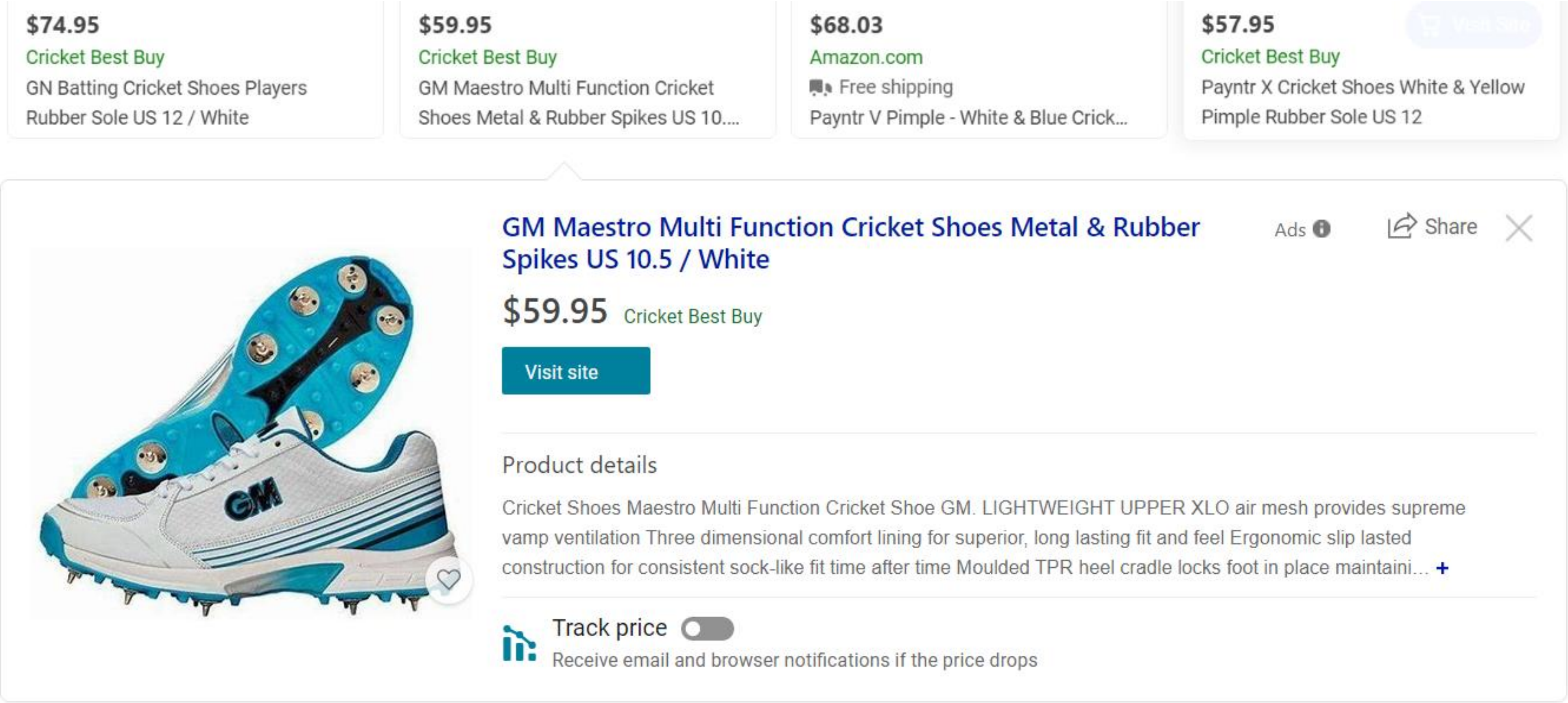
Mentors: Swati Tiwari, Neelesh Khanna, Niranjan Paramkusum

Outline

- Problem statement
- Current pipeline
- Baseline scores
- Multimodal learning
  - Pre-Oscar
  - Oscar
  - VinVL
- Results
  - Object detection examples
  - VinVL scores
- Future work

