What Does BERT with Vision Look At?

Liunian Harold Li
UCLA

Mark Yatskar
AI2

Da Yin
PKU

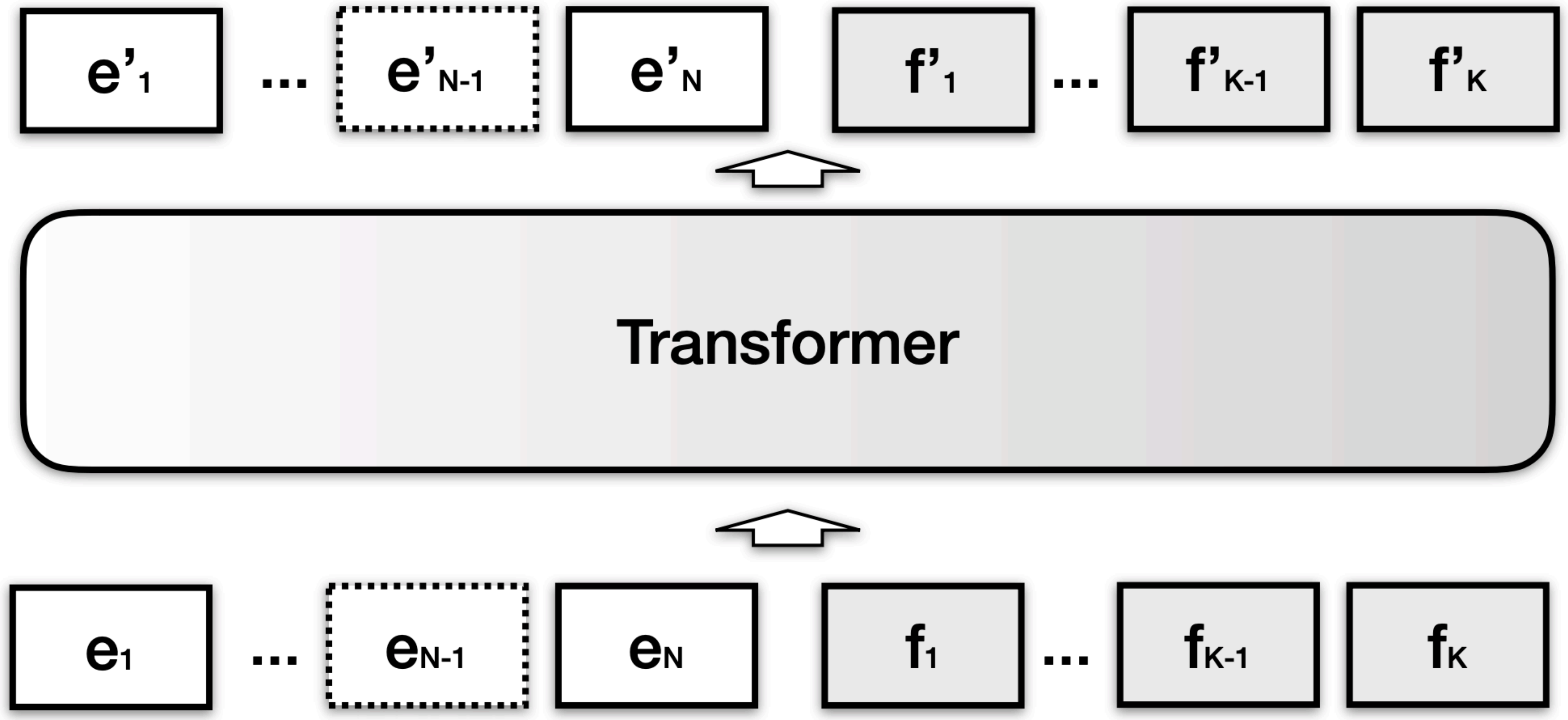
Cho-Jui Hsieh
UCLA

Kai-Wei Chang
UCLA

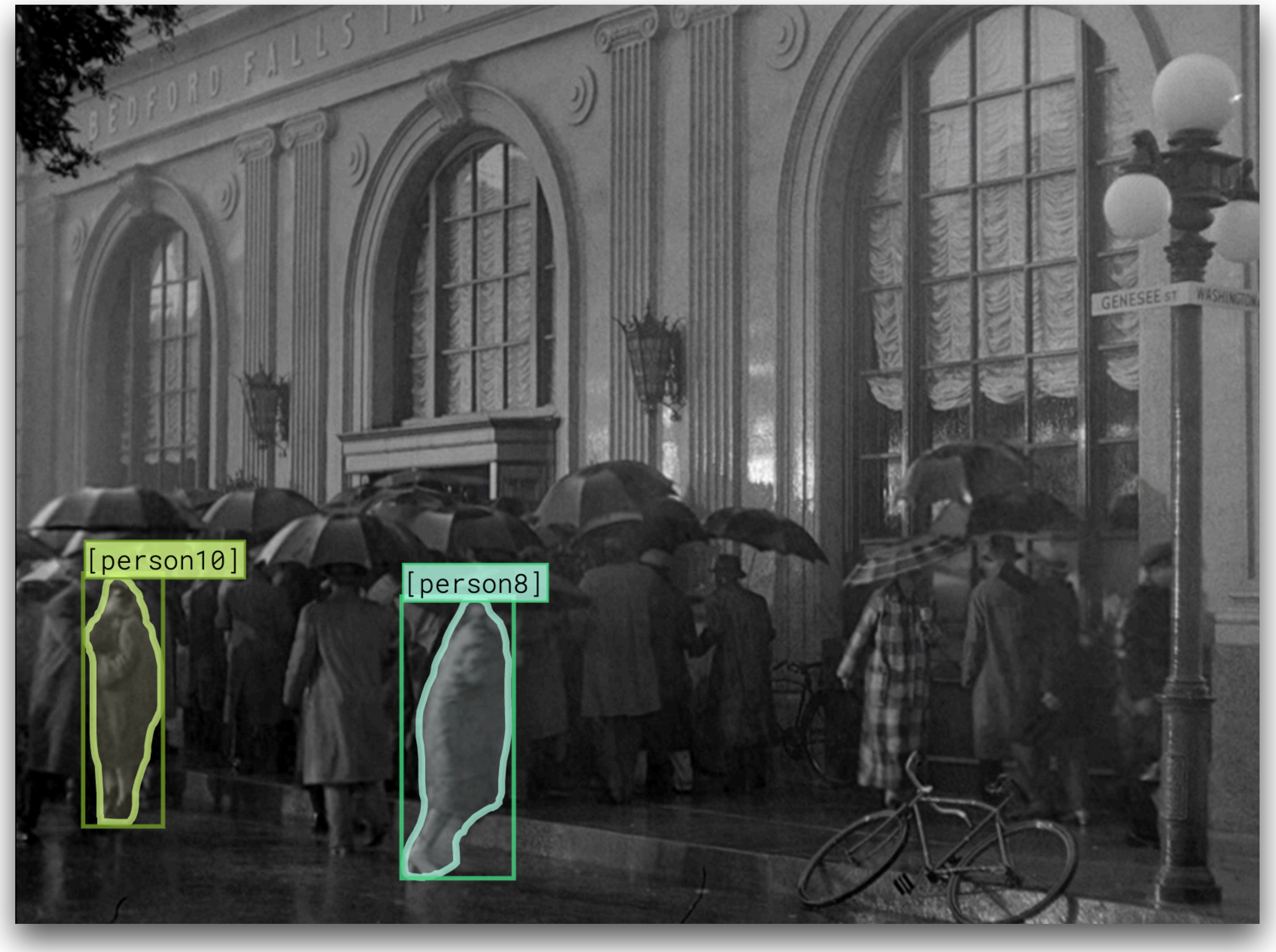
A long version, "VisualBERT: A Simple and Performant Baseline for Vision and Language" is on Arxiv (Aug 2019).

BERT with Vision: Pre-trained Vision-and-language (V&L) Models

Several people walking on a sidewalk in the rain with umbrellas.



Several people [MASK] on a [MASK] in the [MASK] with [MASK].



a) Yes, it is snowing.

