MoVie: Revisiting Modulated Convolutions for Visual Counting and Beyond

Duy-Kien Nguyen

Vedanuij Goswami

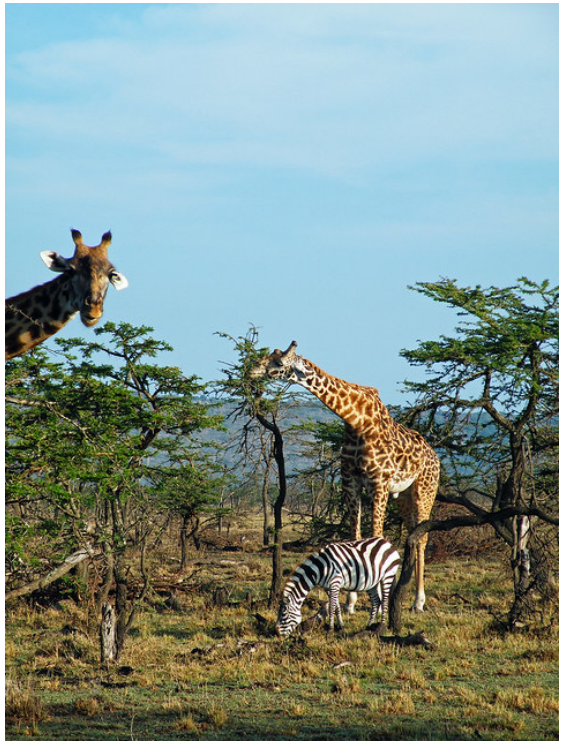
Xinlei Chen

ICLR

facebook AI Research

Visual Counting Problems

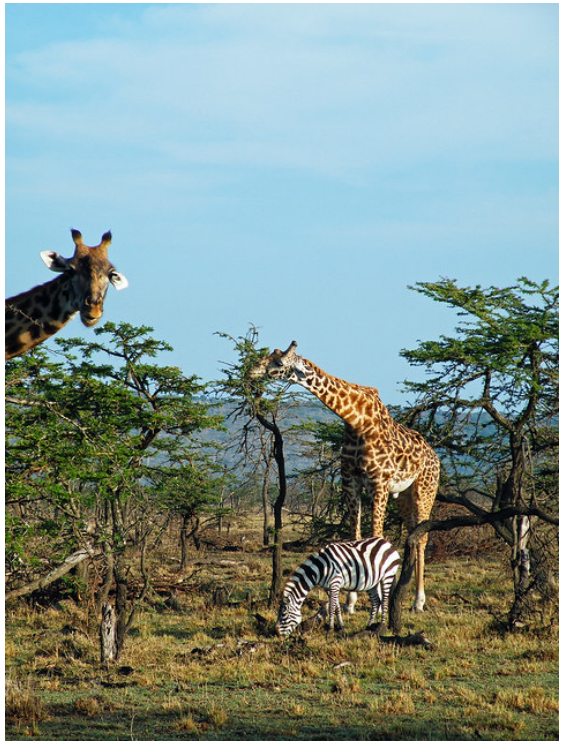
Open-Ended Counting



Query (Question)
How many giraffes are eating grass?
1

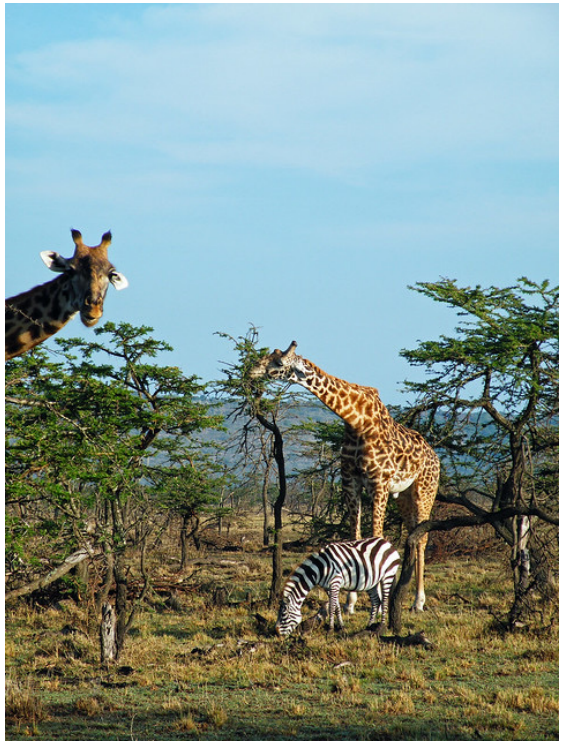
Visual Counting Problems

