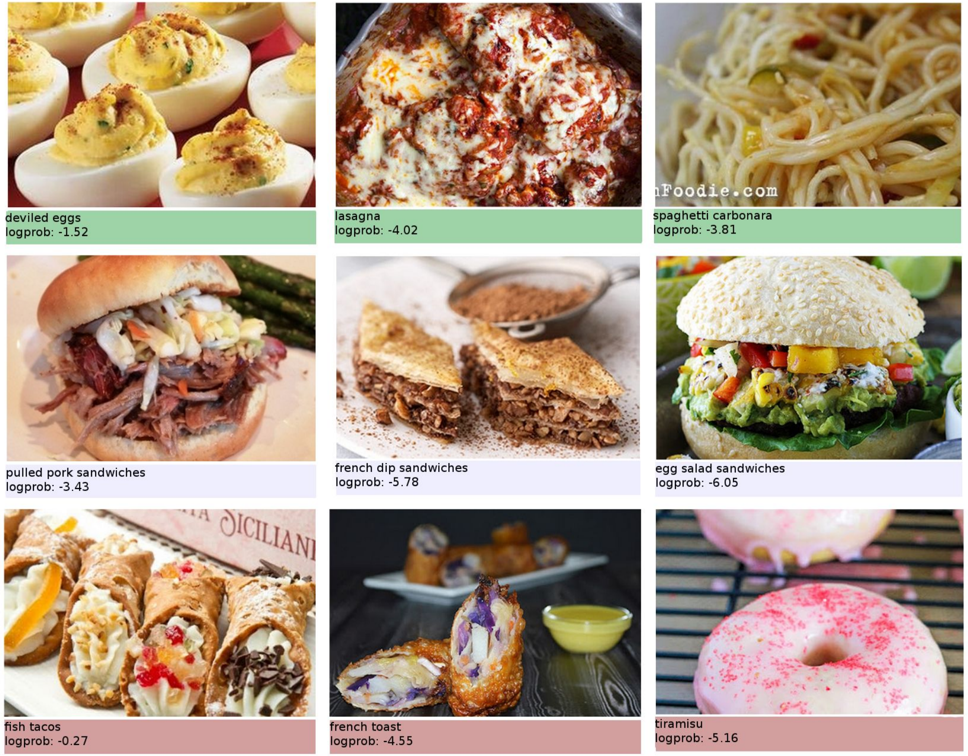
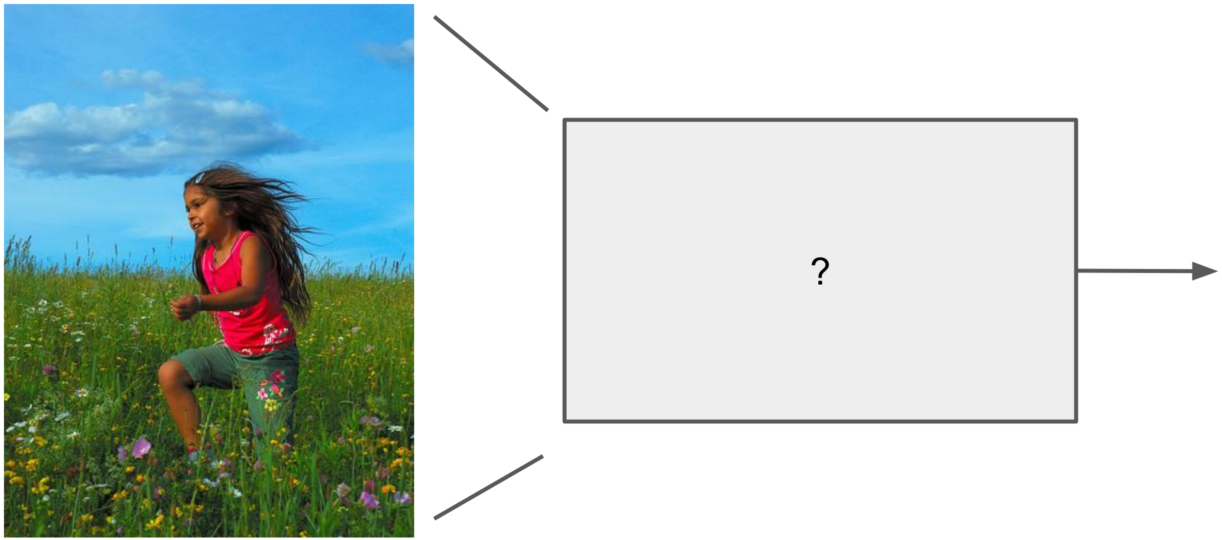


Image Representations and New Domains in Neural Image Captioning

Jack Hessel, Nic Saava, Mike Wilber

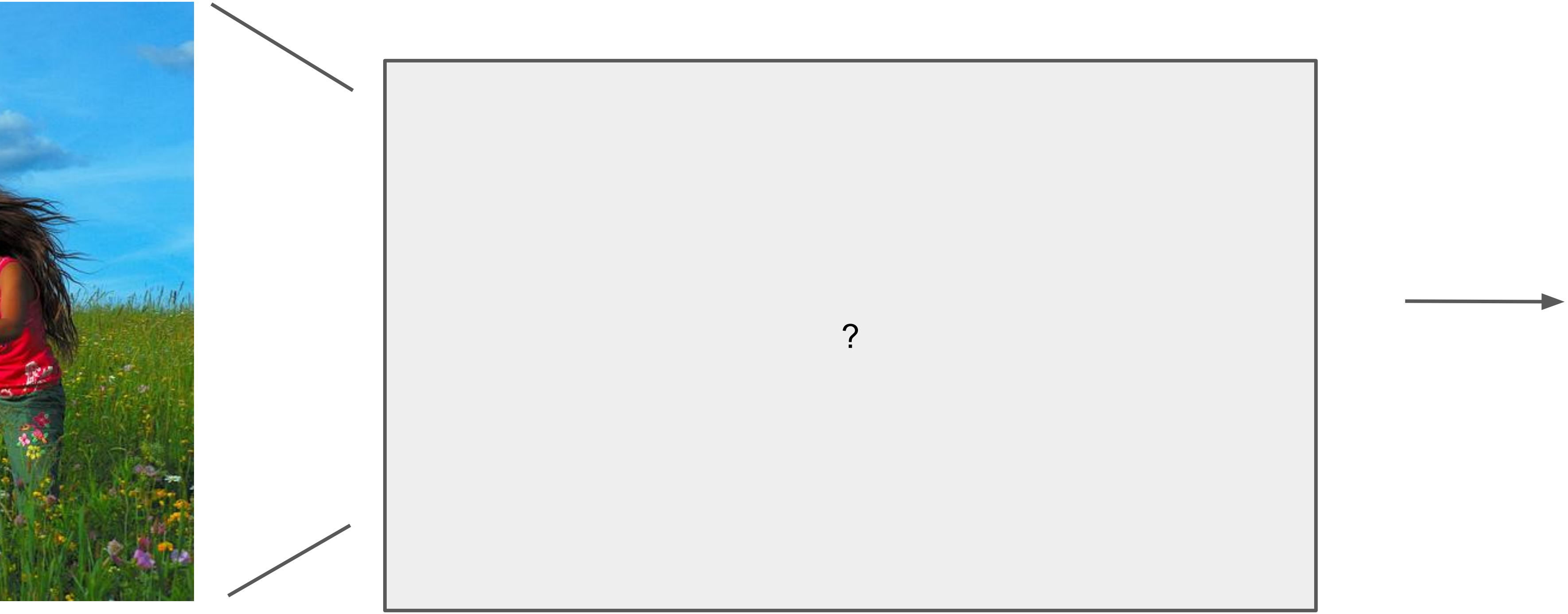
The caption generation problem...



?

A young girl runs through a grassy field.

1



?

A you
throug

1

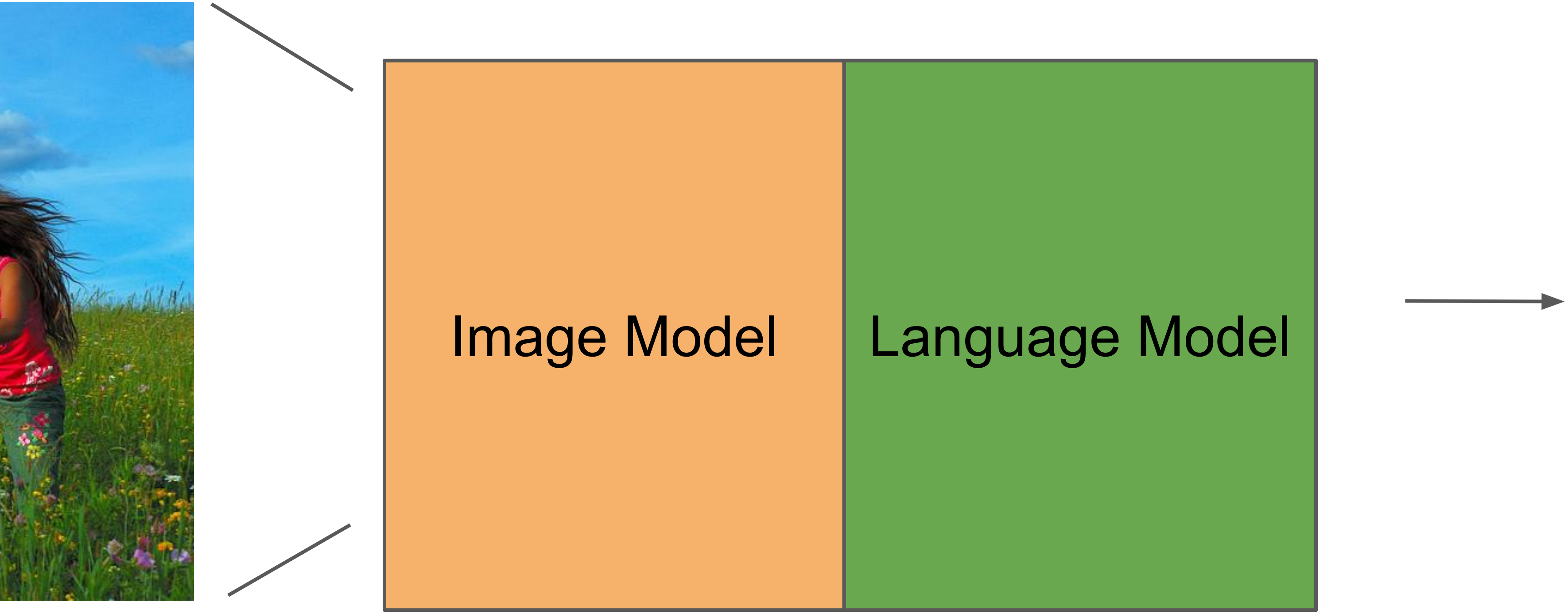
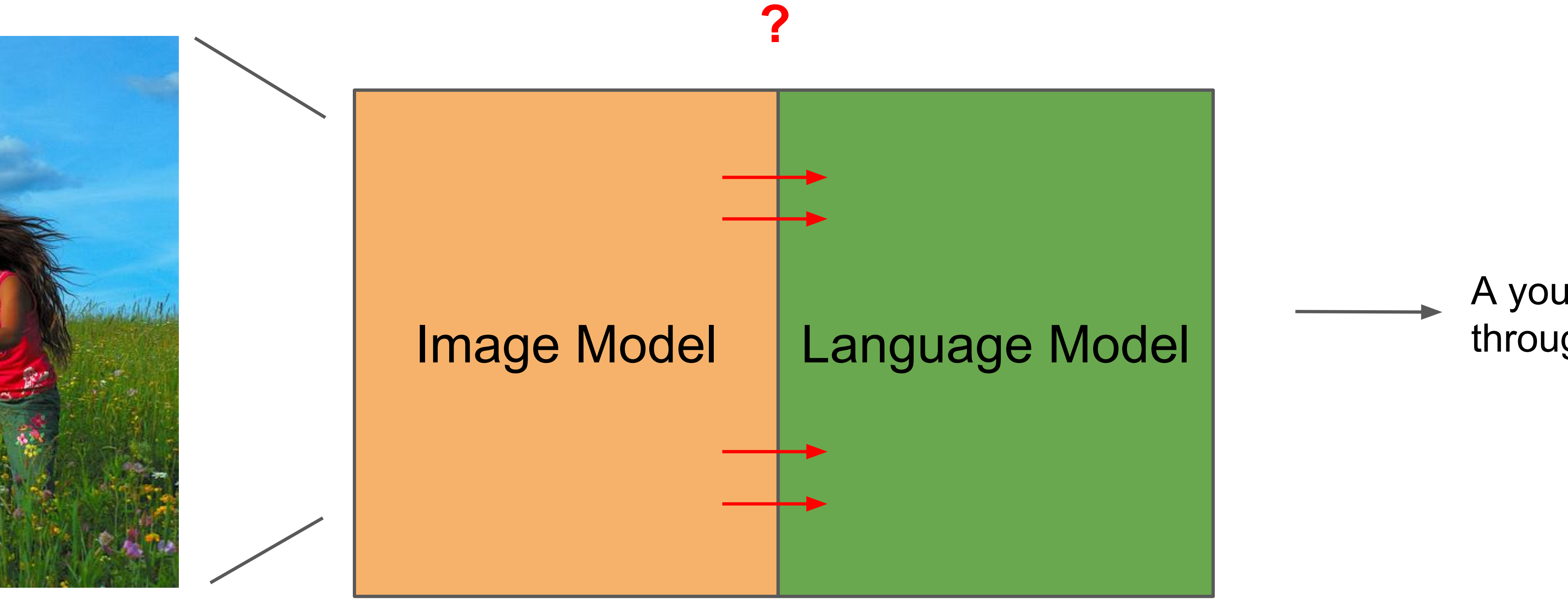


Image Model

Language Model

A you
throug

1



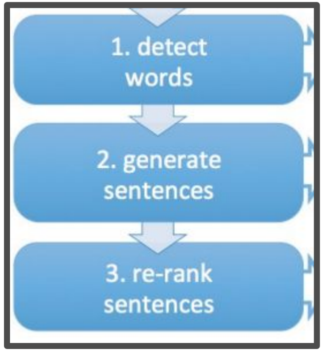
?