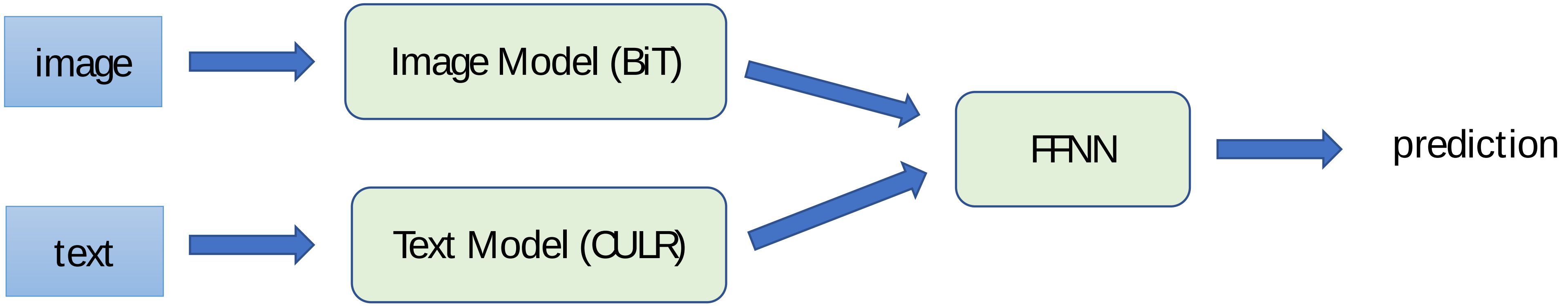
The Problem Statement

➢ Product Ads (PA) classification – adult, weapon



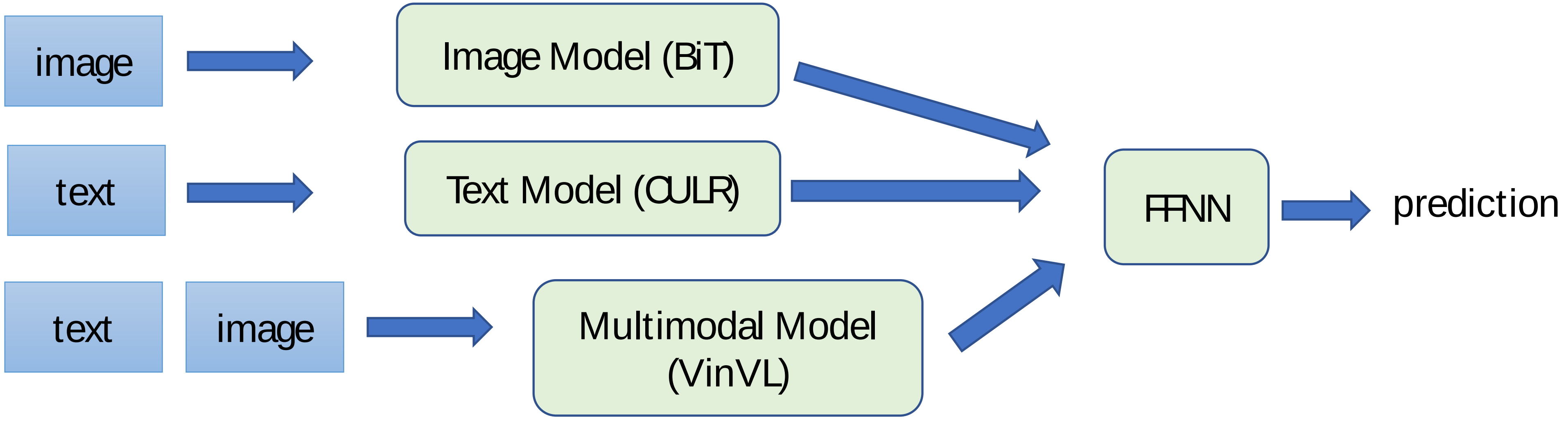
The Problem Statement

- Product Ads (PA) classification – adult, weapon
- PA - Image (raw and thumbnail)
  - Text (Product name, Merchant name, Description, etc.)
  - Labels (TextDisallowed, ImageDisallowed, OverallAdDisallowed)
- Unimodal pipeline



The Problem Statement

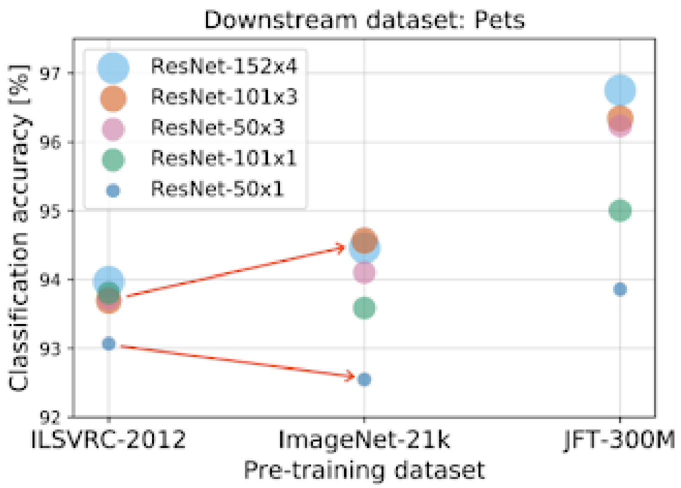
- ➢Product Ads (PA) classification – adult, weapon
- ➢PA - Image (raw and thumbnail)
  - Text (Product name, Merchant name, Description, etc.)
  - Labels (TextDisallowed, ImageDisallowed, OverallAdDisallowed)
- ➢Multimodal Pipeline:

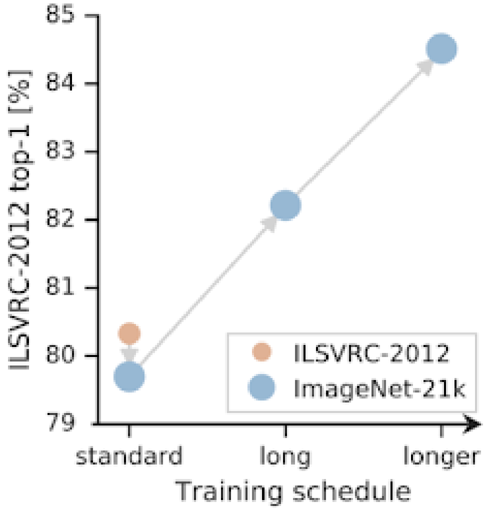


BiT (Big Transfer)[1]

- A pre-training recipe
- Large datasets, large models
- Replace Batch Normalisation with Group Normalisation and Weight Standardization
- We use BiT-M-R152x2

Model	ILSVRC2012	ImageNet ReaL
SOTA	90.45	91.12
BiT-L	87.54	90.54
BiT-M	85.39	89.02
BiT-S	81.30	86.21





[1] Big Transfer (BiT): General Visual Representation Learning https://arxiv.org/pdf/1912.11370.pdf
Images source: https://ai.googleblog.com/2020/05/open-sourcing-bit-exploring-large-scale.html

Baseline scores

Model	Test Target	Adult PRAUC	Weapon PRAUC
Text	Text label	0.9200	0.9477
Text	Overall label	0.7539	0.9462
Image	Image label	0.9328	0.8813
Image	Overall label	0.8755	0.8651
Combined	Overall label	0.9522	0.9381