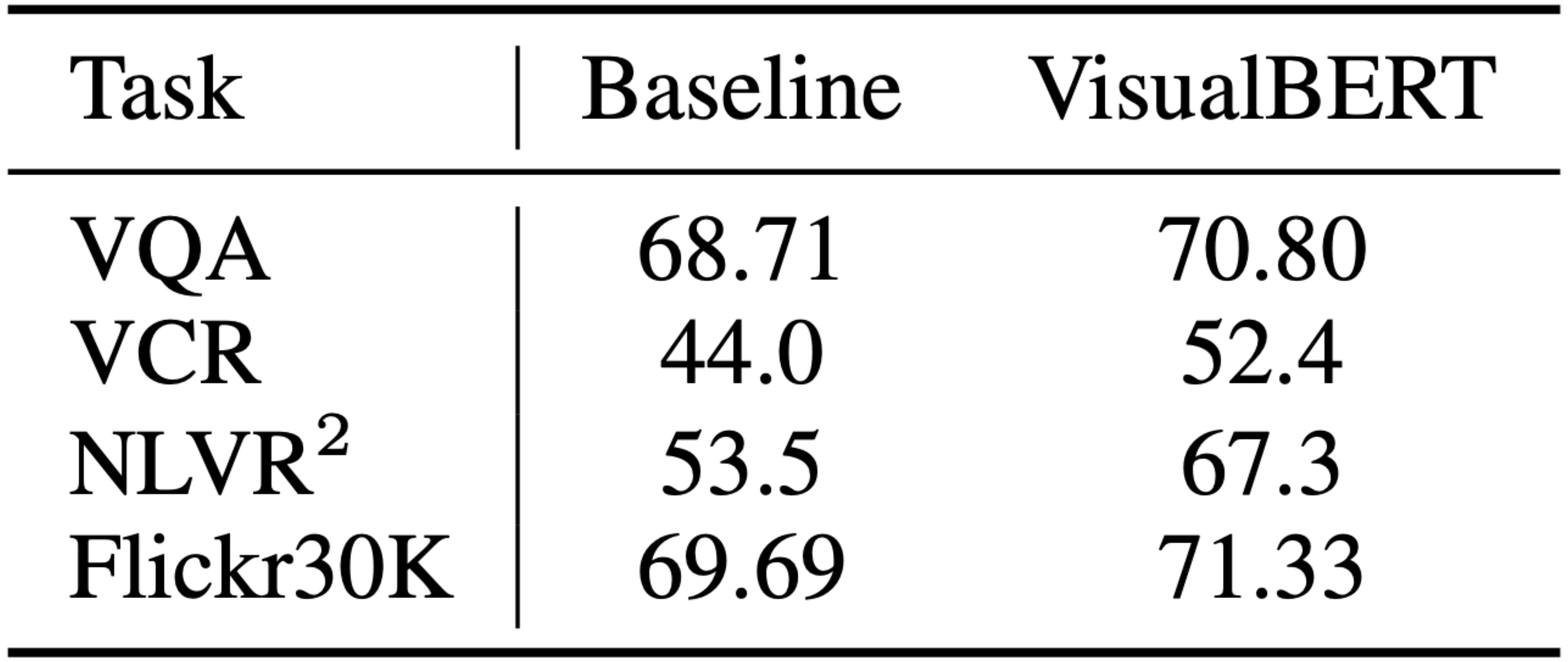
b) Yes, [person8] and [person10] are outside.

c) No, it looks to be fall.

d) Yes, it is raining heavily.

Pre-train on image captions and transfer to visual question answering

BERT with Vision: Pre-trained Vision-and-language (V&L) Models



Task	Baseline	VisualBERT
VQA	68.71	70.80
VCR	44.0	52.4
NLVR^2	53.5	67.3
Flickr30K	69.69	71.33