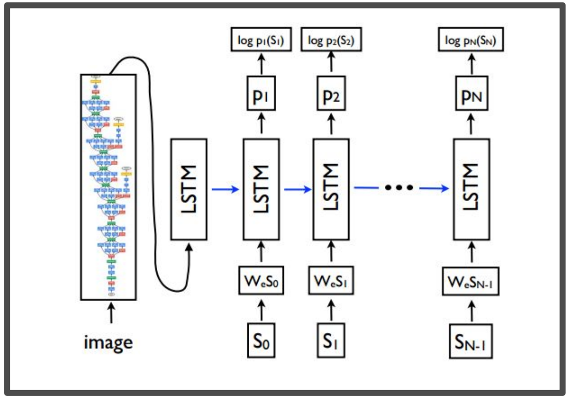
Image Model

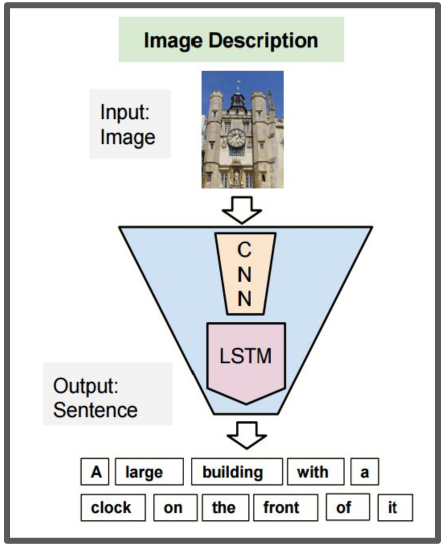
Language Model

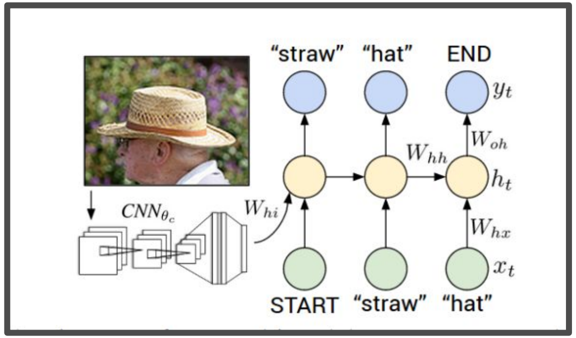
A you
throug

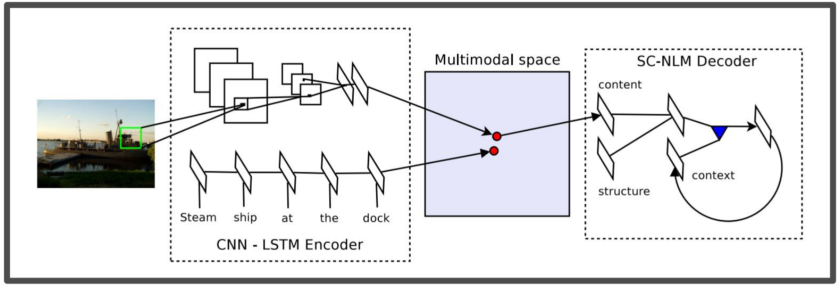
The world of neural caption generation models...

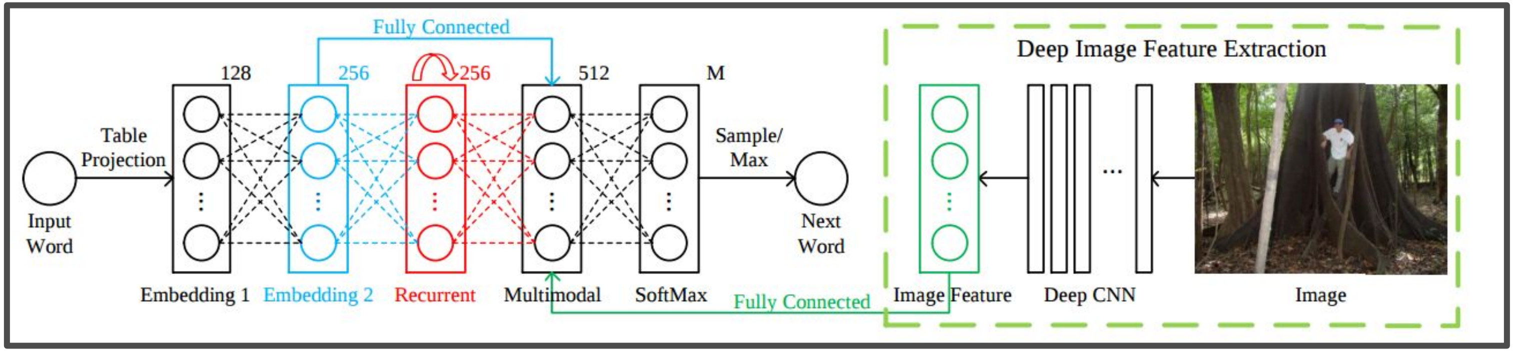




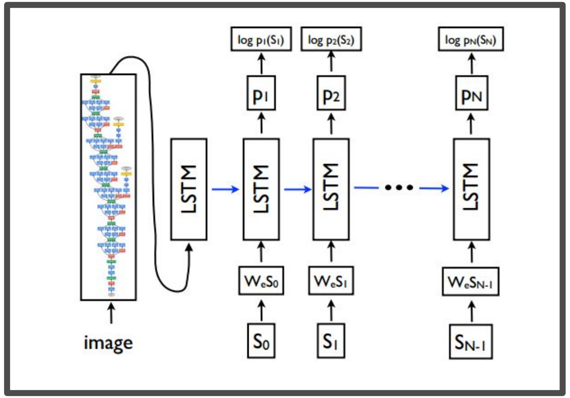






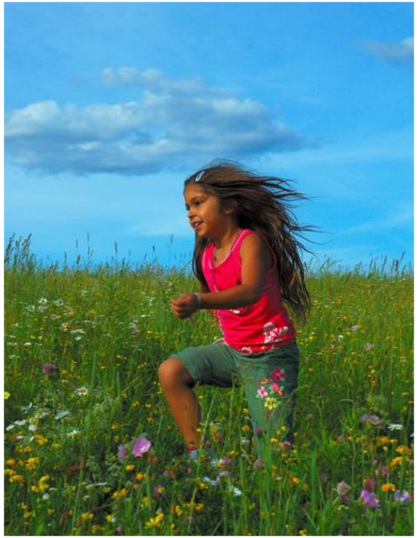


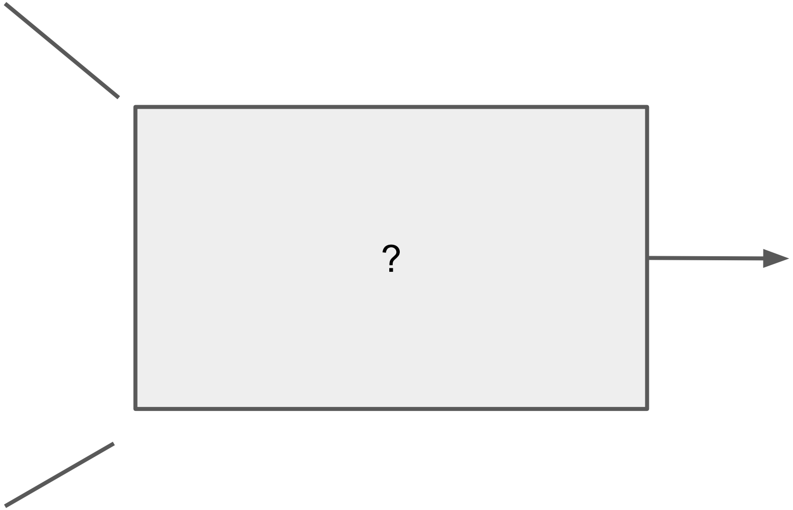
[Karpathy and Li 2014]
[Mao et al. 2014]
[Vinyals et al. 2014]
[Kiros et al. 2014]
[Donahue et al. 2014]
[Fang et al. 2014]



[Vinyals et al. 2014]

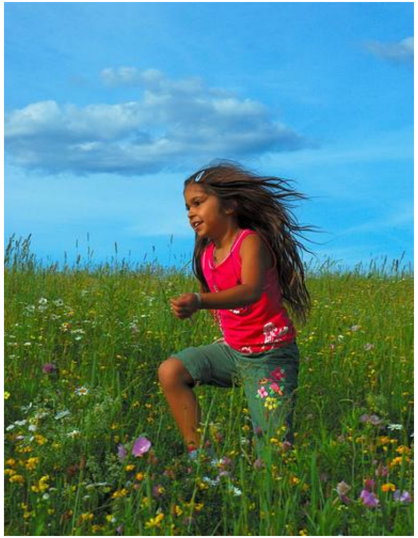
Neural Image Caption (Vinyal’s et. al. 2014)





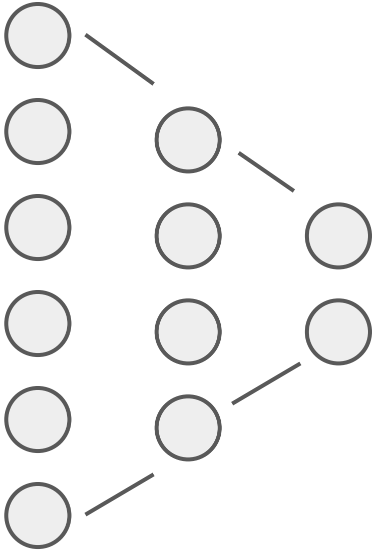
A young girl runs
through a grassy field.

Neural Image Caption (Vinyal’s et. al. 2014)



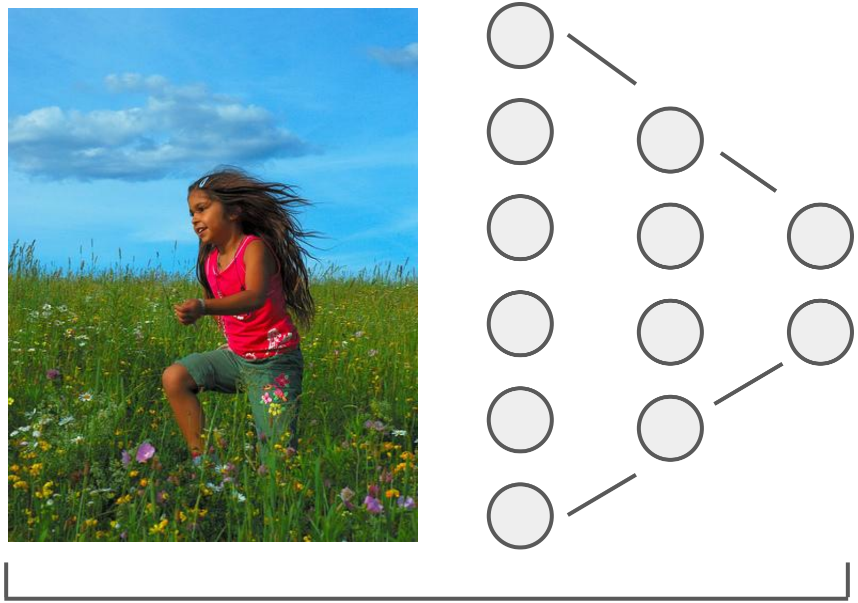
A young girl runs
through a grassy field.

Neural Image Caption (Vinyal’s et. al. 2014)



A young girl runs
through a grassy field.

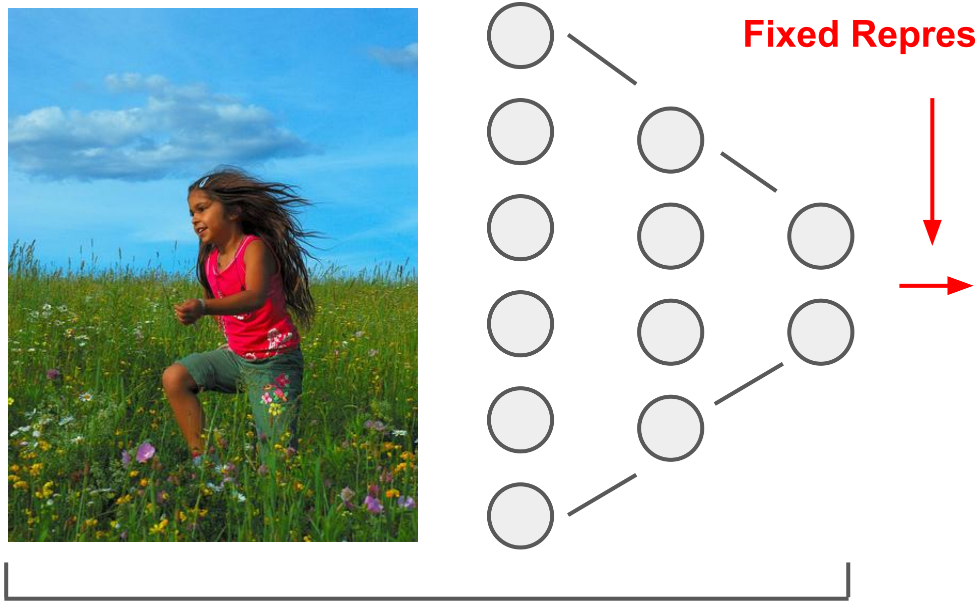
Neural Image Caption (Vinyal’s et. al. 2014)



Convolutional Neural Network

A young girl runs
through a grassy field.

Neural Image Caption (Vinyal’s et. al. 2014)



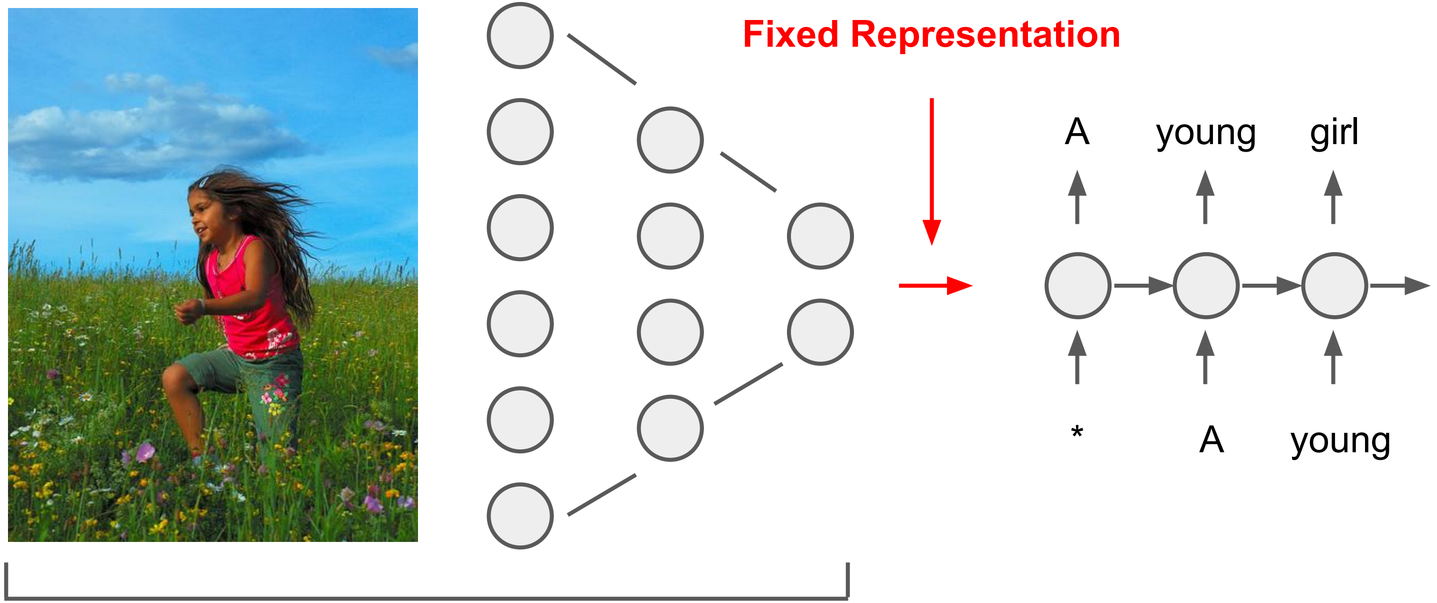
Fixed Representation

Fixed
+
Pretrained

Convolutional Neural Network

A young girl runs
through a grassy field.