Open-Ended Counting



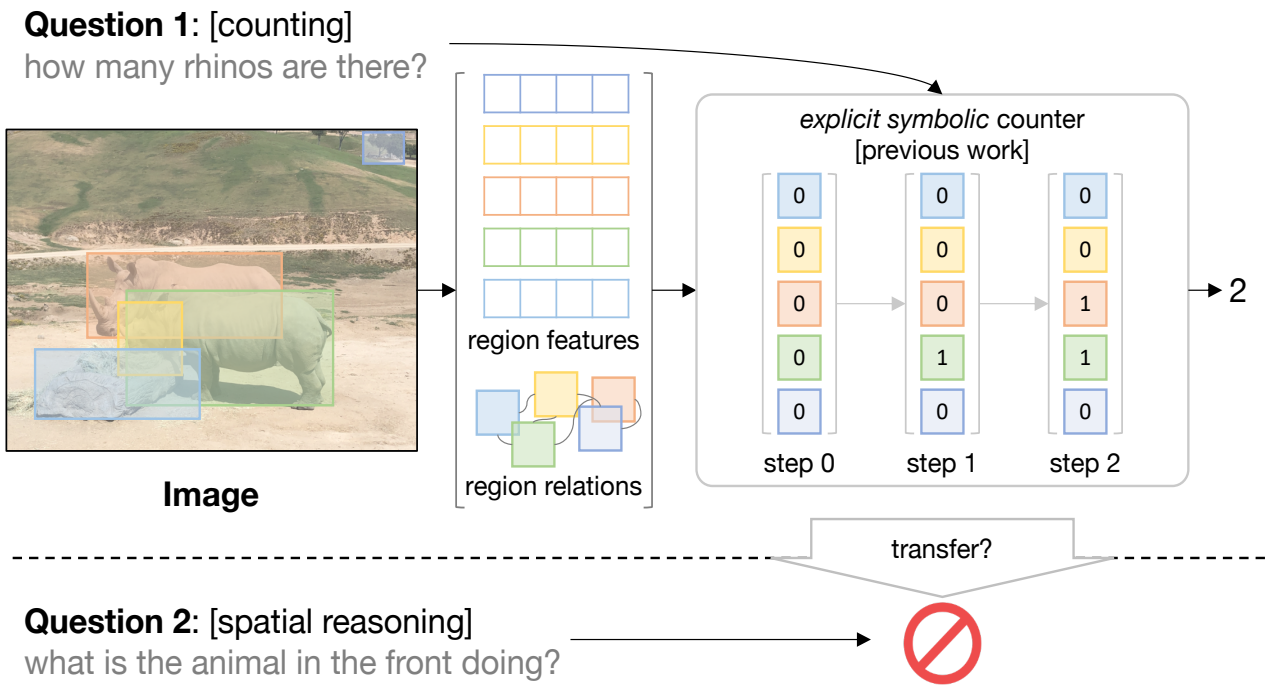
Query (Question)
How many giraffes are eating grass?
1

Common Objects Counting

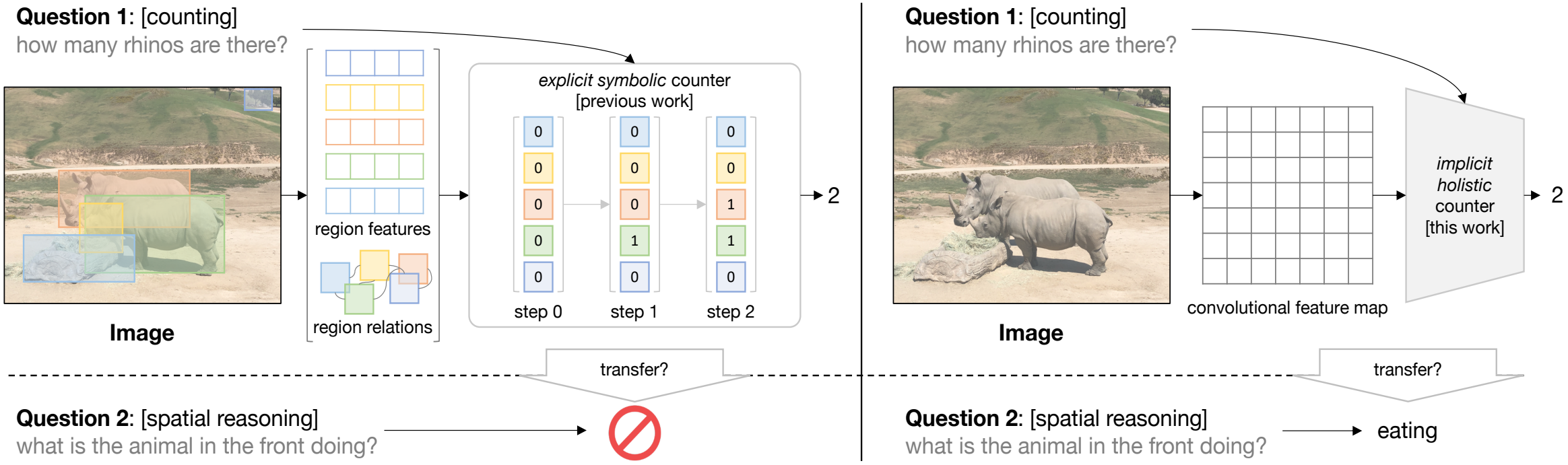


Query (Category)	
Giraffe	2
Zebra	1
People	0

Visual Counting Approaches



Visual Counting Approaches



Motivation

CNN

Pooling

Fusion

Answer
banana

Question
What is the mustache made of?