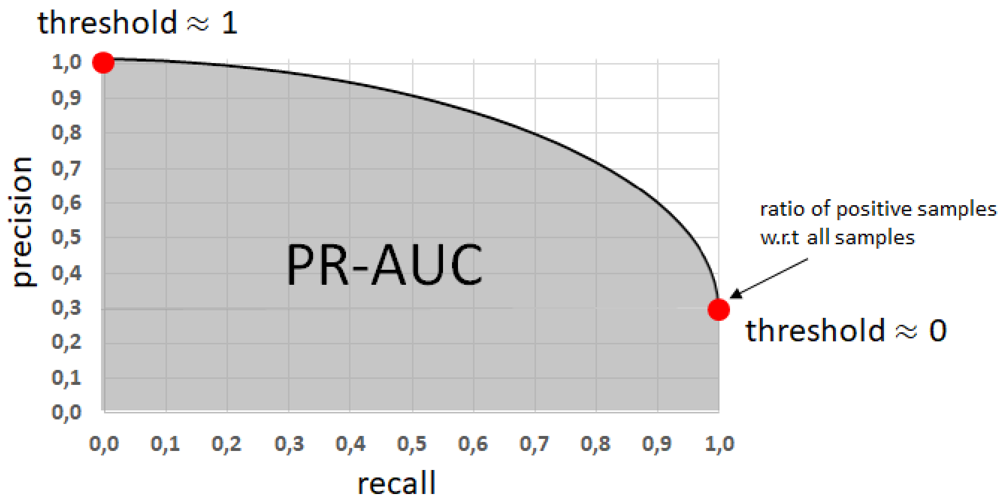


Image source: https://towardsdatascience.com/gaining-an-intuitive-understanding-of-precision-and-recall-3b9df37804a7

Details

- ➢ Text model: CULR-Large v1, a multilingual BERT-like model
- ➢ Data: ~8.6% adult, ~6.3% weapon

Multimodal Learning

Image captioning

The man at bat readies to swing at the pitch while the umpire looks on.

Image source: COCO Captions 2015, https://cocodataset.org/#captions-2015

Visual Question Answering

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Image source: VQA Dataset, https://visualqa.org/

Multimodal Learning

- Learn cross-modal representations by large-scale pre-training
- Concatenate image region features and text features to form input
- Use self-attention to implicitly learn image-text semantic alignments
- Image region features extracted from pre-trained Object Detection (OD) models
- Self-supervised pre-training objectives
  - Masked Language Modeling (MLM)
  - Image-caption matching

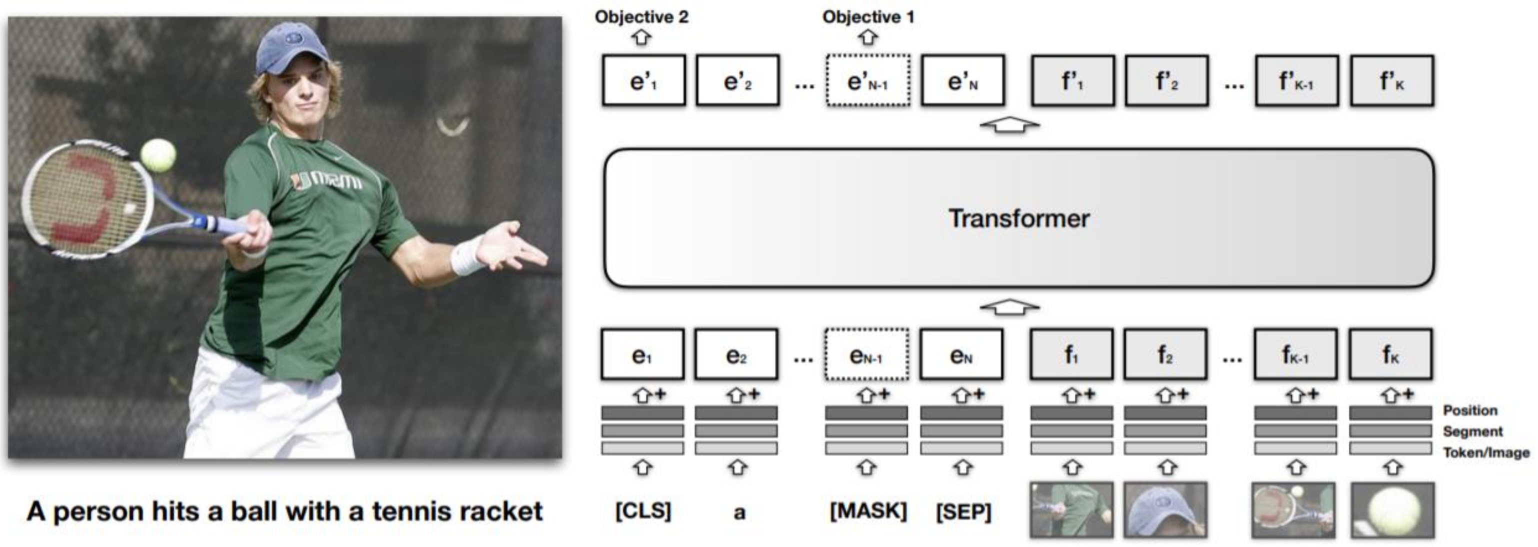
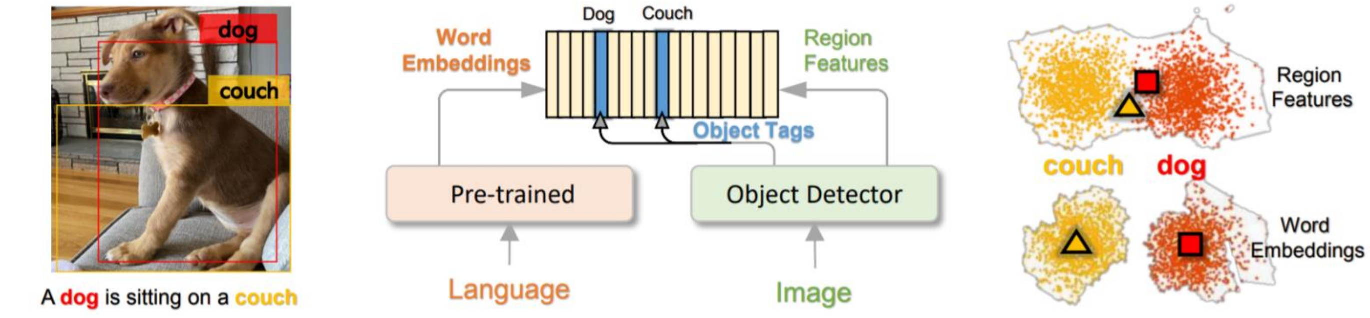