Neural Image Caption (Vinyal’s et. al. 2014)



Fixed Representation

A young girl

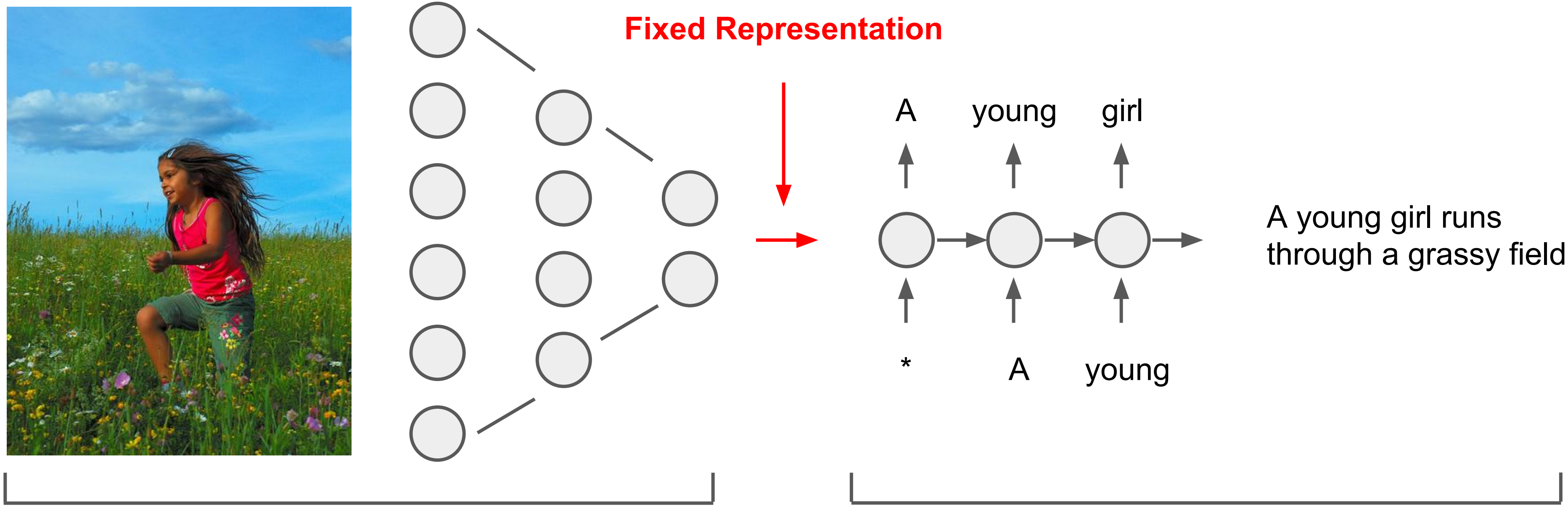
* A young

A young girl runs through a grassy field.

Fixed + Pretrained

Convolutional Neural Network

Neural Image Caption (Vinyal’s et. al. 2014)



Fixed Representation

A young girl runs
through a grassy field.

Fixed
+
Pretrained

Convolutional Neural Network

Long Short Term Memory
Recurrent Neural Network

[Hochreiter and Schmidhuber 1997]

Sutskever et al. 2011

The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show’s agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE...

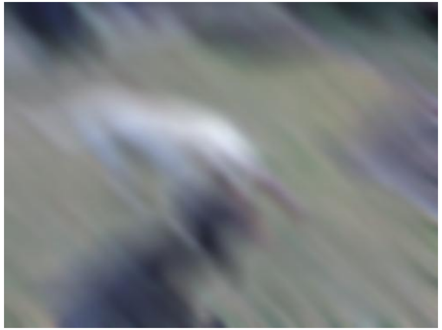
Sutskever et al. 2011

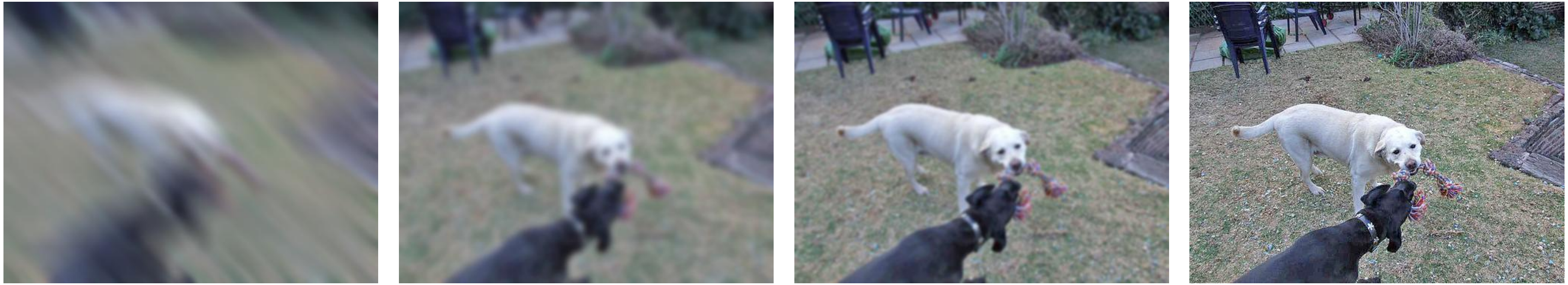
The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show's agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE...

Sutskever et al. 2011

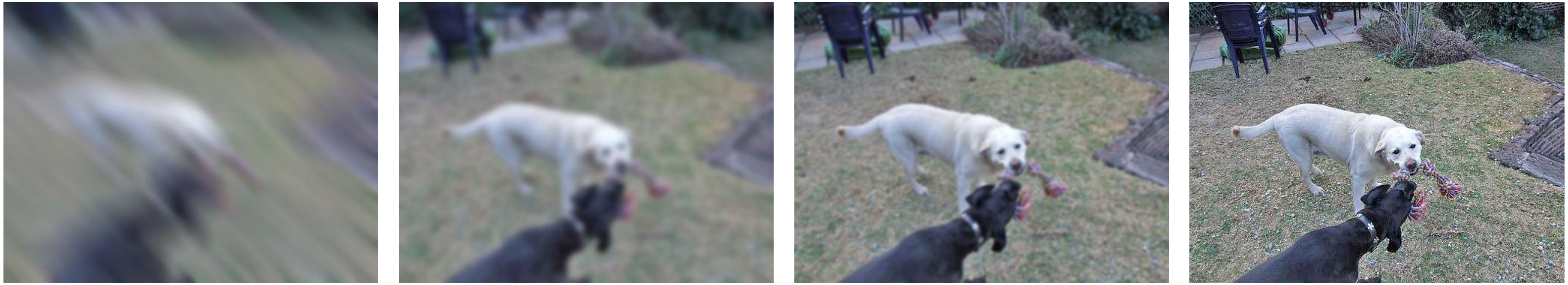
The meaning of life is the tradition of the ancient human reproduction: it is less favorable to the good boy for when to remove her bigger. In the show’s agreement unanimously resurfaced. The wild pasteured with consistent street forests were incorporated by the 15th century BE...