LSTM

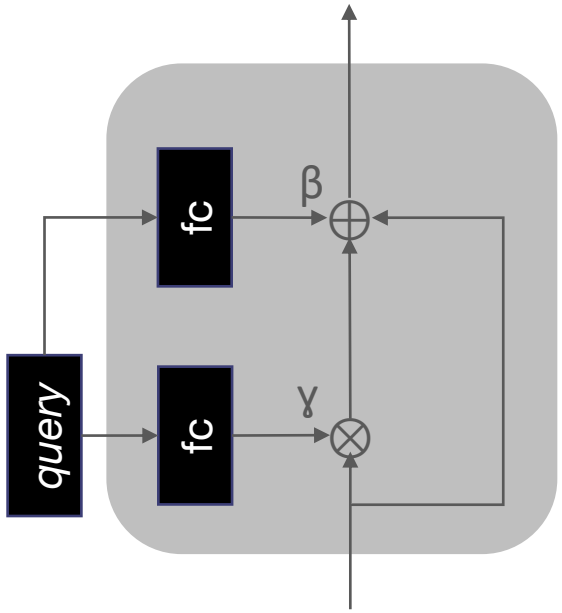
Pooling

Motivation

Counting Network needs:

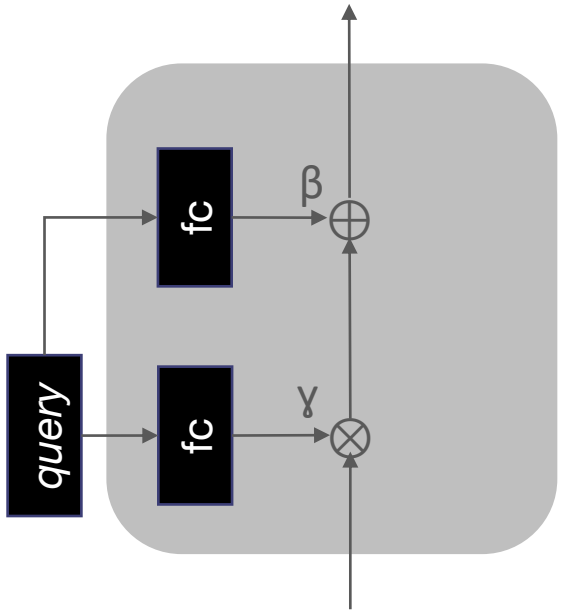
1. Spatial and local information for object occurrences
2. Translation equivariance across convolutional feature map

Feature-wise Linear Modulation [Perez et al]

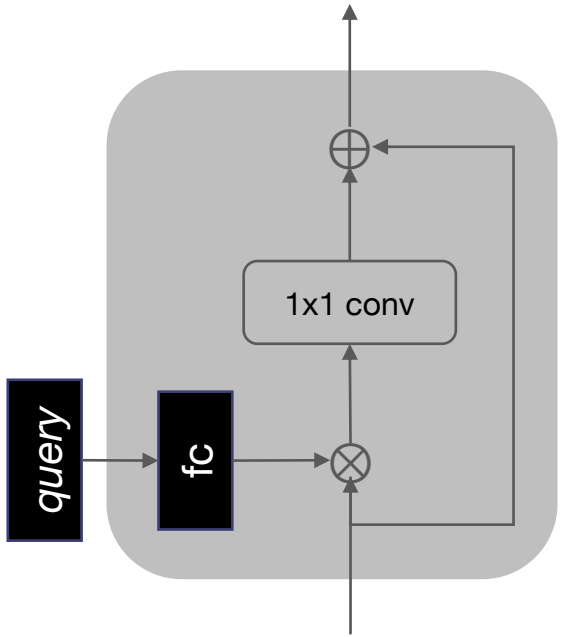


FiLM

Feature-wise Linear Modulation [Perez et al]