Performance of VisualBERT compared to strong baselines

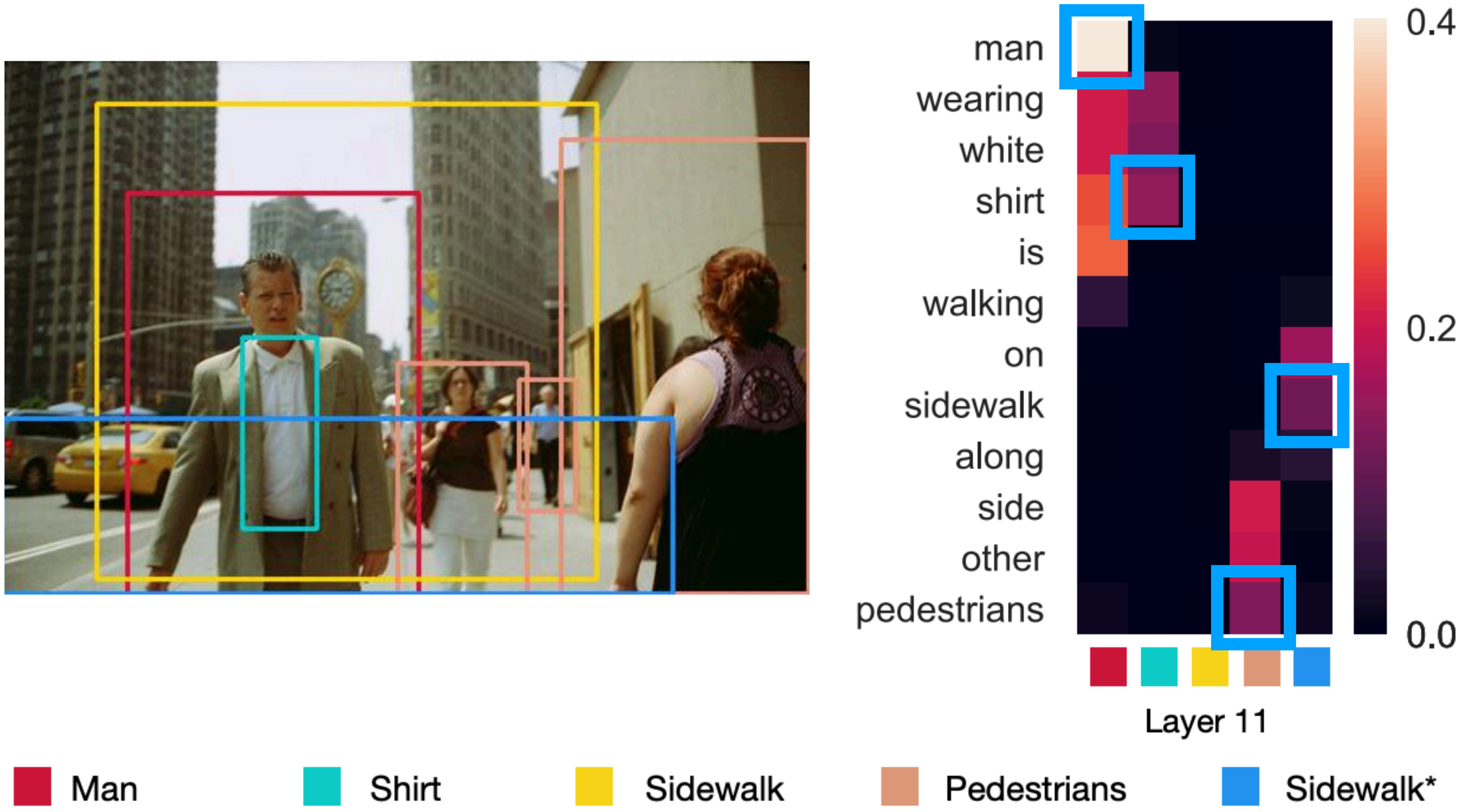
Mask and predict on image captions

Transformer over image regions and texts