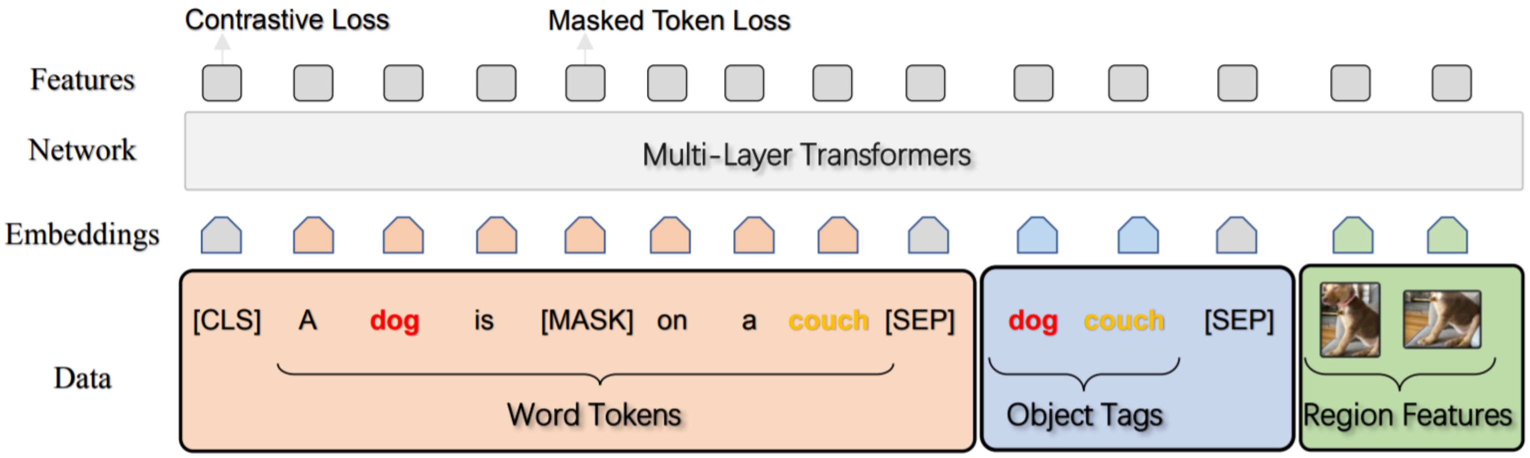
Image source: VisualBERT: A Simple and Performant Baseline for Vision and Language, https://arxiv.org/pdf/1908.03557.pdf

Oscar[1]

- Key idea – use object tags as *anchor points* to ease learning of alignment
- Motivation – salient objects detected in images are often mentioned in the paired text



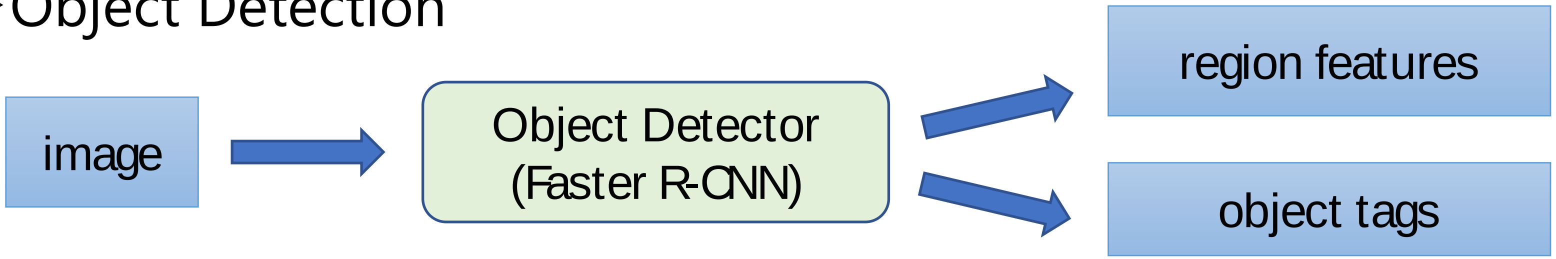
(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces



[1] Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, https://arxiv.org/pdf/2004.06165.pdf

Oscar: Input Representation

- Object Detection



- Input – each image-text pair is represented as a tuple `(w,q,v)`
  - `w` :- sequence of word embeddings of the text
  - `q` :- sequence of word embeddings of the object tags
  - `v` :- sequence of region features
- Idea (to summarize): Both `q` and `w` share the same semantic space (BERT initialisation), so their alignments are easy to identify
- So image regions `v` corresponding to relevant object tags are likely to have higher attention weights when queried by related words in `q`