FiLM



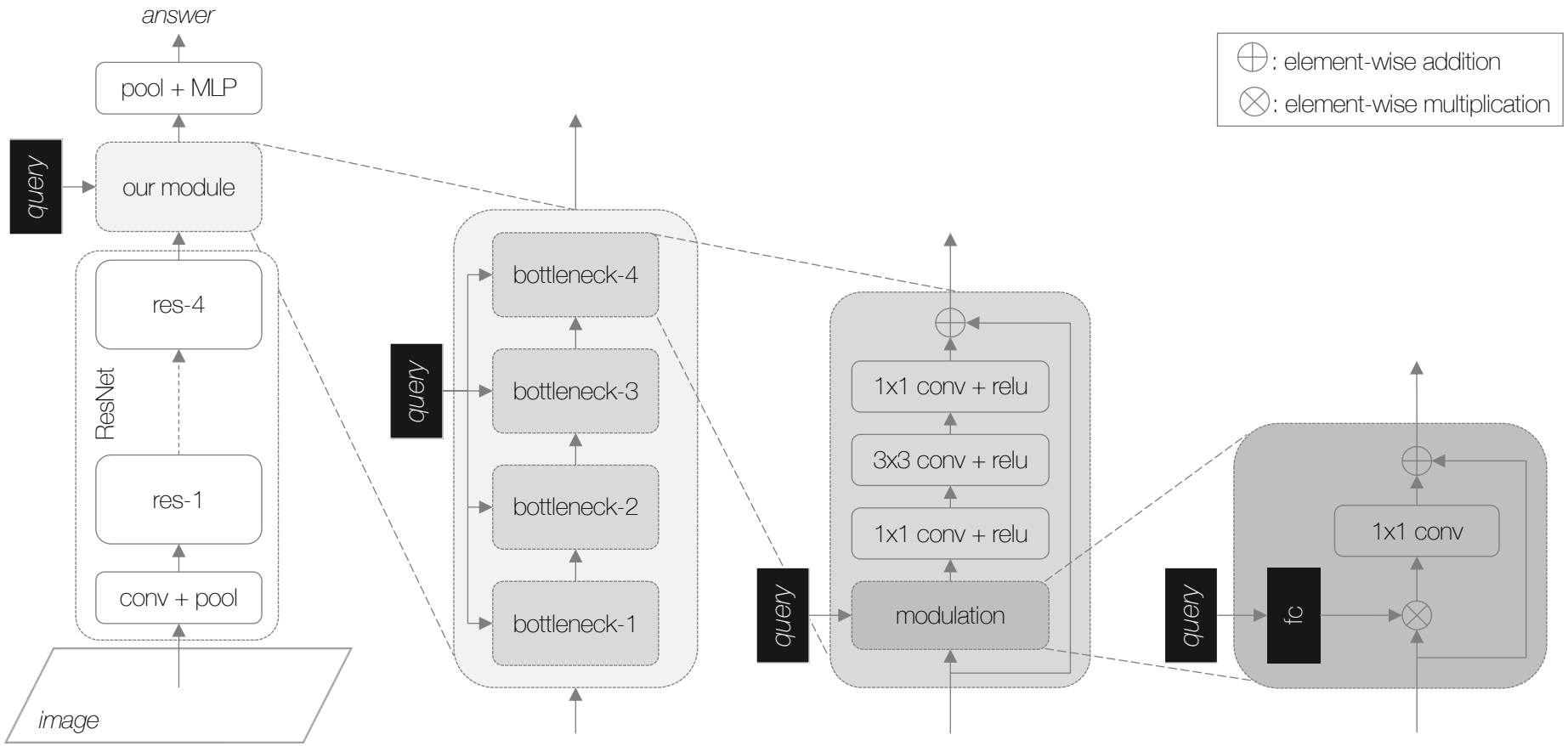
Ours

Modulated ConVolution Bottleneck

Advantages:

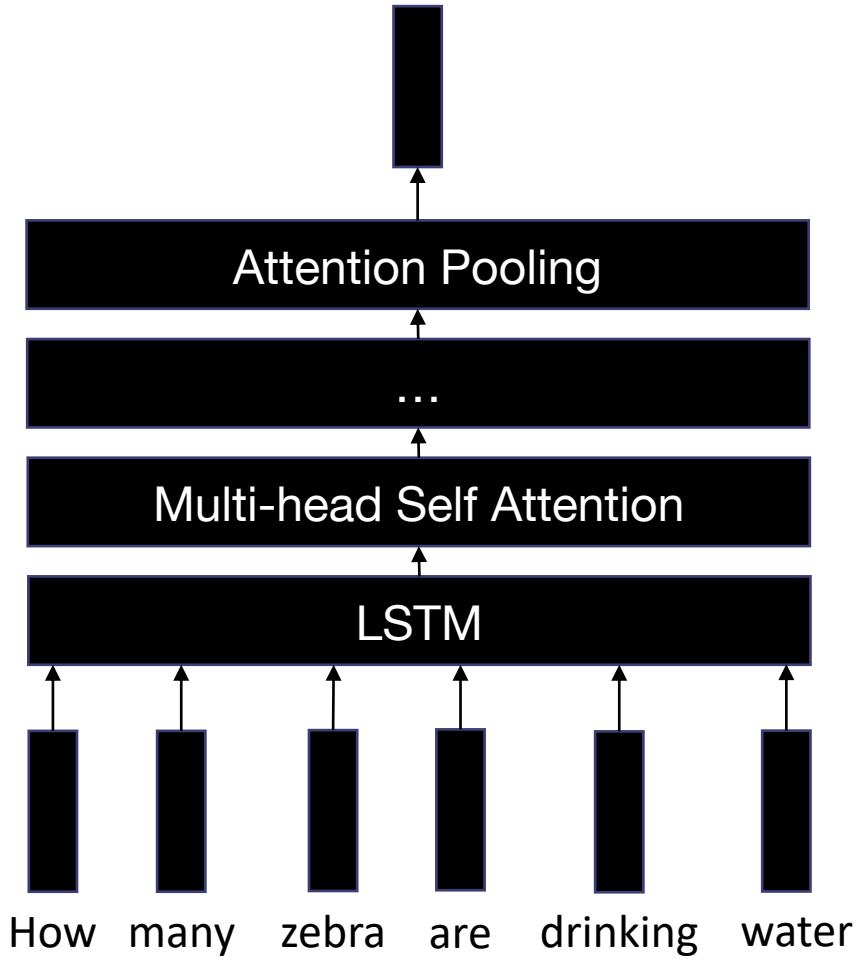
1. Translation Equivariant
2. Light-weight and Fast
3. Easy to attach to pretrained Resnet bottleneck
4. Accept different types of query in counting