Significant improvement over baselines

ViLBERT, B2T2, LXMERT, VisualBERT, Unicoder-VL, VL-BERT, UNITER, …

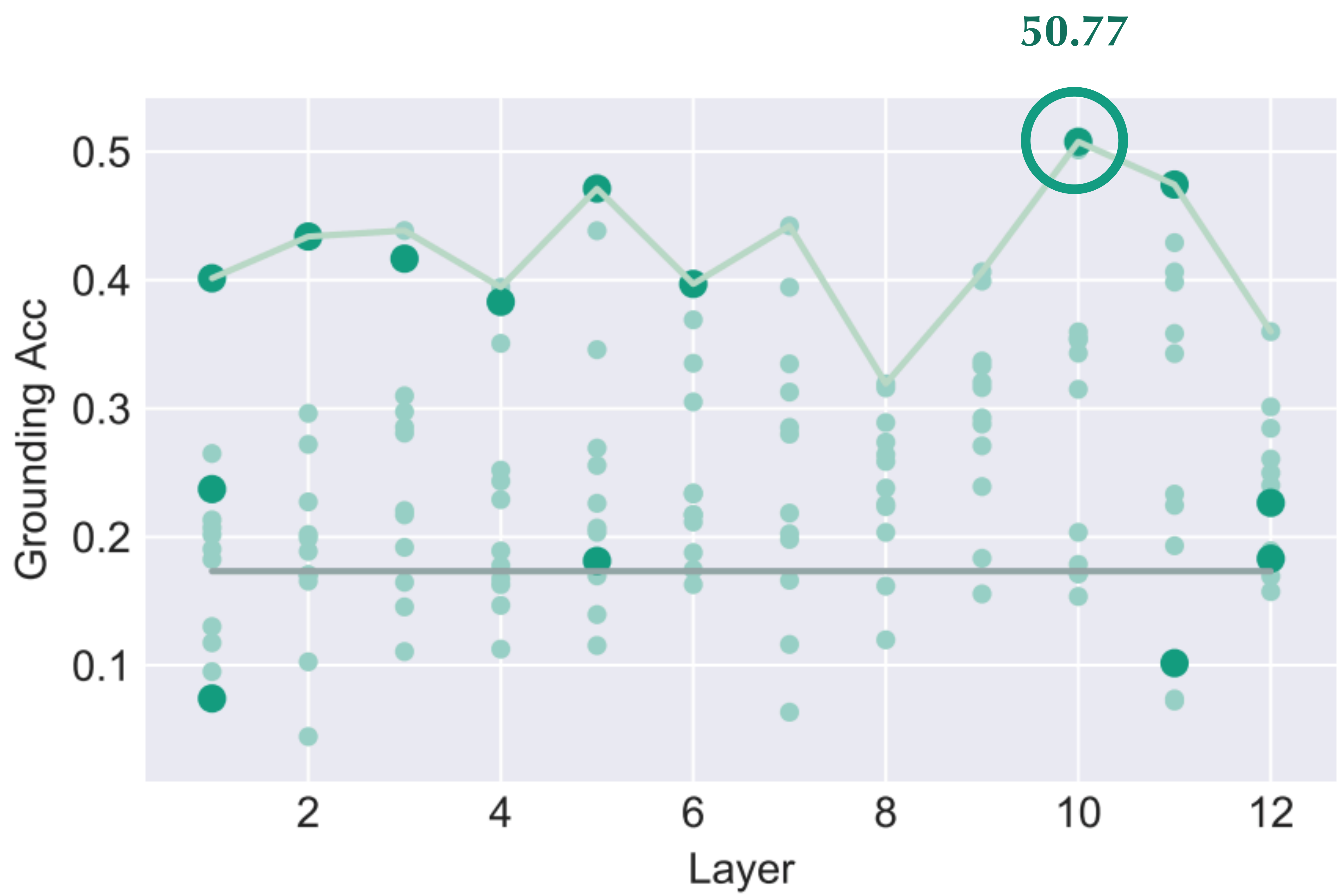
What does BERT with Vision learn during pre-training?