6

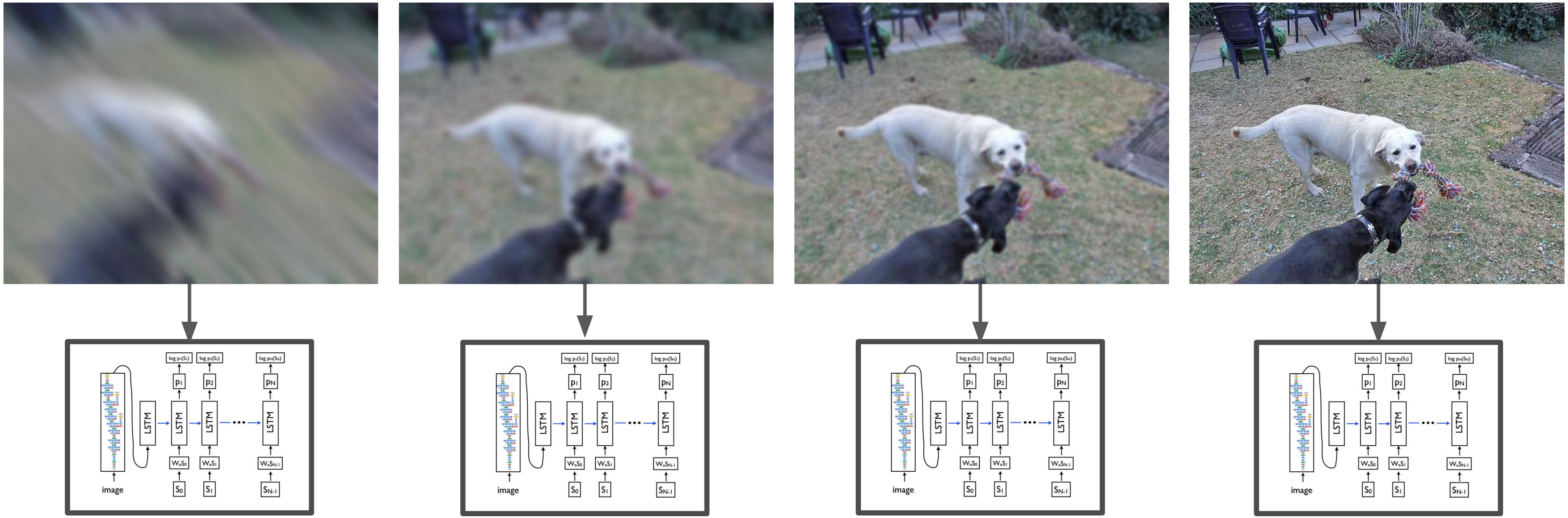




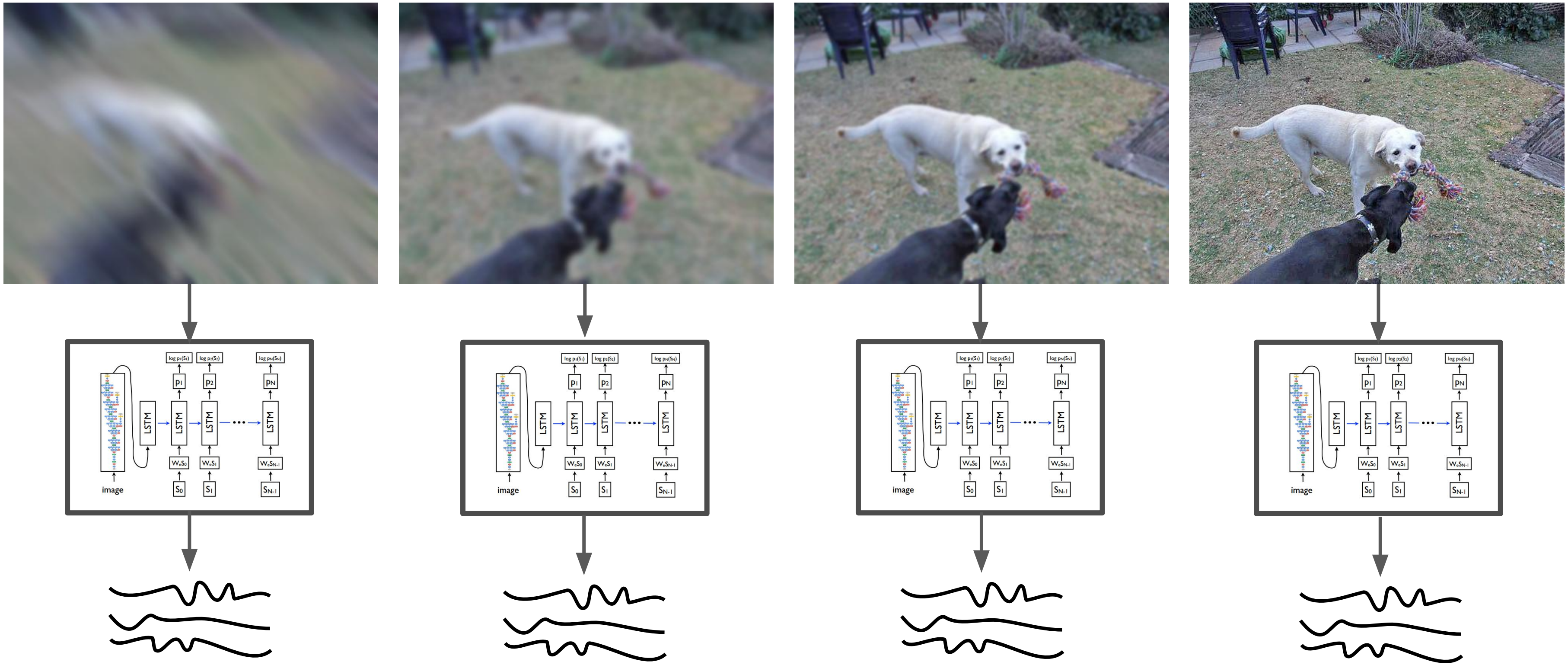
Higher Quality Representations



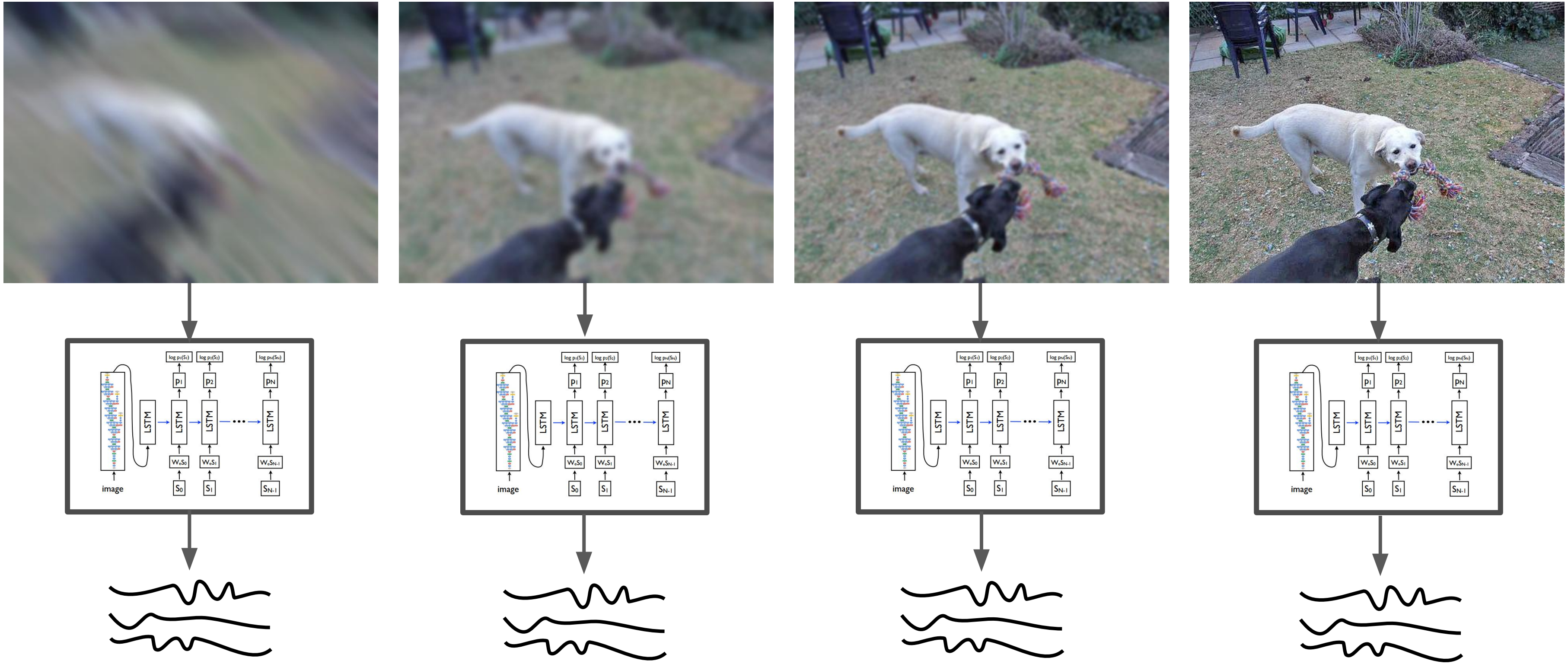
Higher Quality Representations



Higher Quality Representations

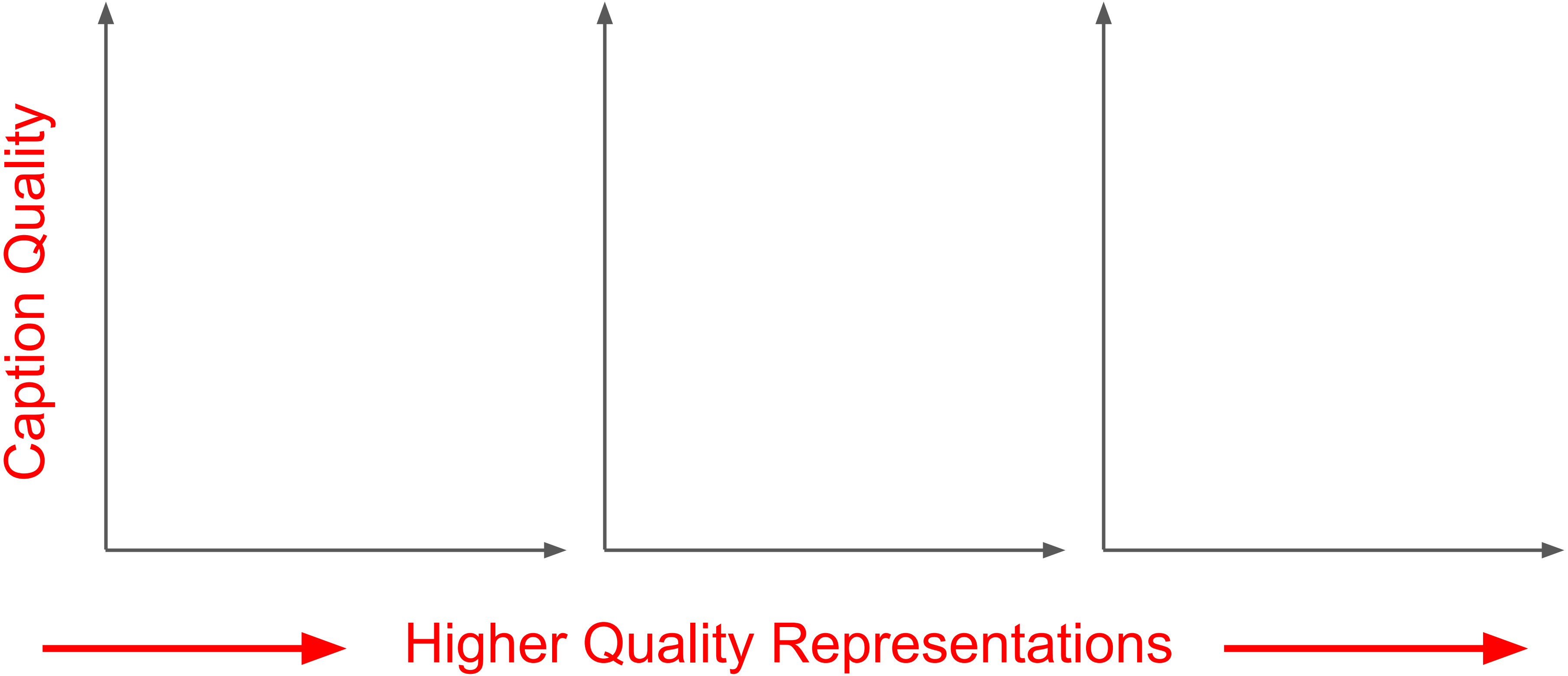


Higher Quality Representations

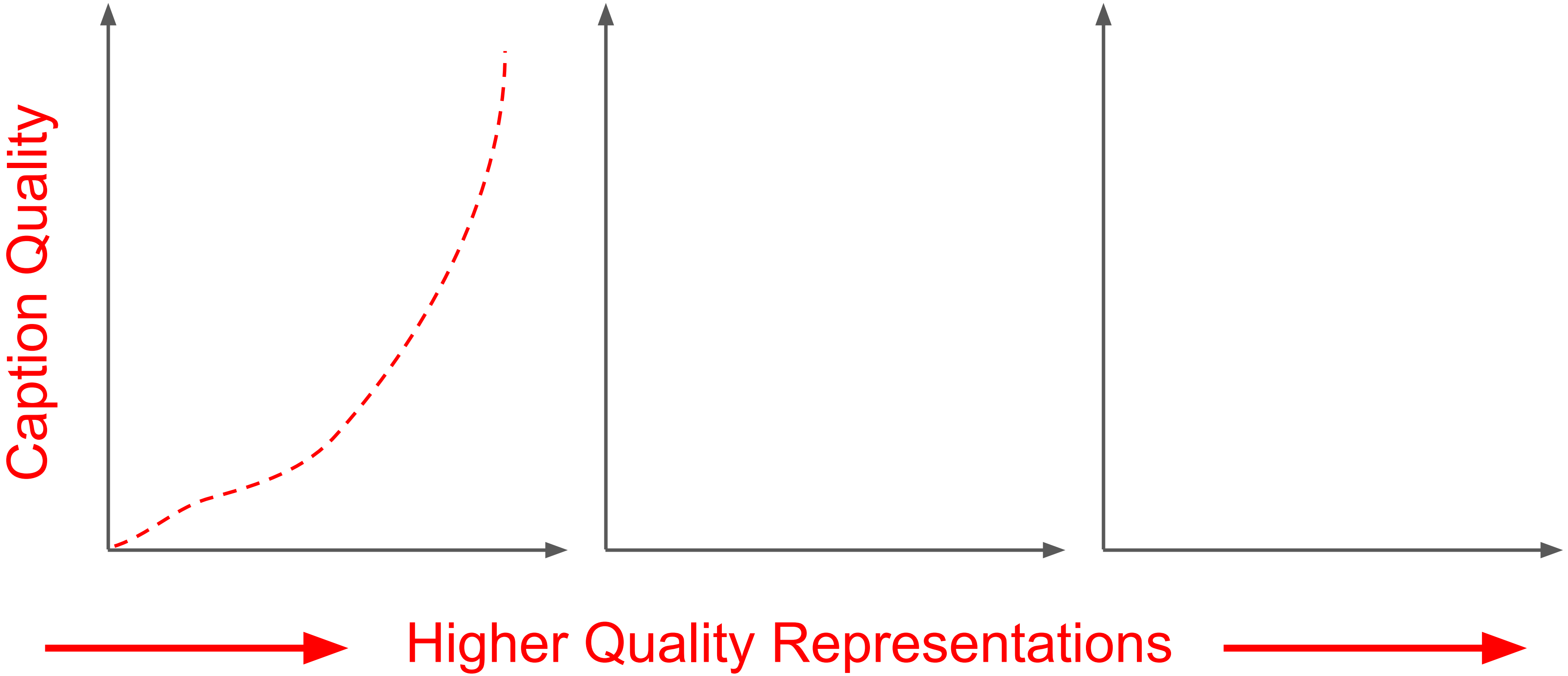


?? Higher Quality Captions ??

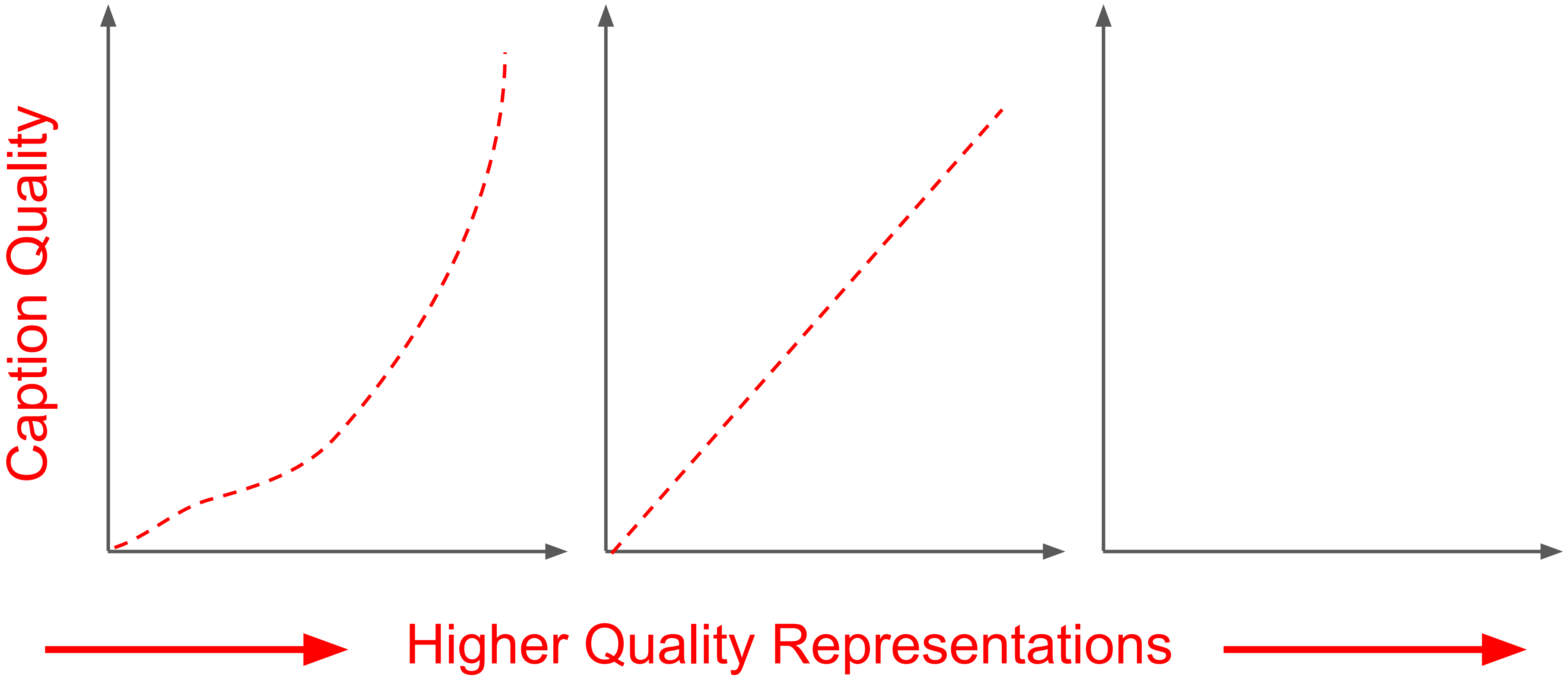
Plausible outcomes....



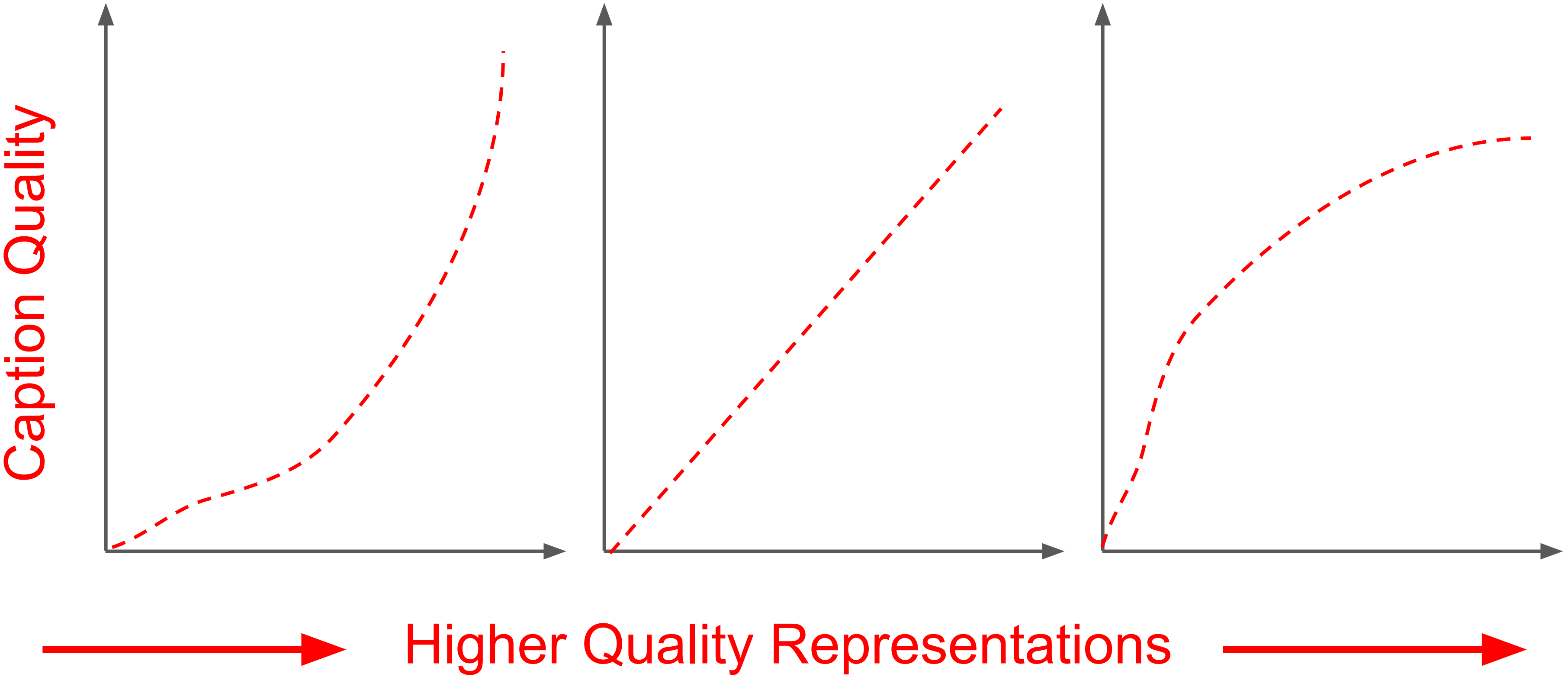
Plausible outcomes....