Modulated ConVolution Bottleneck

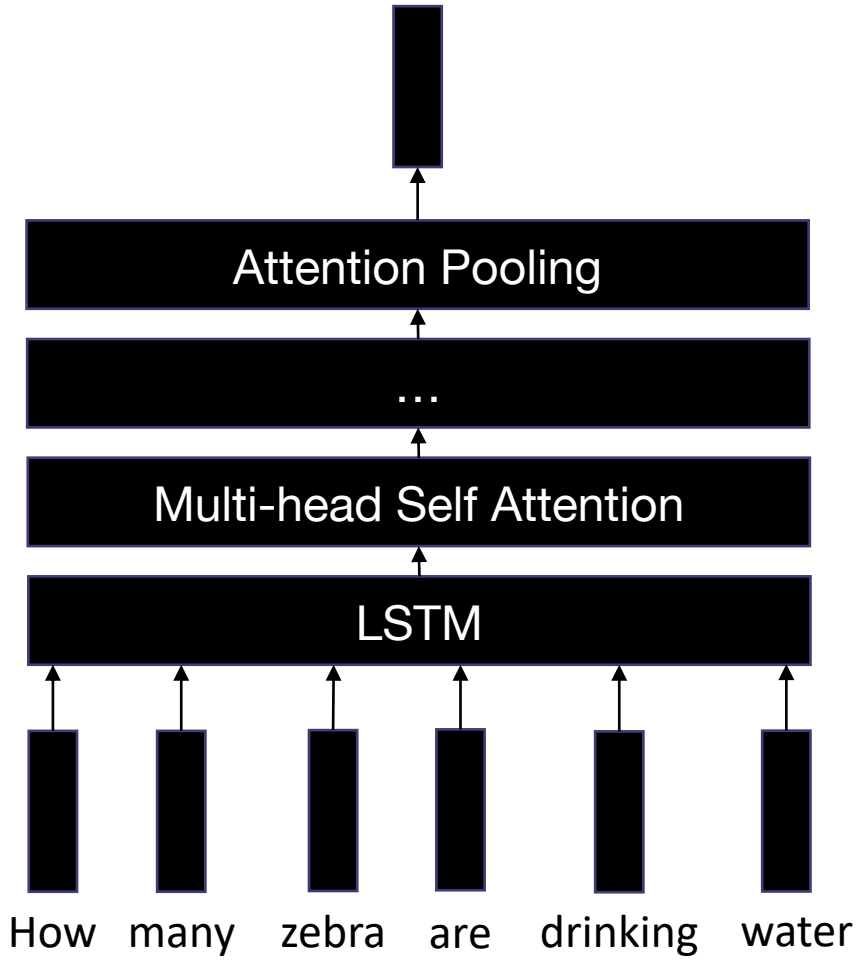


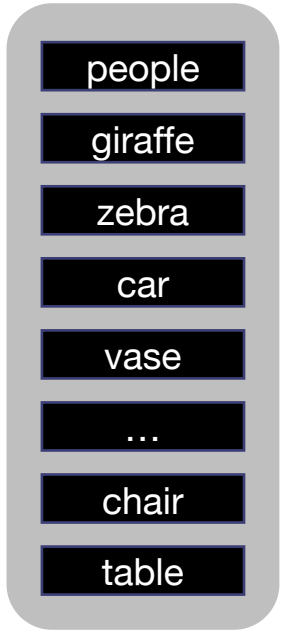
(a) pipeline (b) module (c) bottleneck (d) modulation