Oscar: Pre-training Objectives

Masked Token Loss (Dictionary view) - L_M

- Mask each input token in h=concat(w,q) with 15% probability
- Goal of training – predict masked tokens using surrounding (partial) text and image context

Contrastive loss (Modality view) - L_C

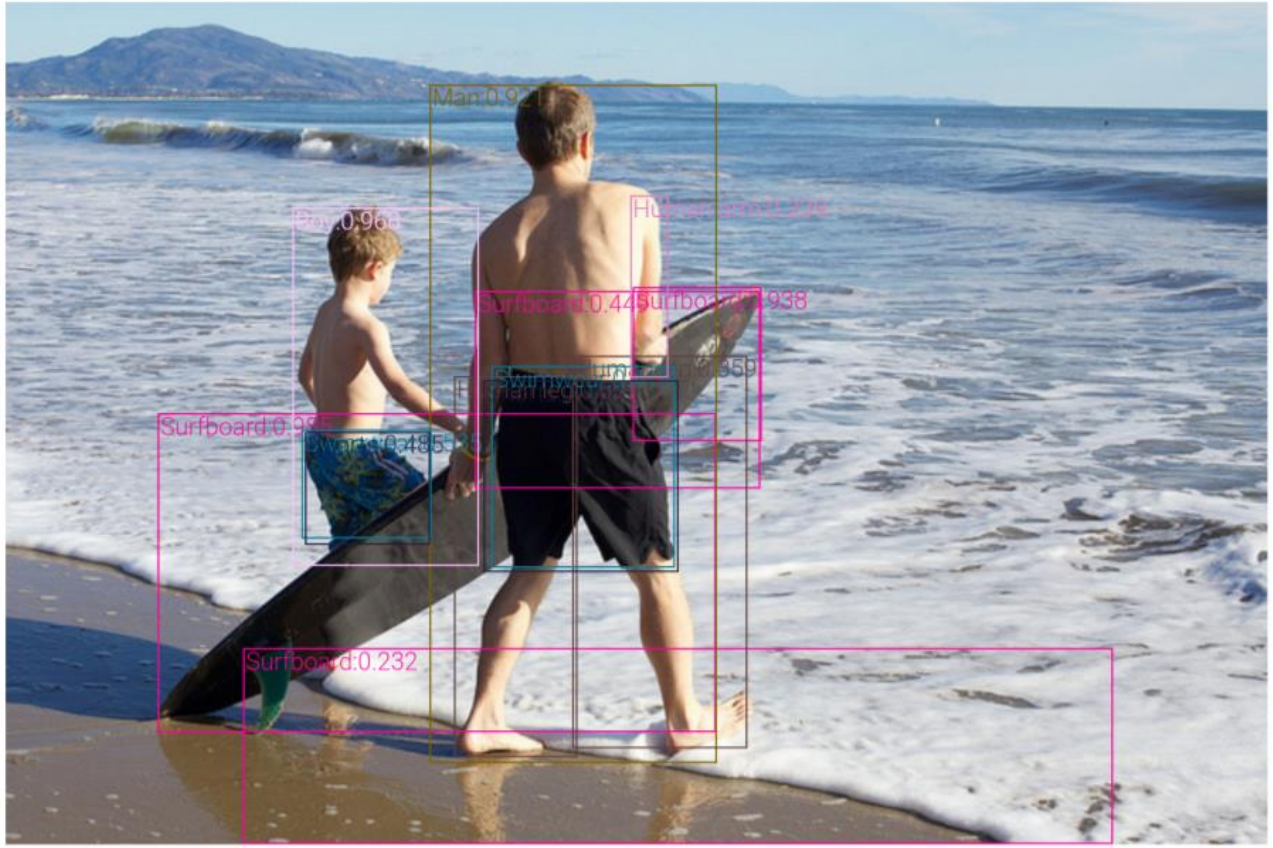
- Sample a set of “polluted” image representations by replacing q with 50% probability with a randomly sampled tag sequence q′
- Idea – Utilize object tags as a proxy for images
- Goal of training – predict whether triplet is polluted (using the [CLS] token)

Full pre-training objective

- $L_{pre\text{-}training} = L_M + L_C$

Oscar+ novel 3-way contrastive loss – L_{CL3}

- Create two types of polluted training samples – (w′,q,v) and (w,q′,v)
- Idea – (w′,q,v) is polluted captions, (w,q′,v) is polluted answers
- Goal of training – predict type of pollution in triplet (using the [CLS] token)



VinVL[1]

- ➢An improved OD model for VL
- ➢Model based on ResNeXt-152 C4 architecture
- ➢Trained on a much larger composite dataset – 1848 classes
- ➢Datasets employed – COCO, OpenImages, Objects365, VisualGenome

[1] VinVL: Revisiting Visual Representations in Vision-Language Models, https://arxiv.org/pdf/2101.00529.pdf