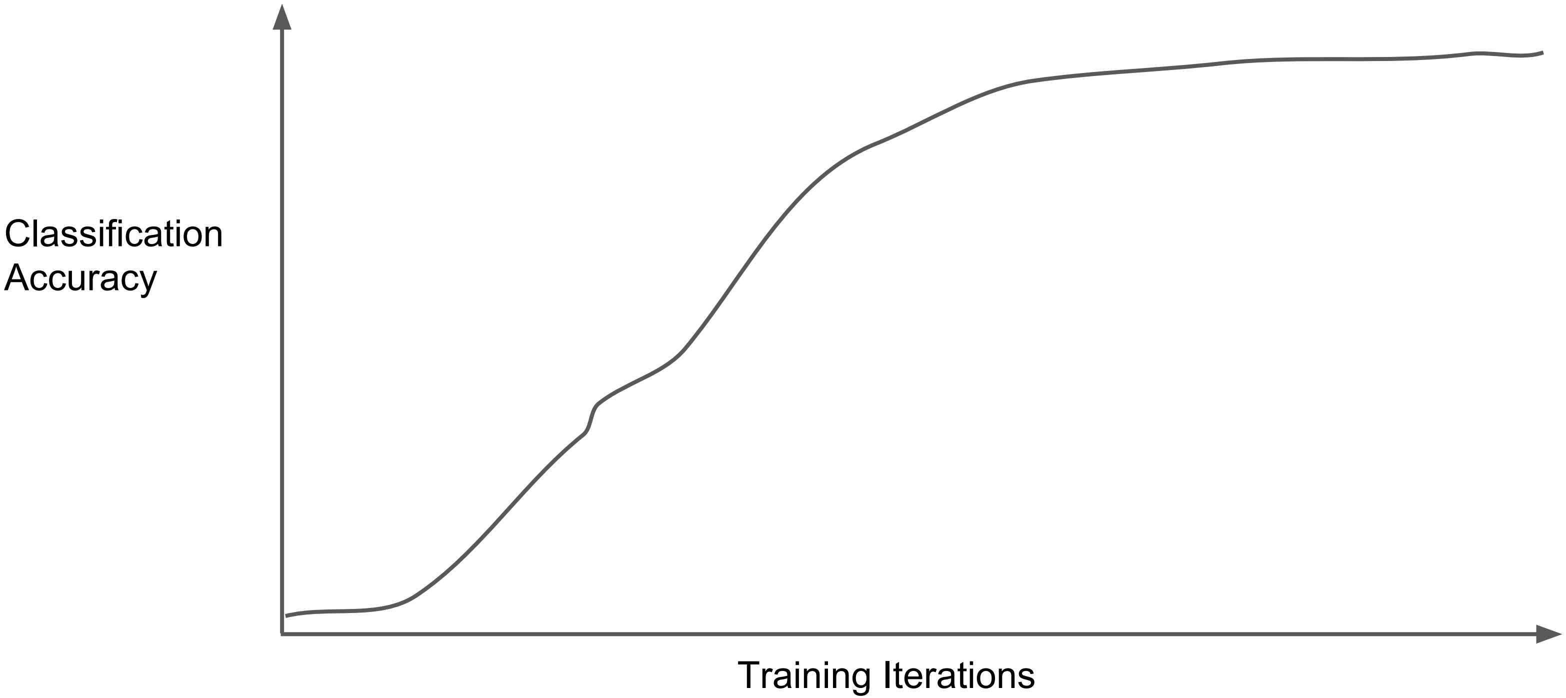
Plausible outcomes....

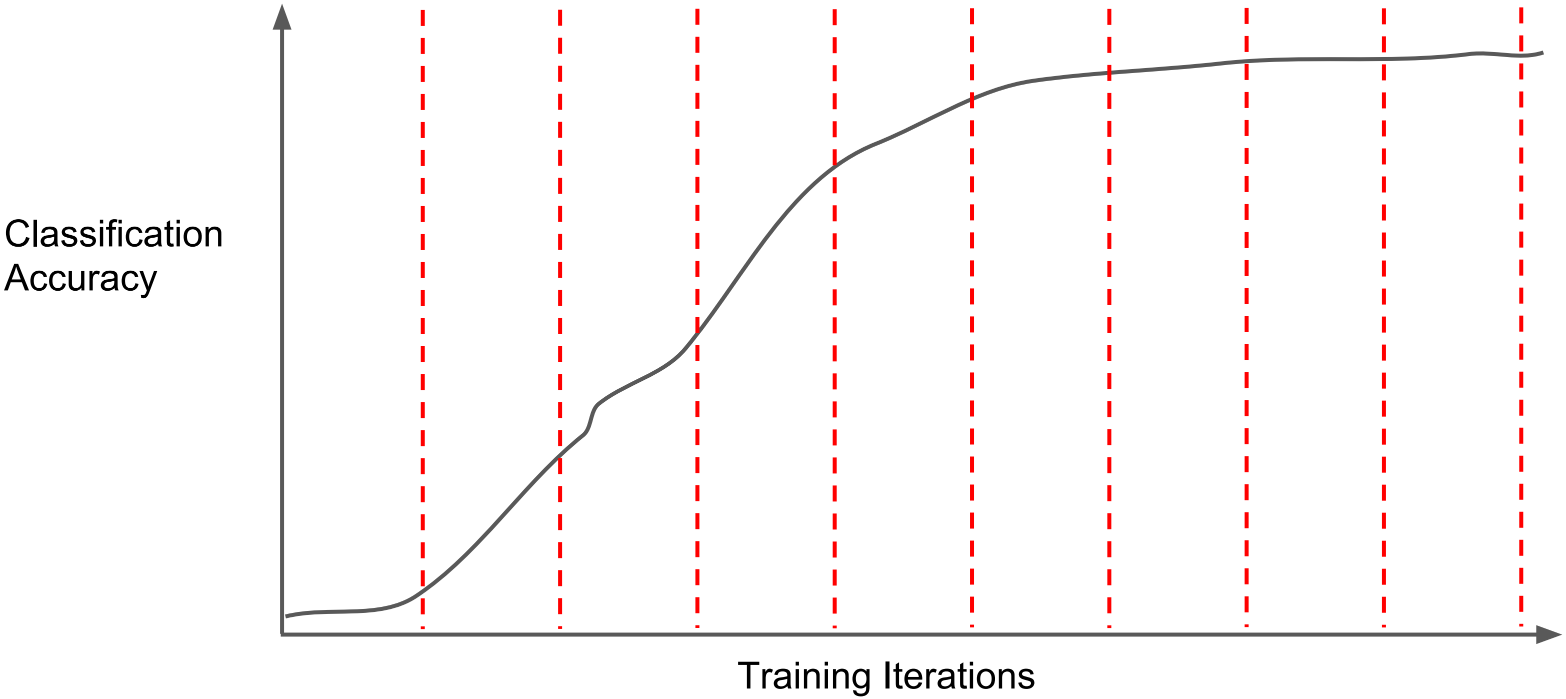


Plausible outcomes....

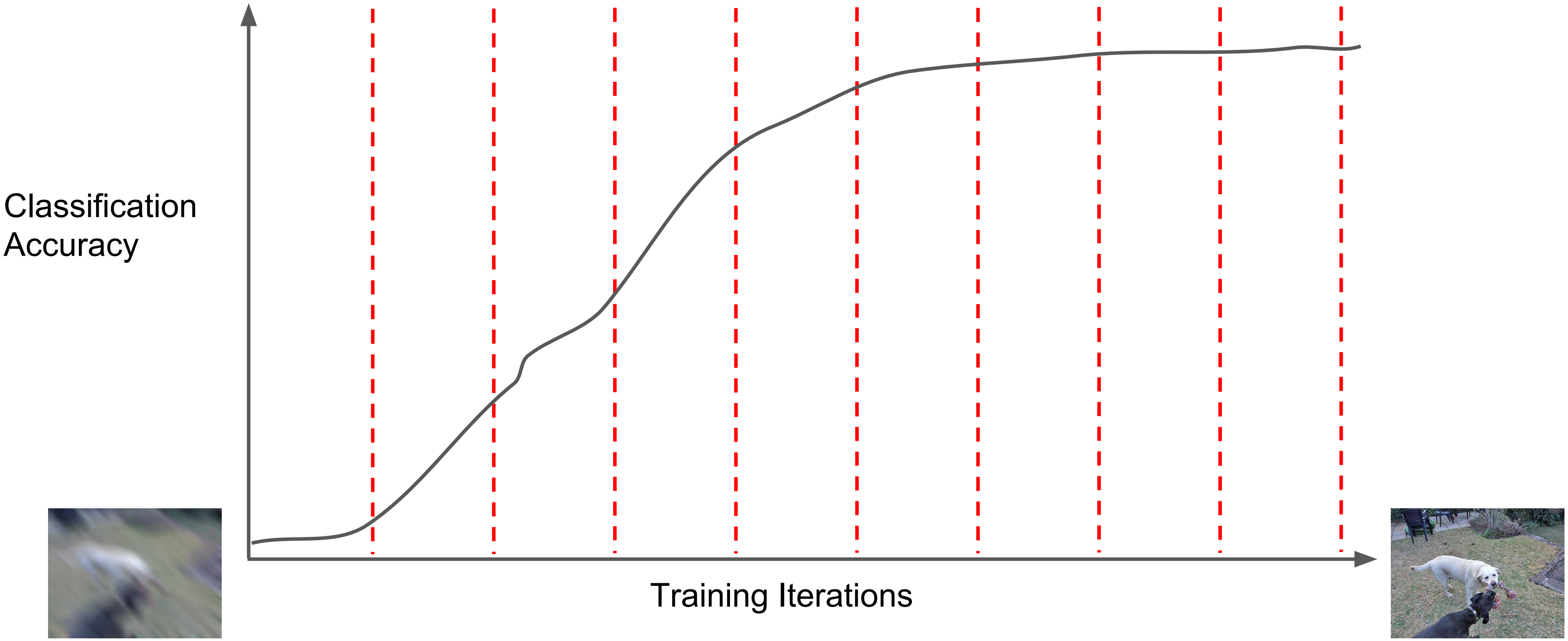


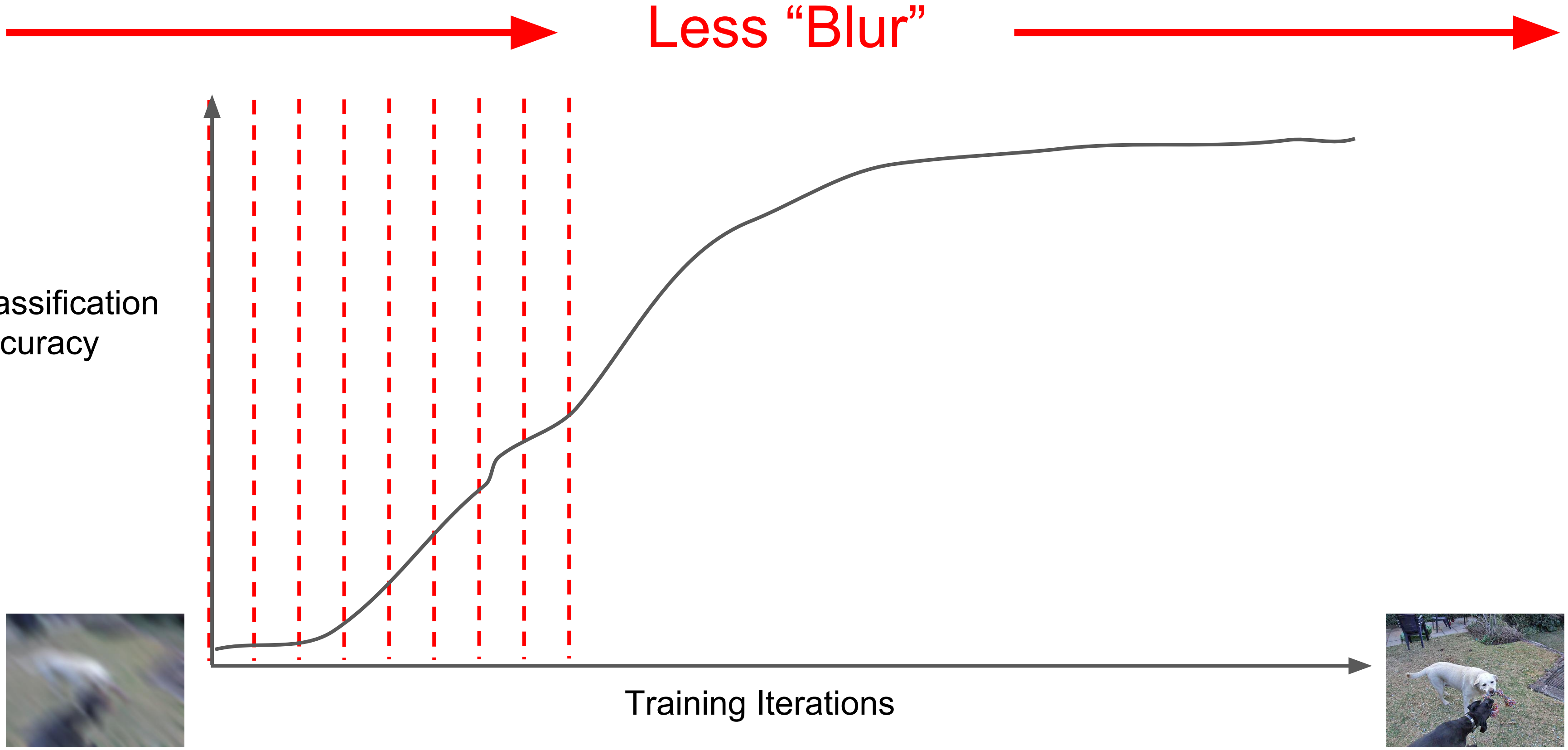
Classification
Accuracy
Training Iterations