Entity grounding

Map entities to regions

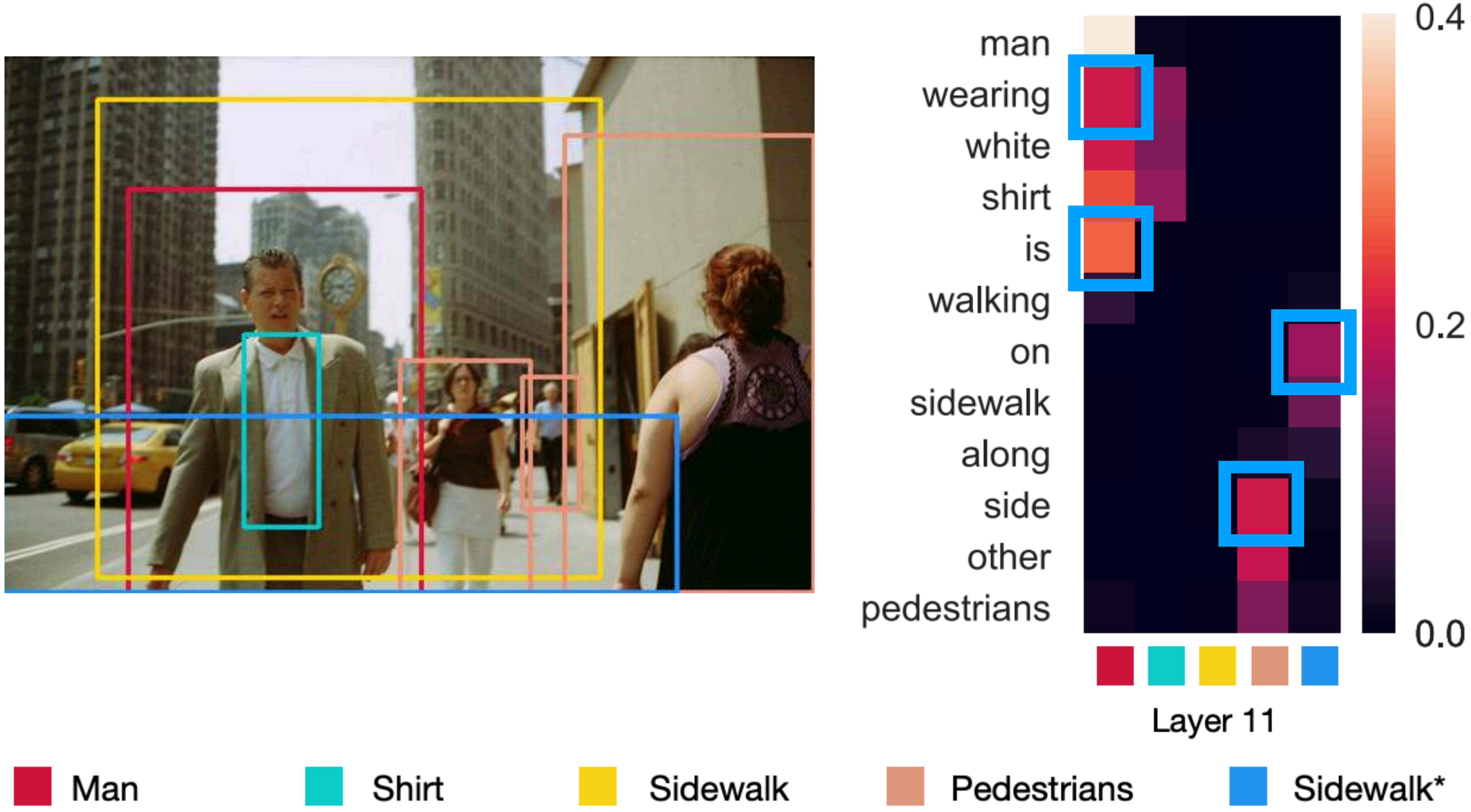
Probing attention maps of VisualBERT: Entity Grounding



Certain heads can perform entity grounding

Accuracy peaks in higher layers

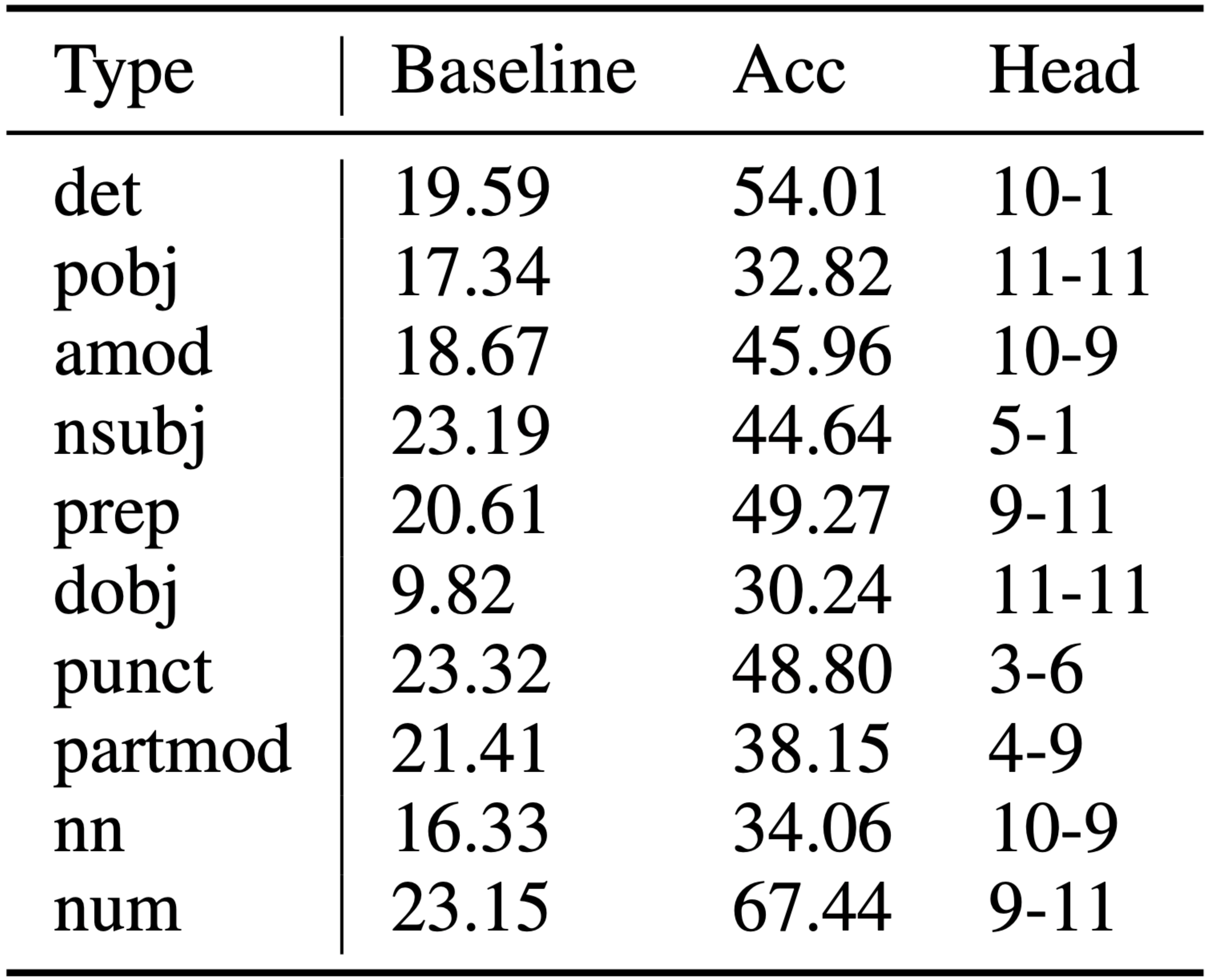
What does BERT with Vision learn during pre-training?