Object Detection Examples

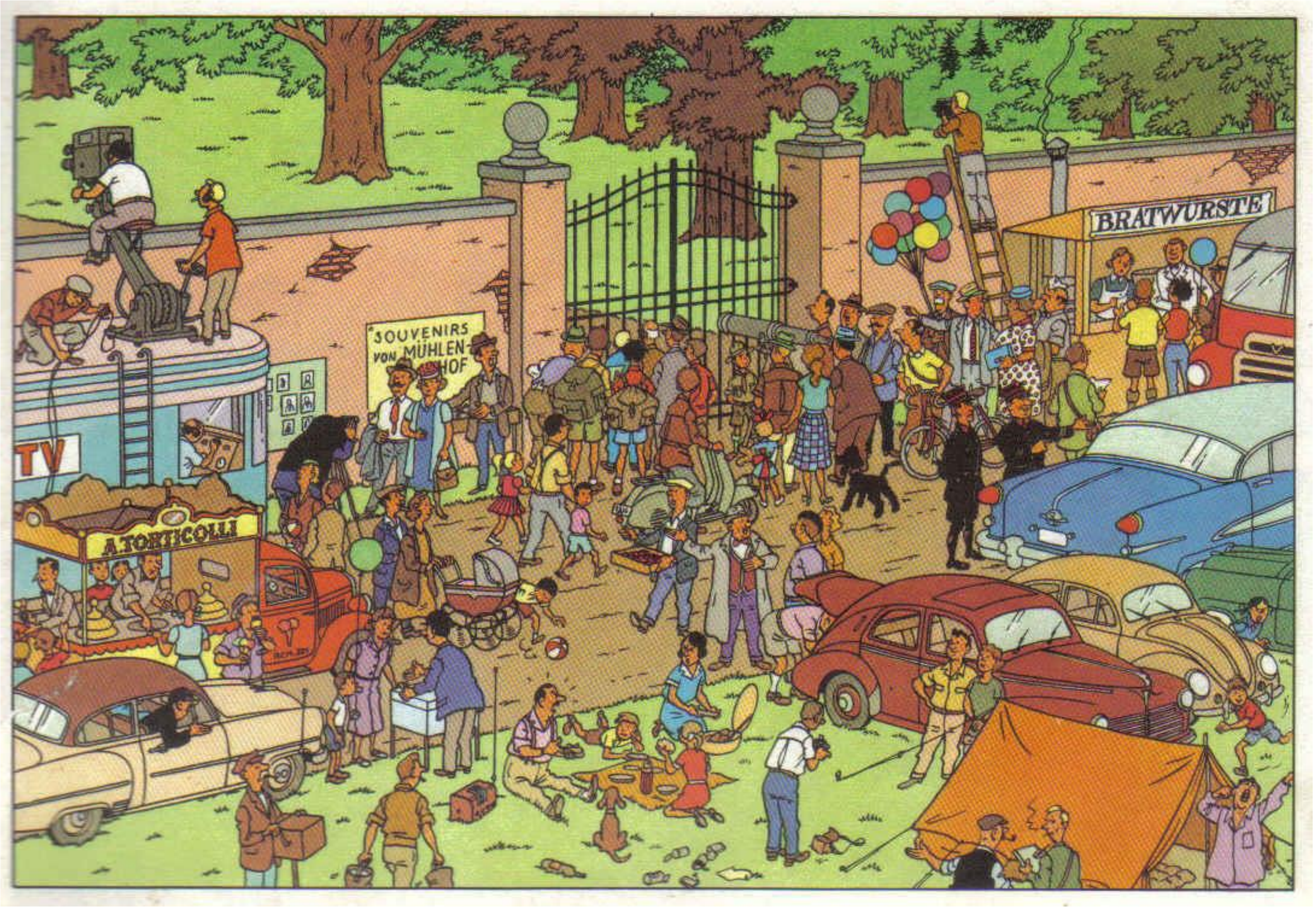
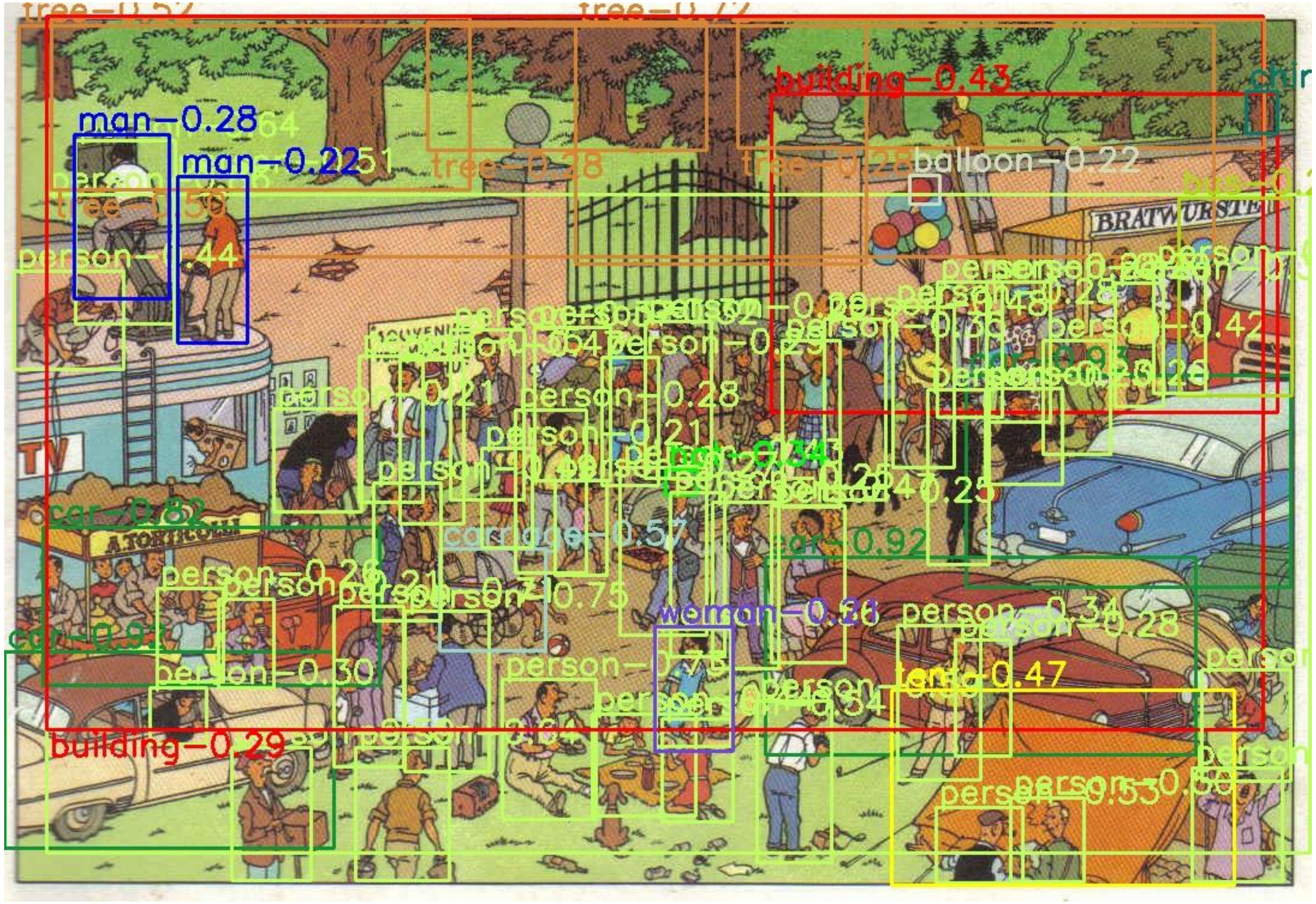