Query Representation



Query Representation





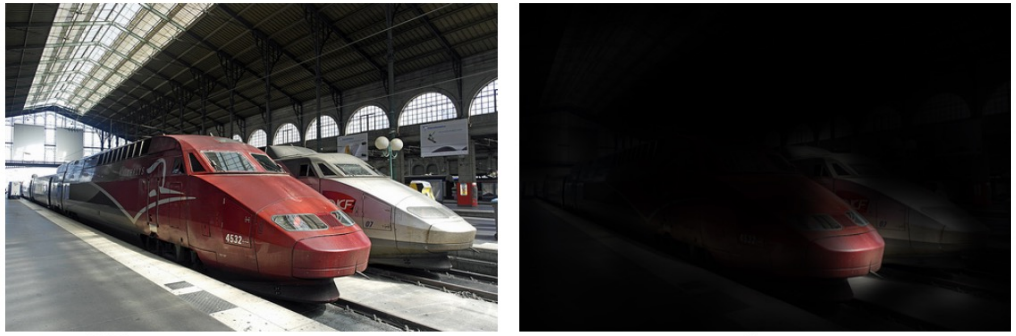
Embedding Vectors

Visual Counting Performance

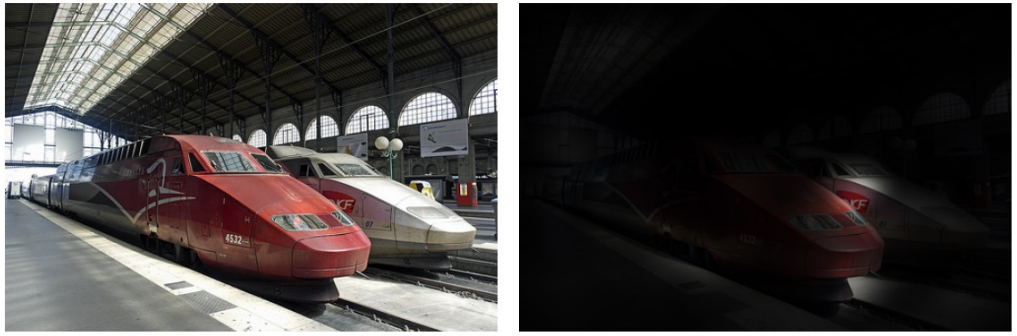
Performance on TallyQA and HowManyQA

Method	Backbone	#Params (M)	FLOPS (G)	HowMany-QA		TallyQA-Simple		TallyQA-Complex	
				ACC ↑	RMSE ↓	ACC ↑	RMSE ↓	ACC ↑	RMSE ↓
Counting module [Zhang et al.]	R-101	44.6	-	54.7	2.59	70.5	1.15	50.9	1.58
IRLC [Trott et al.]	R-101	44.6	1883.5	56.1	2.45	-	-	-	-
TallyQA [Acharya et al.]	R-101 + 152	104.8	1790.9	60.3	2.35	71.8	1.13	56.2	1.43
MoVie	R-50	25.6	176.1	61.2	2.36	70.8	1.09	54.1	1.52
MoVie	R-101	44.6	306.9	62.3	2.30	73.3	1.04	56.1	1.43
MoVie (ResNeXt)	X-101	88.8	706.3	64.0	2.30	74.9	1.00	56.8	1.43

Visualization in Counting



How many red trains are in the picture
How many red trains are in the picture
Pred: 1, Ans: 1



How many trains are in the picture
How many trains are in the picture
Pred: 2, Ans: 2

How many animals are laying down
How many animals are laying down
Pred: 1, Ans: 1

How many animals are there
How many animals are there
Pred: 8, Ans: 8

Visualization in Counting

How many of the birds are in the air
How many of the birds are in the air
Pred: 2, Ans: 1

How many road sign poles are there
How many road sign poles are there
Pred: 2, Ans: 3

How many of the pillows are gray
How many of the pillows are gray
Pred: 2, Ans: 0

How many seals are on the bench
How many seals are on the bench
Pred: 0, Ans: 1

Visual Counting Performance

Performance on COCO counting

Method	Instance Supervision	RMSE ↓	RMSE-nz ↓	rel-RMSE ↓	rel-RMSE-nz ↓
LC-ResFCN [Laradji et al.]	✓	0.38	2.20	0.19	0.99
glance-noft-2L [Chattopadhyay et al.]	✗	0.42	2.25	0.23	0.91
CountSeg [Cholakkal et al.]	✗	0.34	1.89	0.18	0.84
Faster R-CNN (Detectron2) [Wu et al.]	✓	0.35	1.88	0.18	0.80
MoVie	✗	0.30	1.49	0.19	0.67

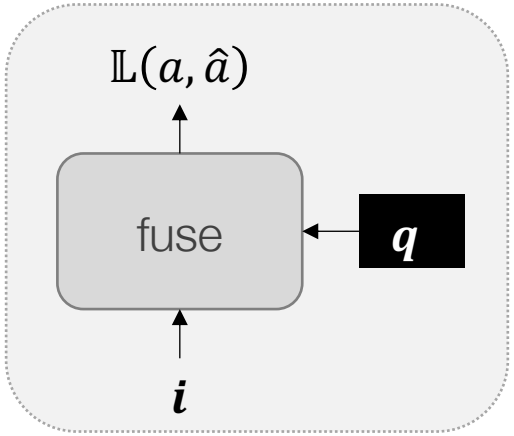
MoVie as a Counting Module in VQA

MoVie is attached to SotA VQA models in order to boost the counting performance

Pythia
(Winner VQA
Challenge 2018)

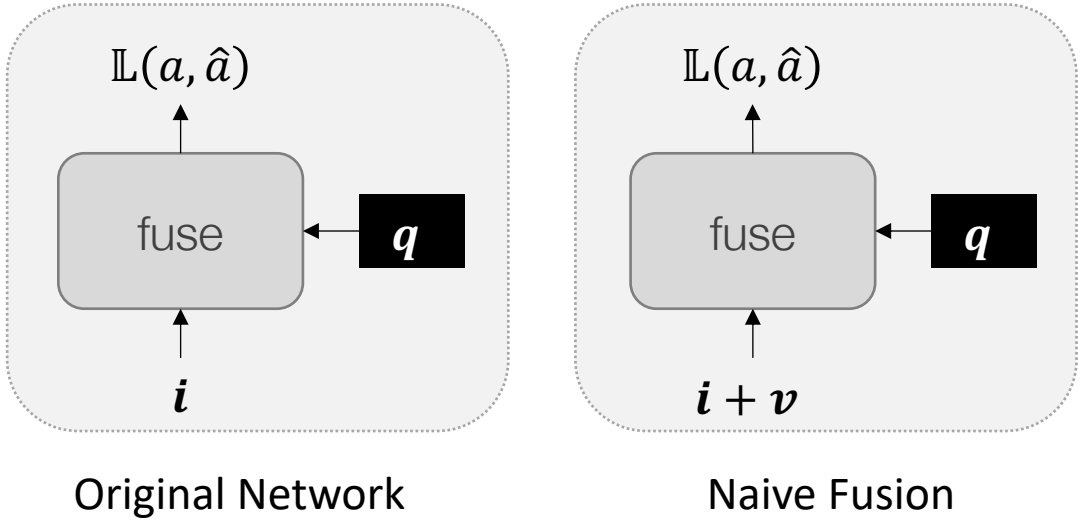
MCAN
(Winner VQA
Challenge 2019)