Syntactic grounding

Map w_1 to regions of w_2, if $w_1 \overset{r}{\longleftrightarrow} w_2$

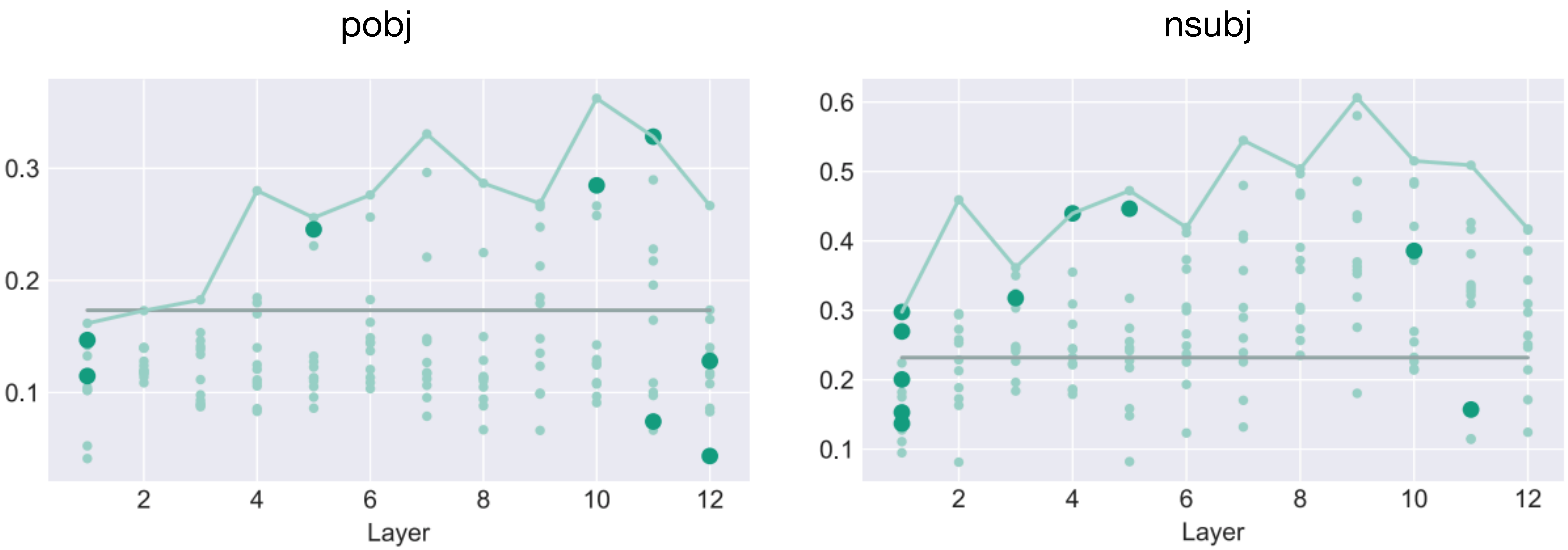
Probing attention maps of VisualBERT: Syntactic Grounding