Image source: https://www.pinterest.com.au/pin/607774912215821850/

Object Detection Examples



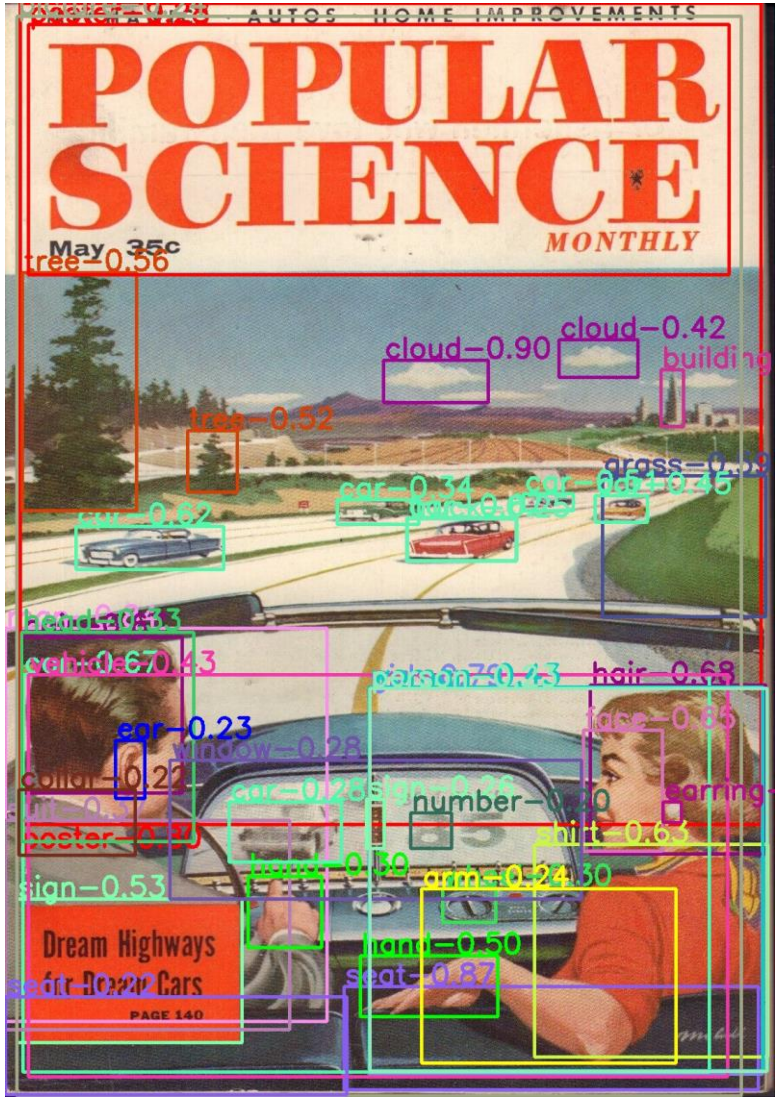


Image source: https://www.pinterest.com.au/pin/607774912215821850/

Object Detection Examples

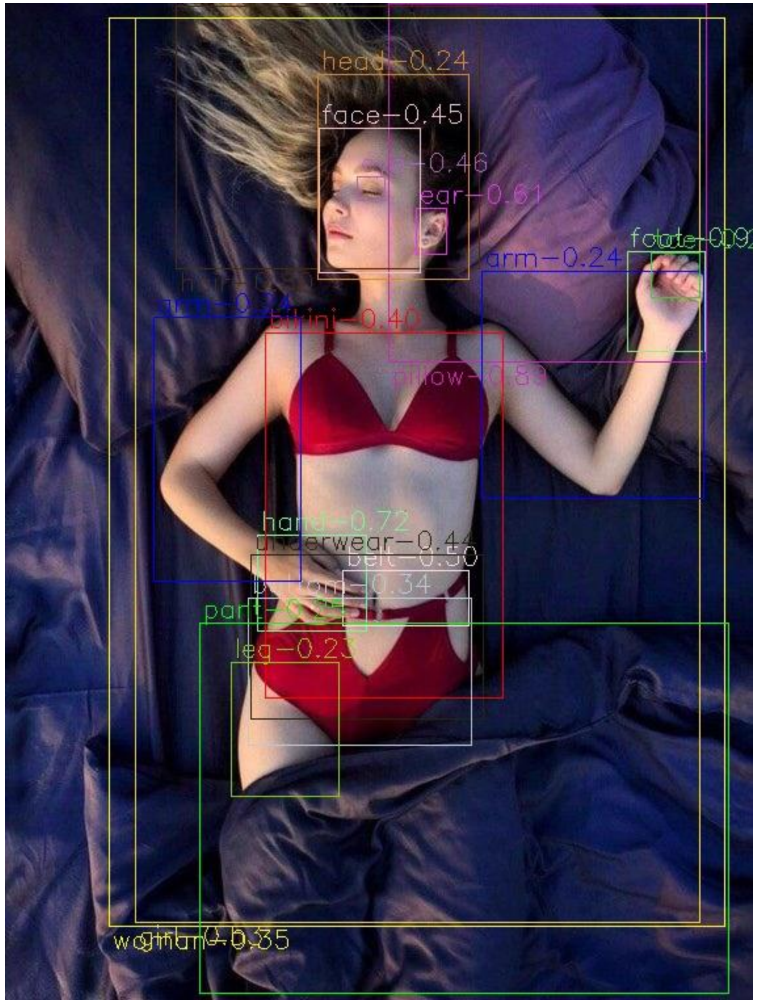
- Most ads have gun accessories – OD model not built to detect