MoVie as a Counting Module in VQA



Original Network

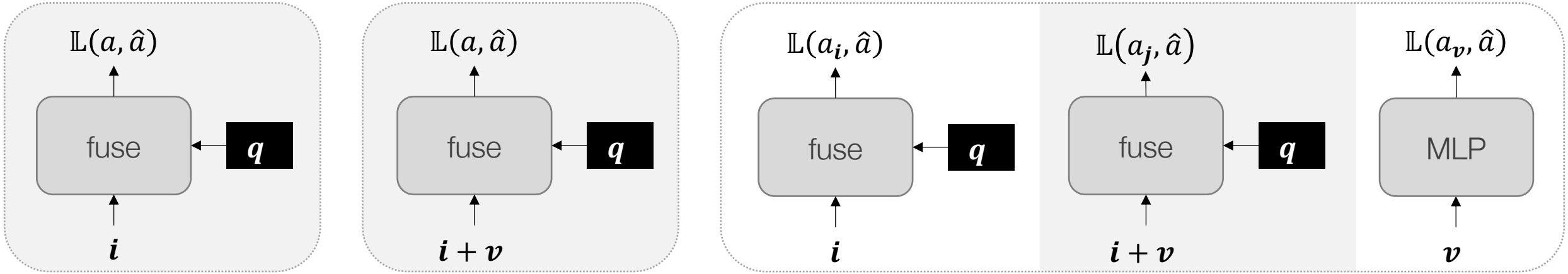
MoVie as a Counting Module in VQA



Original Network

Naive Fusion

MoVie as a Counting Module in VQA



Original Network

Naive Fusion

Three-branch fusion

MoVie as a Counting Module in VQA

Accuracy on VQA 2.0 val set

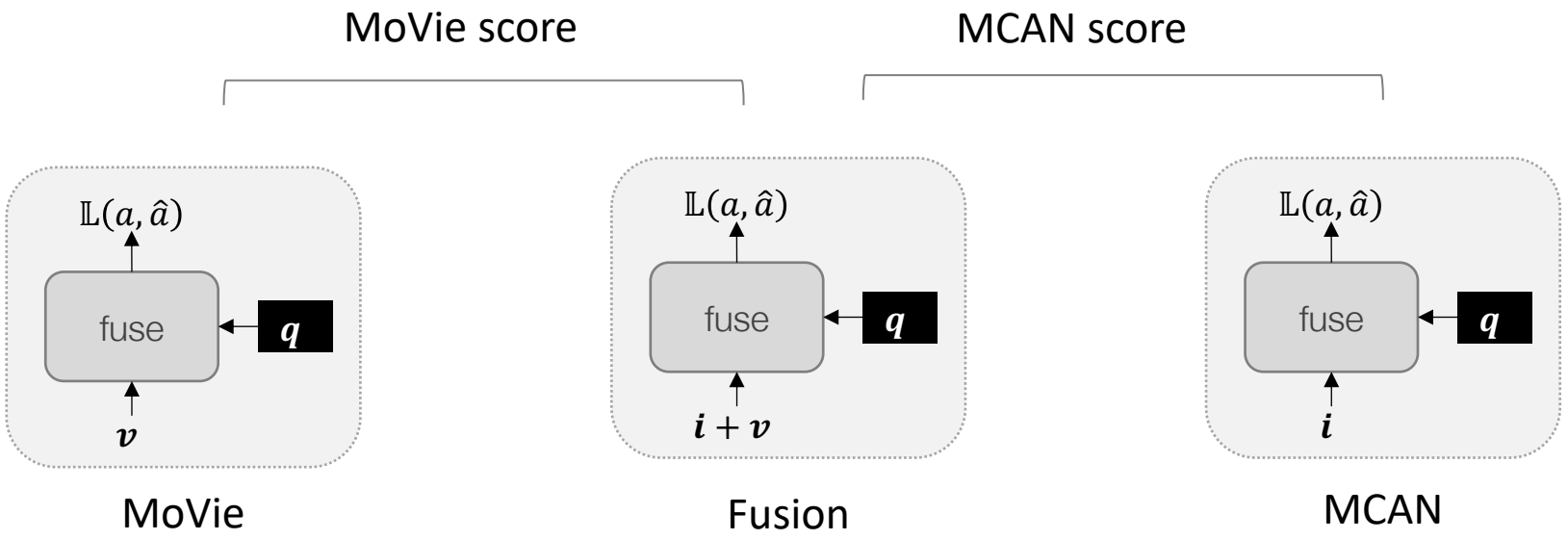
Method	Yes/No ↑	Number ↑	Other ↑	Overall ↑
MoVie	82.48	49.26	54.77	64.46
MCAN-S [Yu et al]	83.59	46.71	57.34	65.81
MCAN-S + MoVie (naive)	83.25	49.36	57.18	65.95
MCAN-S + MoVie (three-branch)	84.01	50.45	57.87	66.72