Classification Accuracy
Training Iterations

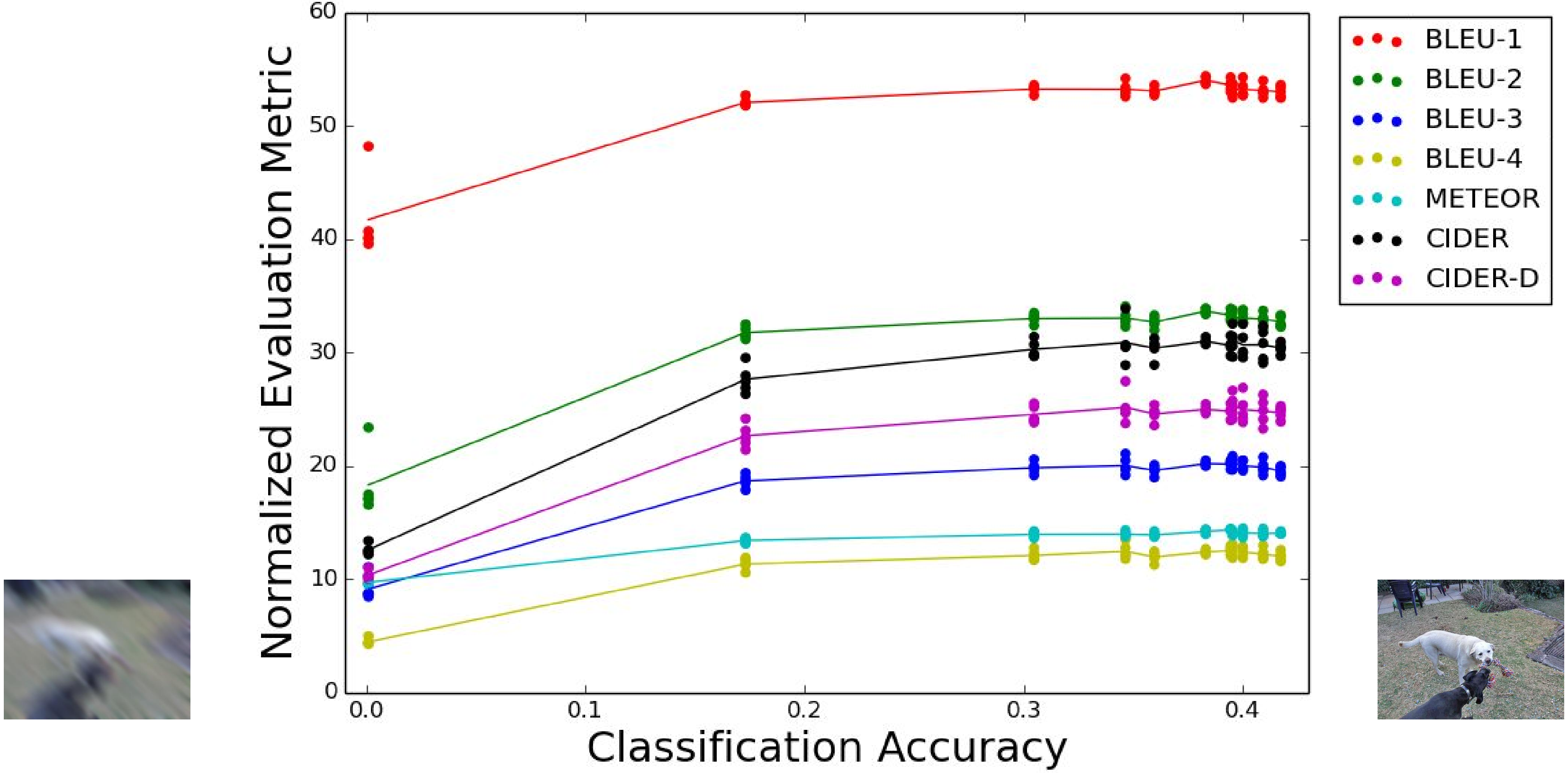

Classification Accuracy
Training Iterations

Less “Blur”

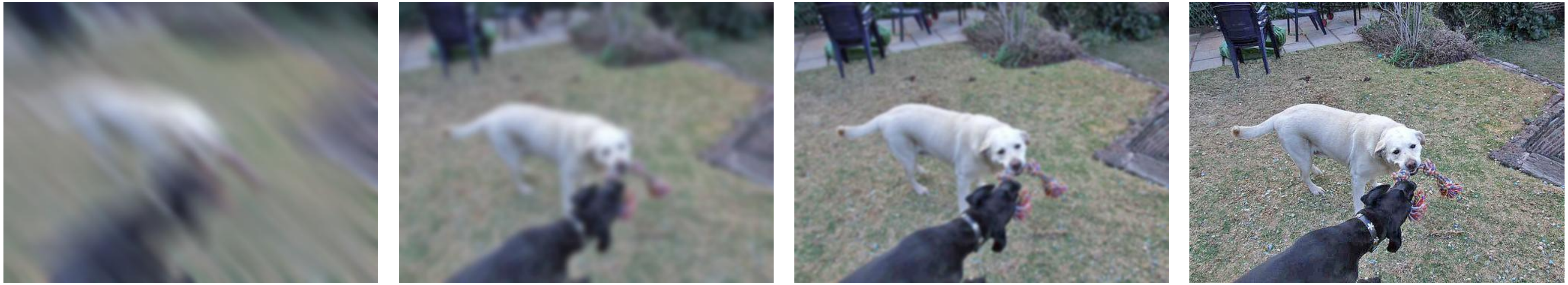



Less “Blur”
Classification Accuracy
Training Iterations

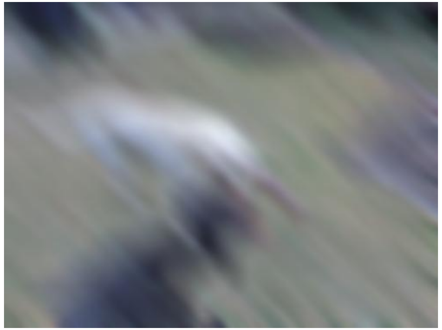
Directly learned: Flickr8k [Hodosh et al. 2013]

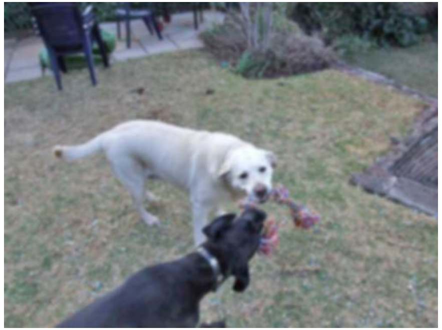


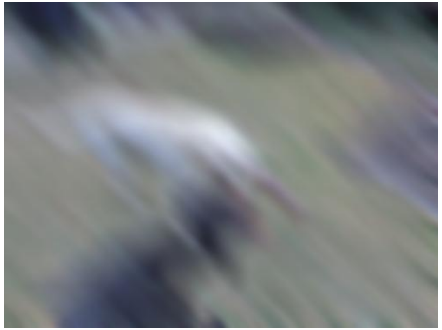
10

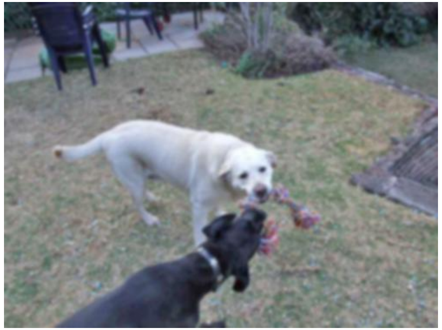


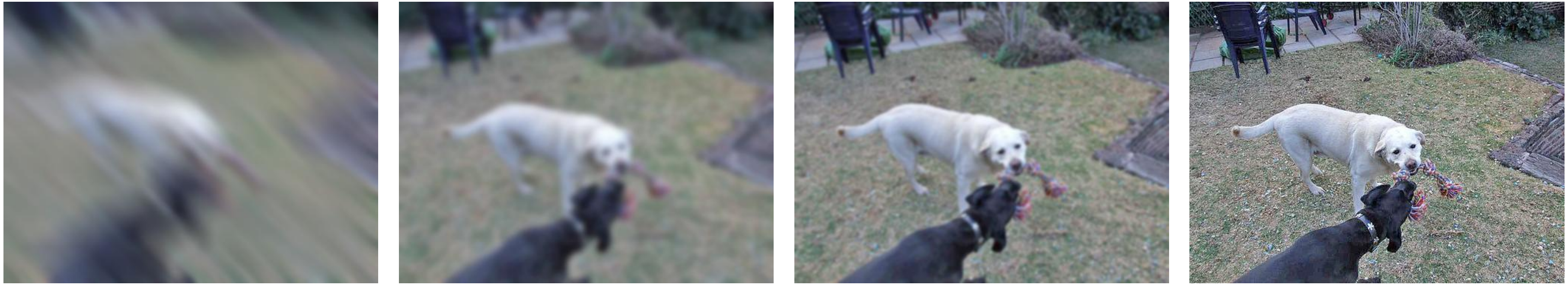
10

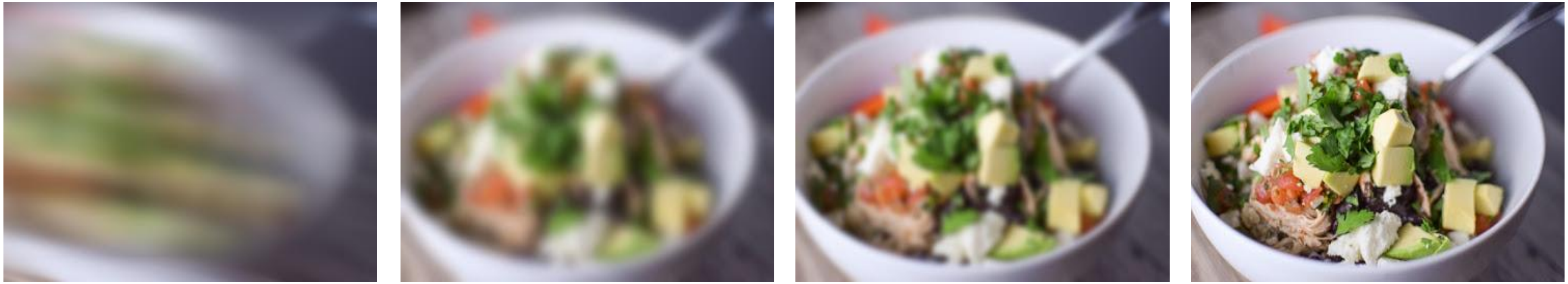










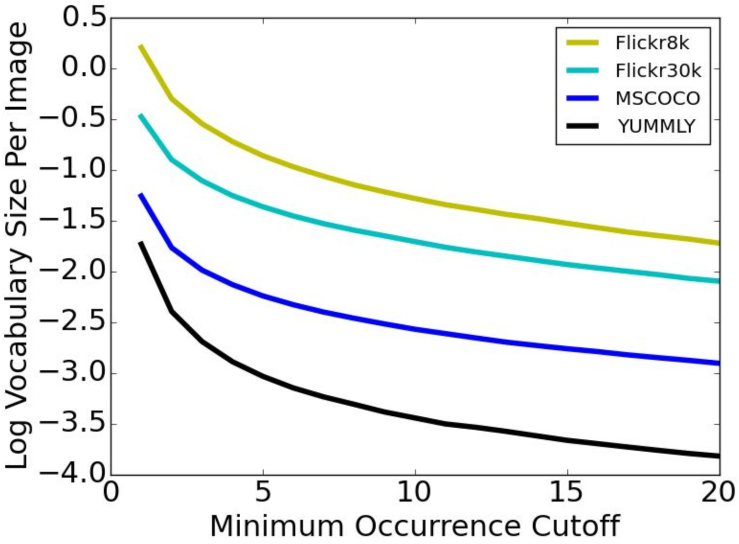


→ “Malaysian Spicy Noodles”

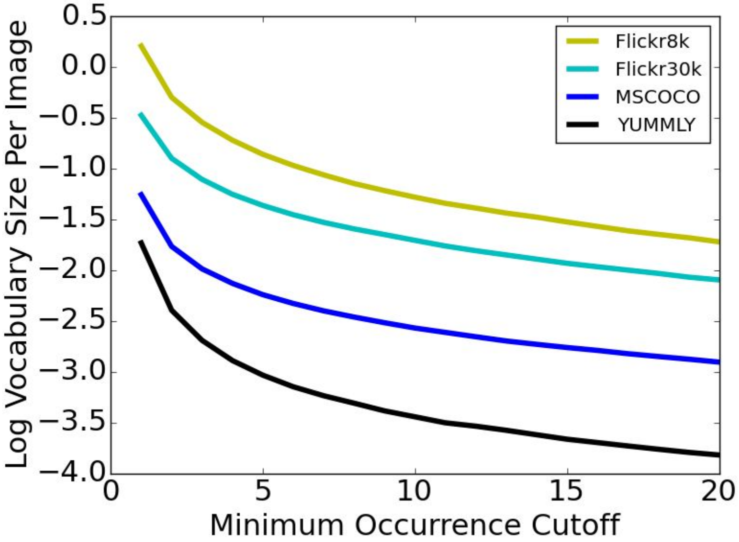
66K Image/recipe pairs, courtesy Yummly.com

Is this really a new *language* task?

Is this really a new *language* task?

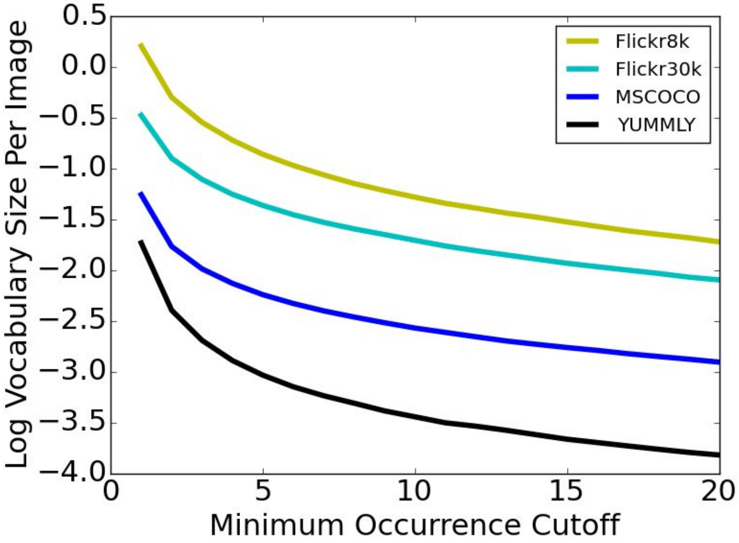


Is this really a new *language* task?



Less variety overall

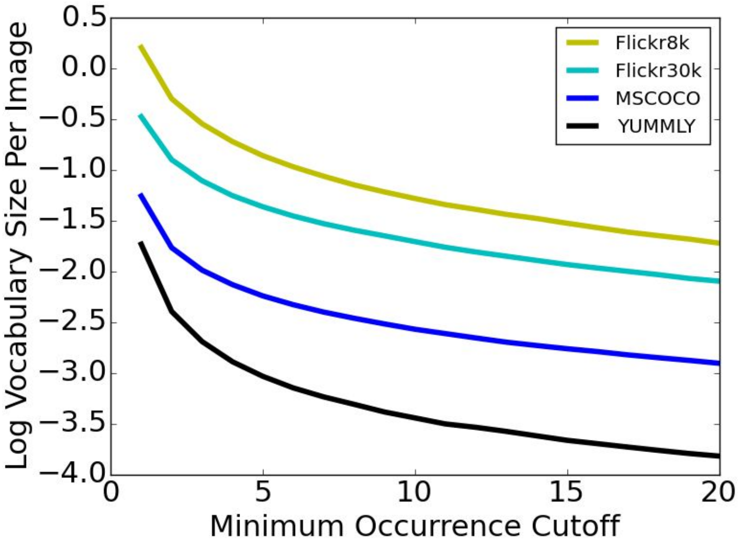
Is this really a new *language* task?



malaysian spicy noodles
chicken curry with lime
cheesy pizza bagels with mushrooms
...