screen light hole device button handle box

box cap pen word sticker word letter logo

gun bracket cap table screw toy handle light

Object Detection Examples



PA Editorial Experiments

Model	OD tag len	LR (With Linear Schedule)	Adult PRAUC	Weapon PRAUC
VinVL base	0	2e-5	0.9088	0.9319
VinVL base	10	2e-5	0.9071	0.9335
VinVL base	20	2e-5	0.9078	0.9331
VinVL base	20	2.5e-5	0.9161	0.9310
VinVL large	20	2e-5	0.9110	0.9322

PA Editorial Results

Model	Target : Overall Disallowed	
	Adult PRAUC	**Weapon PRAUC**
Unimodal NN baseline	0.9521	0.9380
VinVL base	0.9078	0.9331
Multimodal base	0.9518	0.9364
VinVL large	0.9110	0.9322
Multimodal large	0.9516	0.9365

- Weapon Class: obtaining similar results
- Adult Class: significant dip
- Adult is heavily dependent on Image Modality and there is scope of improvement there.
    - Objects tags are oversampled and not refined for Ads data.
    - Object Detector is not finetuned on Ads data.
- VinVL uses BERT initialization and our baselines are using CULR as the text model

Future Work

- Do Oscar+ pre-training on newer models (XLM-R, CULR, etc)
- Explore adding an image token representing whole image (BiT penultimate layer)
- Policy-based domain-specific object tags

Thanks

Open To Questions

Appendix

Multimodal Learning

- Attention weights of some selected heads in VisualBERT
  - Implicit grounding of visual concepts in higher layers (eg "man wearing white shirt")
  - Refinement of understanding (eg "walking"-*Pedestrians*)
  - Correction of alignment (eg "shirt"-*Man*)

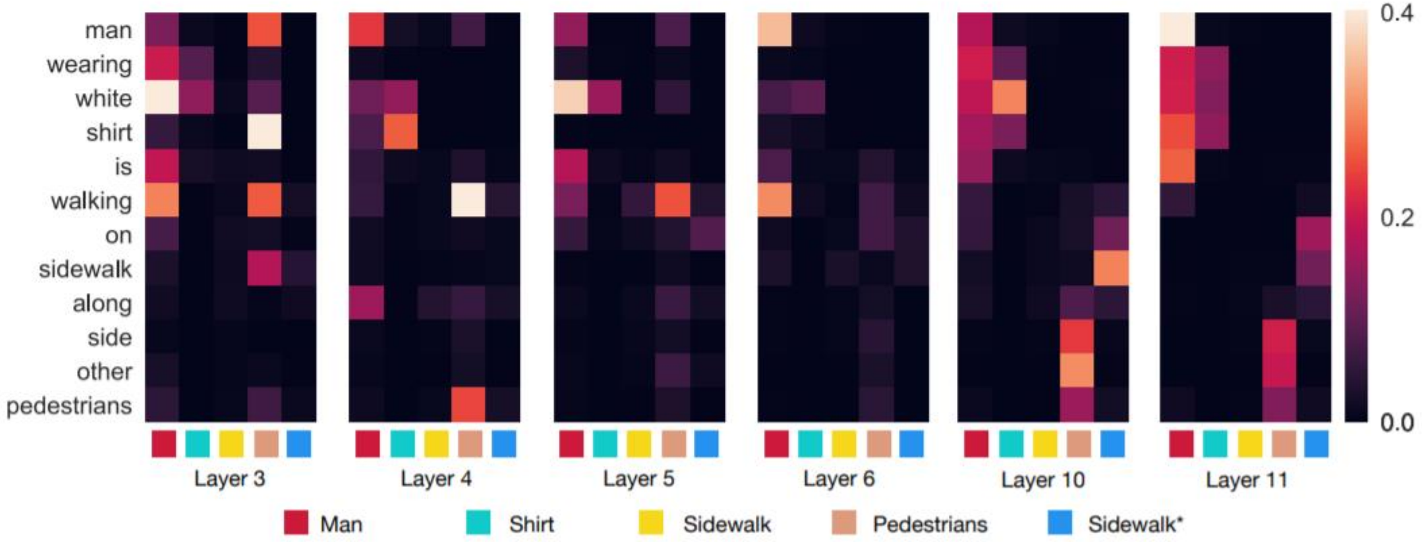


Image source: VisualBERT: A Simple and Performant Baseline for Vision and Language, https://arxiv.org/pdf/1908.03557.pdf

Data Distribution

Overall AdDisallowed	Train	Val	GoldenTest
Compliance	950K	230K	59k
Adult	148K	27K	6.3k
Weapon	72K	16K	5k