MoVie as a Counting Module in VQA

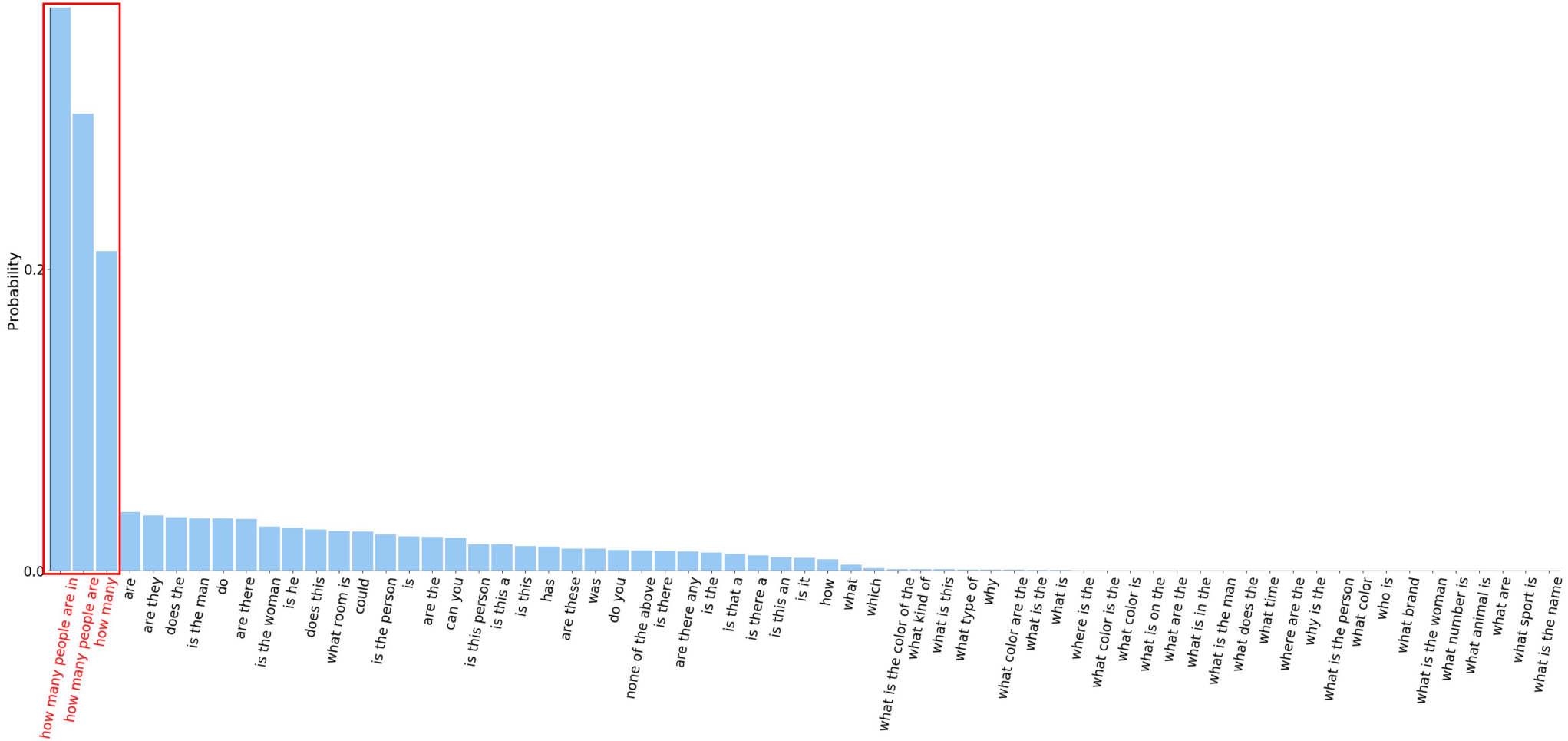
Accuracy on VQA 2.0 test-dev set

Method	Yes/No ↑	Number ↑	Other ↑	Overall ↑
MCAN-L [Huaizu et al.]	88.46	55.68	62.85	72.59
MCAN-L (Pythia)	88.13	55.40	62.85	72.42
MCAN-L + MoVie	88.39	57.05	63.28	72.91
Pythia	84.13	45.98	58.76	67.76
Pythia + MoVie	85.15	53.25	59.31	69.26

Contribution of MoVie to MCAN



Contribution of MoVie to MCAN



MoVie beyond Visual Counting

Accuracy on GQA (*: using scene-graph annotation)

Method	Overall ↑	Binary ↑	Open ↑
CNN + LSTM	46.6	63.3	31.8
BottomUp [Anderson et al.]	49.7	66.6	34.8
MAC [Hudson et al.]	54.1	71.2	38.9
NSM* [Hudson et al.]	63.2	78.9	49.3
MoVie	57.1	73.5	42.7
Humans	89.3	91.2	87.4

Thank you for your attention