Less variety overall

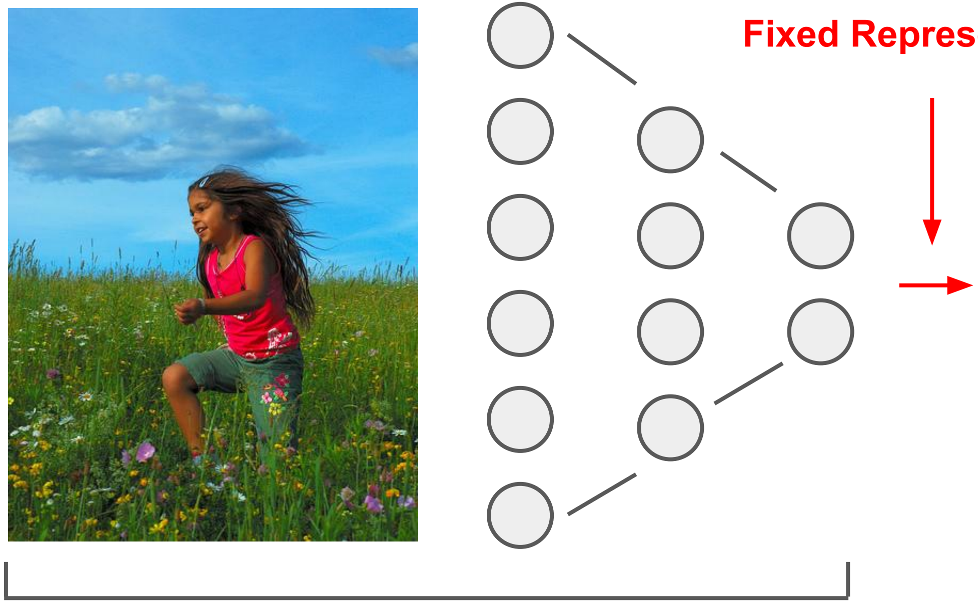
Is this really a new *language* task?



malaysian spicy noodles
chicken curry with lime
cheesy pizza bagels with mushrooms
...

Less variety overall

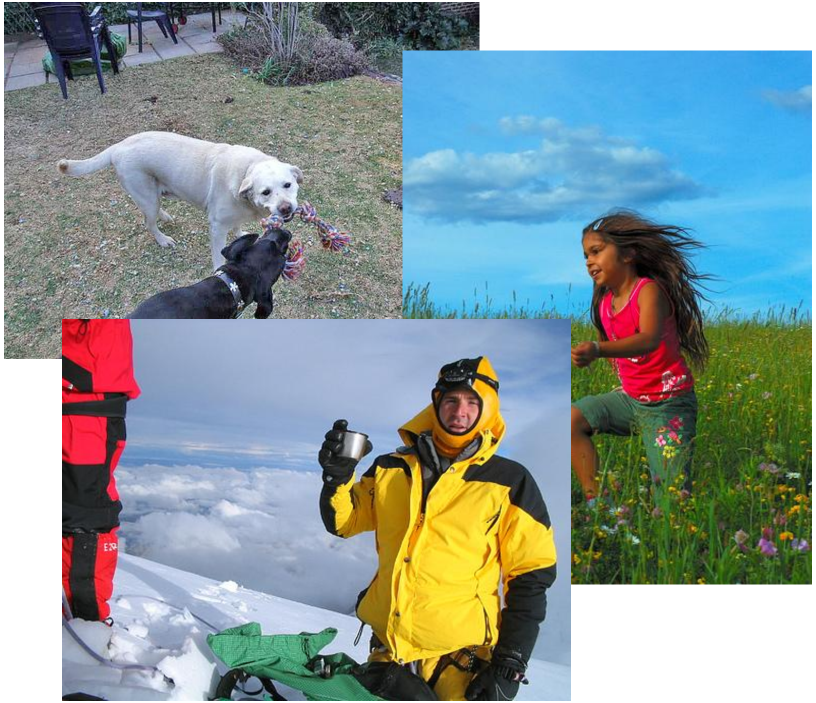
Shorter captions



Fixed Representation

Fixed
+
Pretrained

Convolutional Neural Network



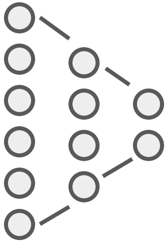
=/=

Data from [Bossard et al. 2014]

101K Labeled Food Images in 101 Classes

Data from [Bossard et al. 2014]

101K Labeled Food Images in 101 Classes

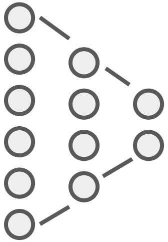


[Krizhevsky et al. 2012]

56.4% Rank-1 Accuracy

Data from [Bossard et al. 2014]

101K Labeled Food Images in 101 Classes



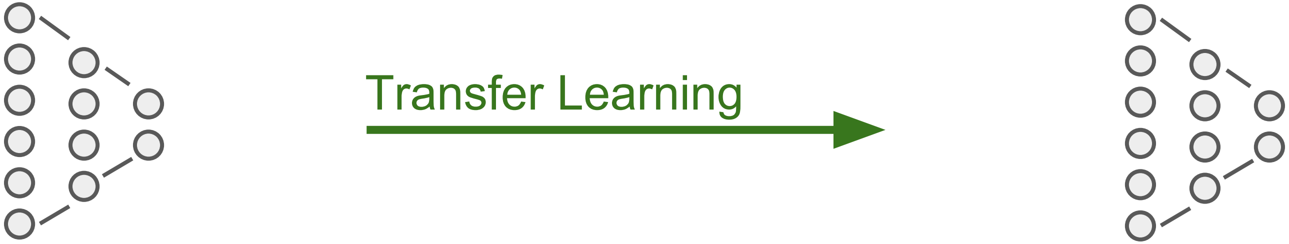
[Krizhevsky et al. 2012]

Transfer Learning

56.4% Rank-1 Accuracy

Data from [Bossard et al. 2014]

101K Labeled Food Images in 101 Classes



Transfer Learning

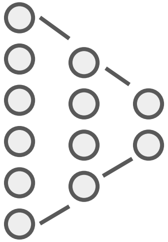
[Krizhevsky et al. 2012]

56.4% Rank-1 Accuracy

66.8% Rank-1 Accuracy

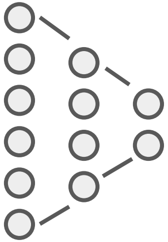
Data from [Bossard et al. 2014]

101K Labeled Food Images in 101 Classes



[Krizhevsky et al. 2012]

Transfer Learning



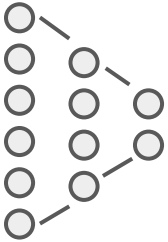
[Krizhevsky et al. 2012]

56.4% Rank-1 Accuracy

66.8% Rank-1 Accuracy

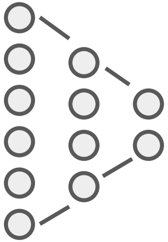
Data from [Bossard et al. 2014]

101K Labeled Food Images in 101 Classes



[Krizhevsky et al. 2012]

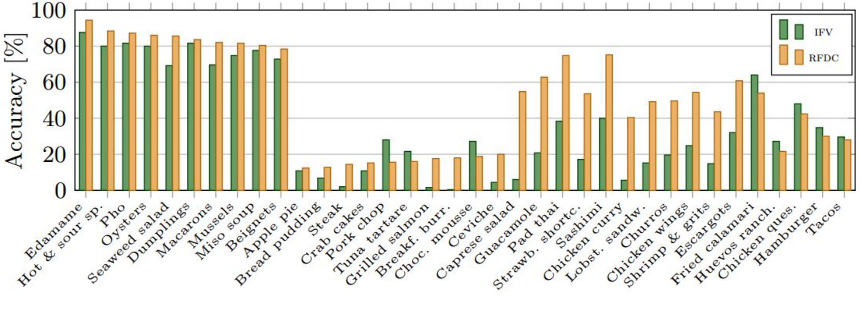
Transfer Learning



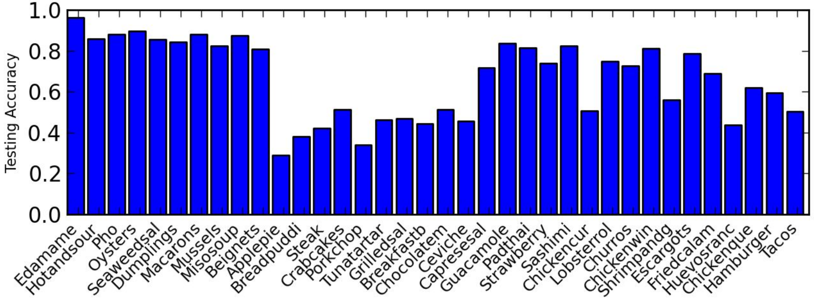
[Krizhevsky et al. 2012]

56.4% Rank-1 Accuracy

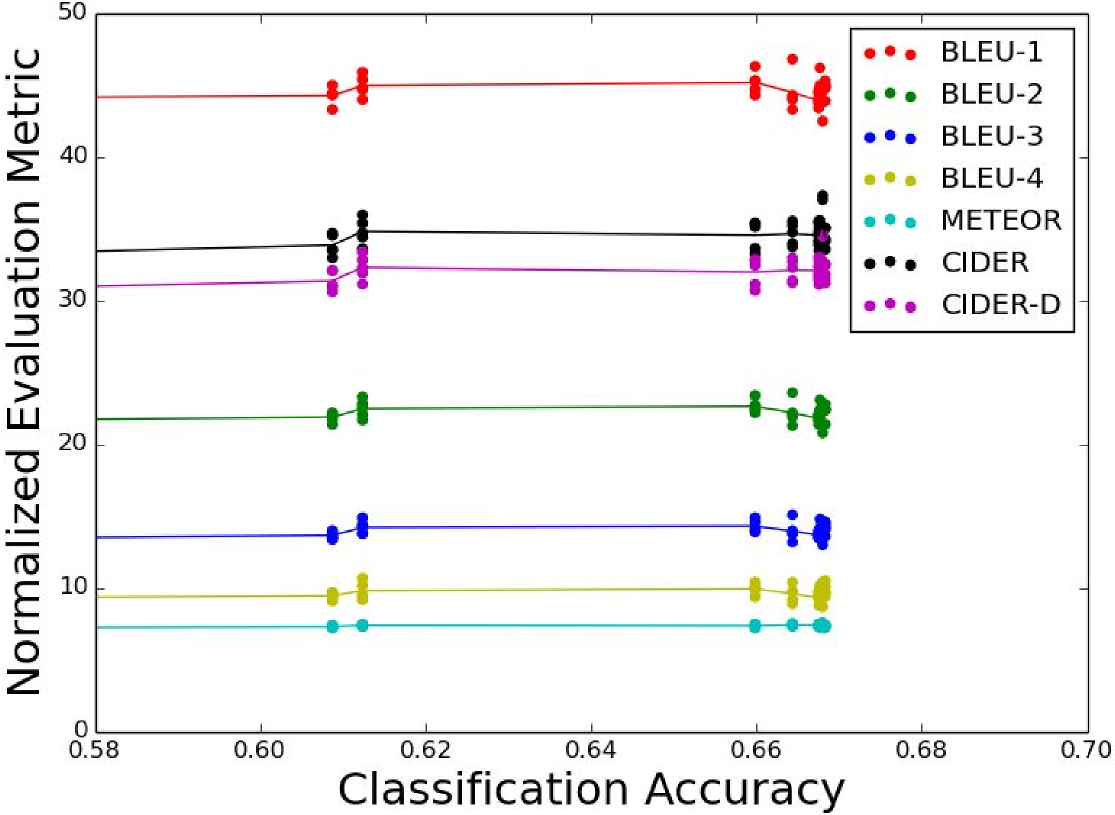
66.8% Rank-1 Accuracy



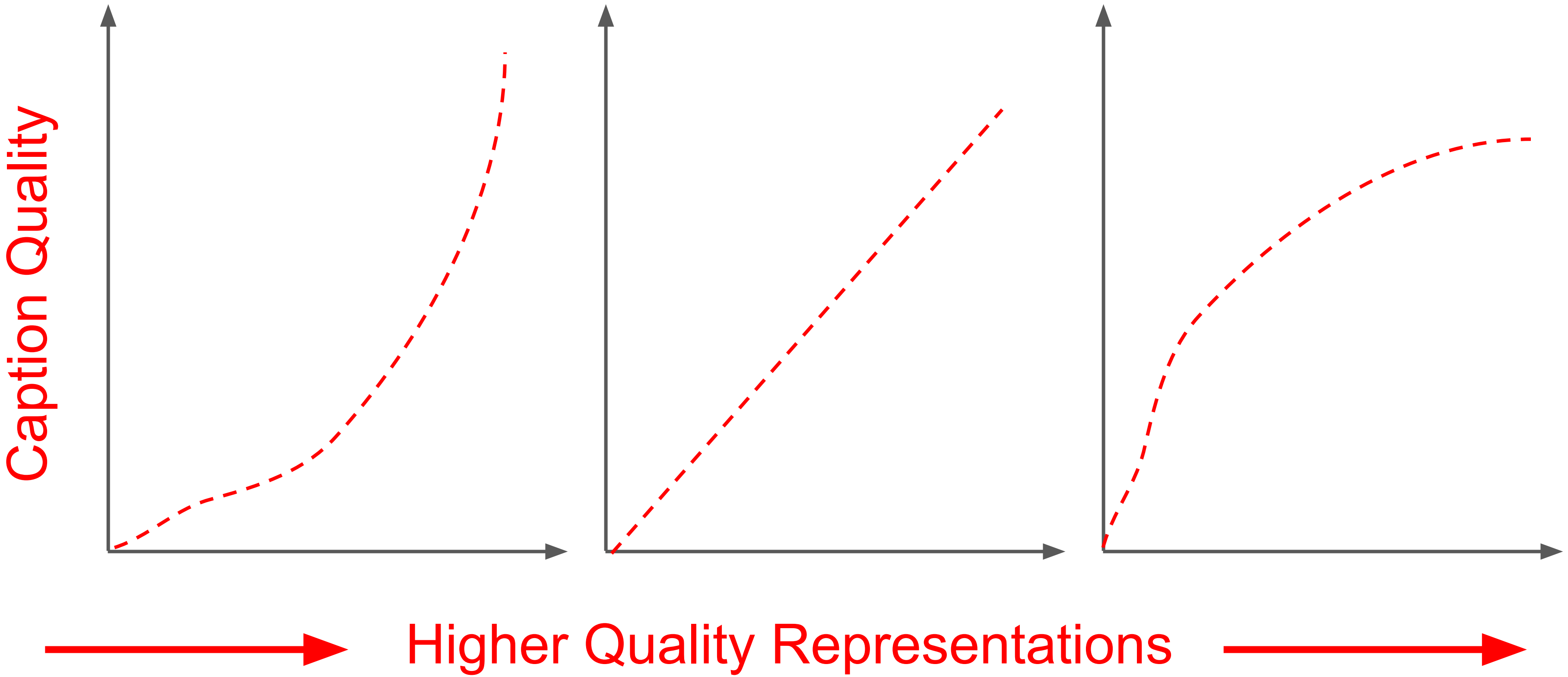
[Bossard et al. 2014]