Type	Baseline	Acc	Head
det	19.59	54.01	10-1
pobj	17.34	32.82	11-11
amod	18.67	45.96	10-9
nsubj	23.19	44.64	5-1
prep	20.61	49.27	9-11
dobj	9.82	30.24	11-11
punct	23.32	48.80	3-6
partmod	21.41	38.15	4-9
nn	16.33	34.06	10-9
num	23.15	67.44	9-11

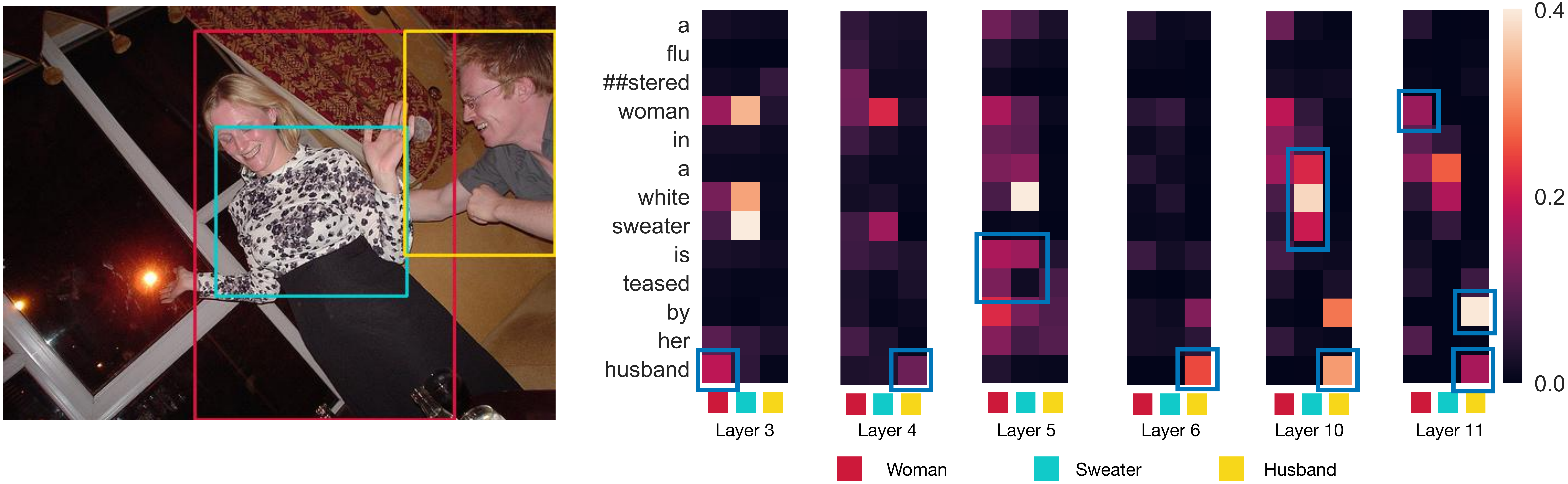
For each dependency relationship, there exists at least one accurate syntax grounding head

Probing attention maps of VisualBERT: Syntactic Grounding



Syntactic grounding accuracy peaks in higher layers

Probing attention maps of VisualBERT: Qualitative Example



Accurate entity and syntax grounding

Refined understanding over the layers

Discussion

Previous work

Pre-trained language models learn the classical NLP pipeline *(Peters et al., 2018; Liu et al., 2019; Tenney et al., 2019)*

Qualitatively, V&L models learn some entity grounding *(Yang et al., 2016; Anderson et al., 2018; Kim et al., 2018)*

Grounding can be learned using dedicated methods *(Xiao et al., 2017; Datta et al., 2019)*

Our paper

BERT with Vision learns grounding through pre-training

We quantitively verify both entity and syntactic grounding

https://github.com/uclanlp/visualbert