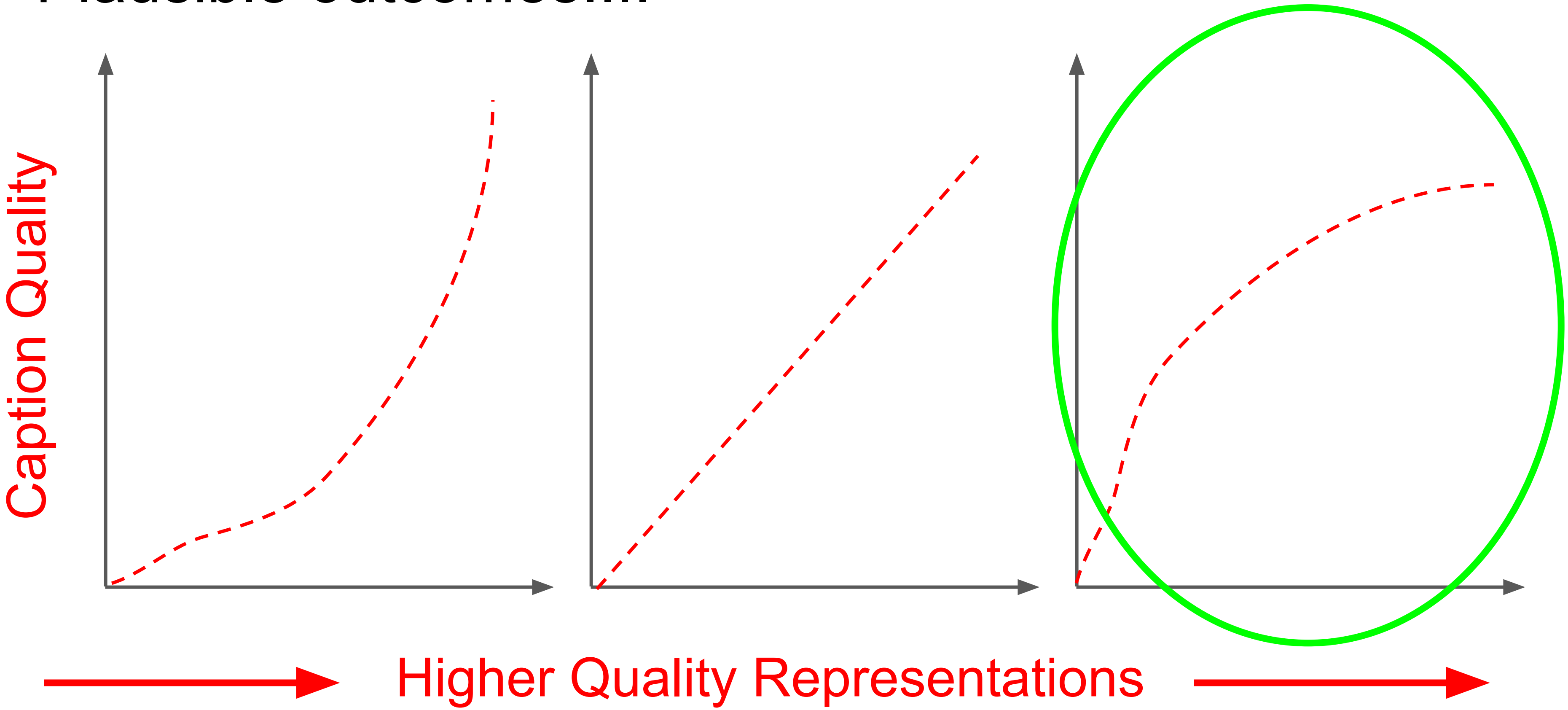
Transfer learned: Yummly



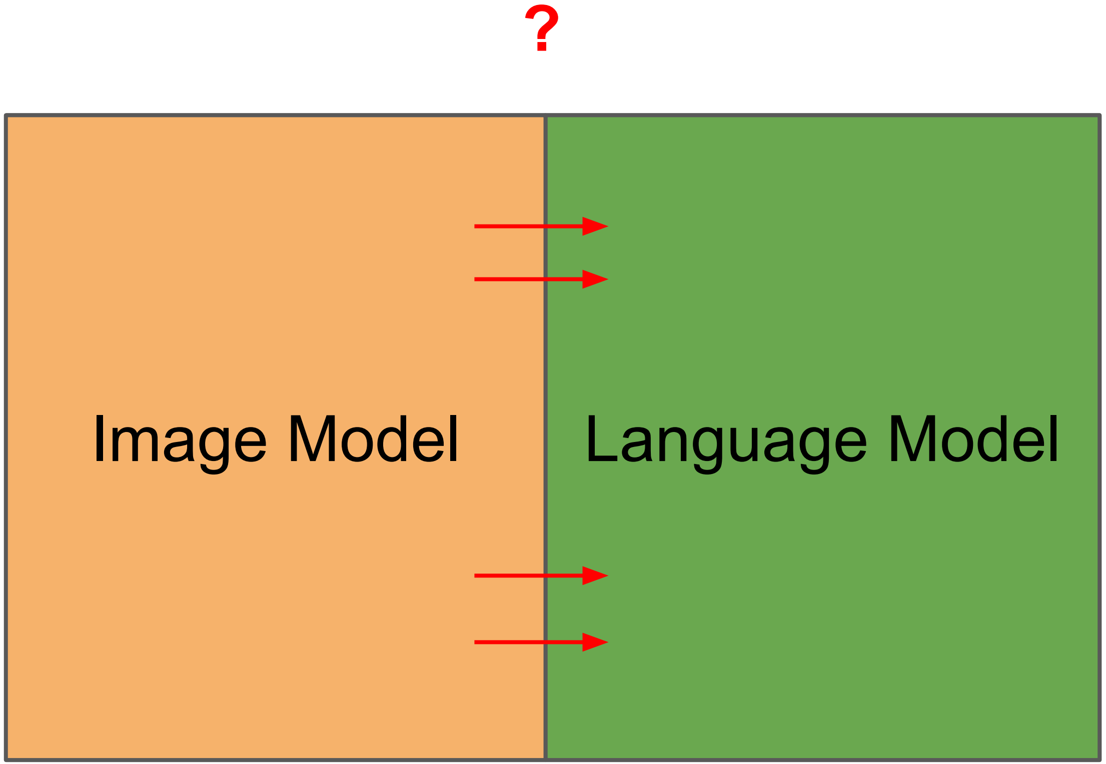
Plausible outcomes....



Plausible outcomes....

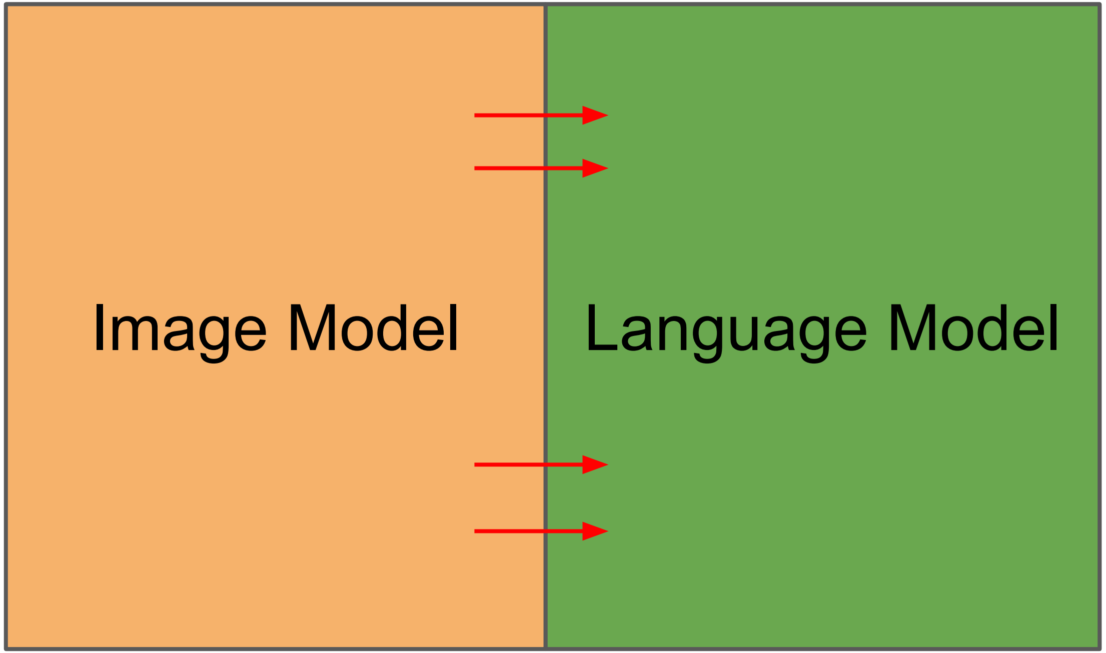


Implications



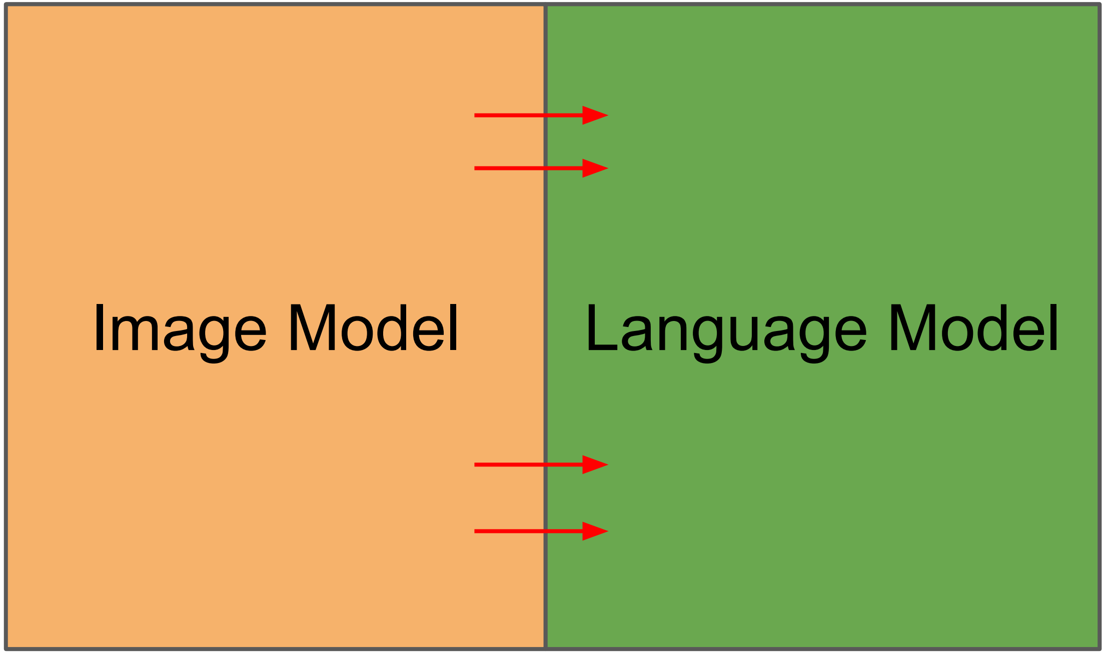
Implications

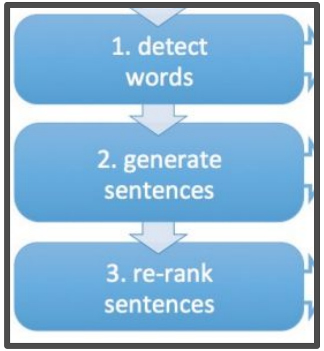
Mostly Coarse-Grained Information



Implications

Mostly Coarse-Grained Information

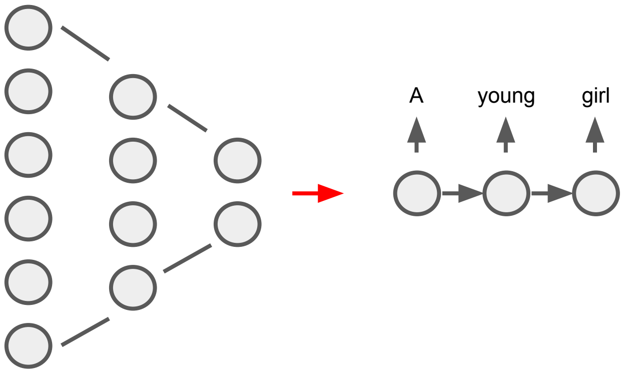




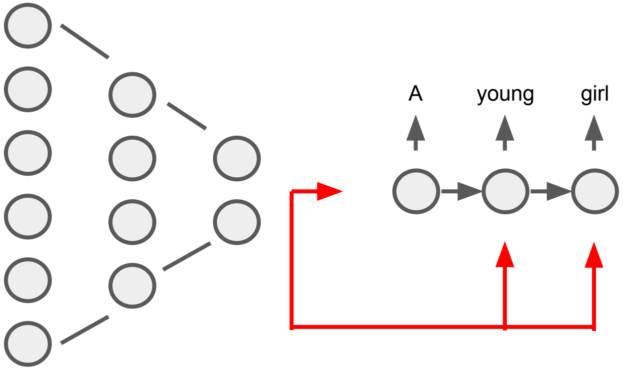
[Fang et al. 2014]

Towards Fine-Grained Captioning

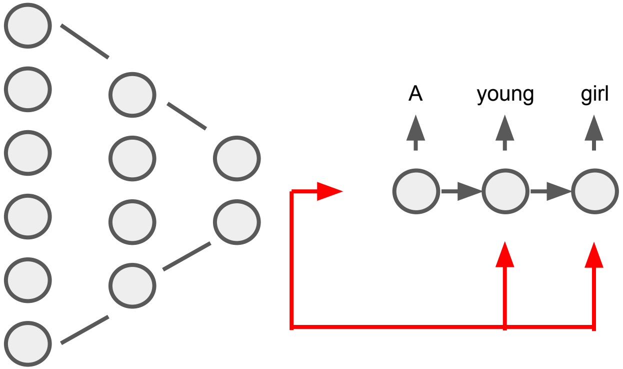
Towards Fine-Grained Captioning



Towards Fine-Grained Captioning



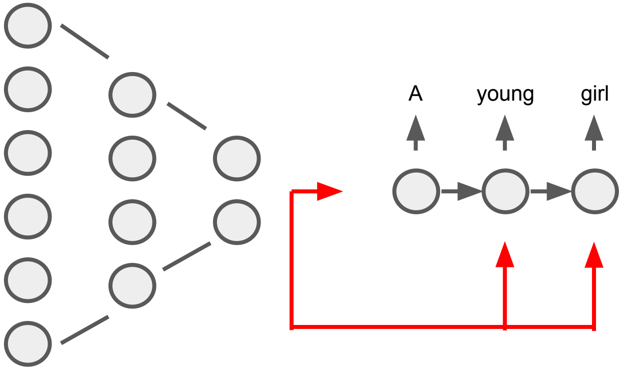
Towards Fine-Grained Captioning



Microsoft
Common Objects in Context

[Lin et al. 2014]

Towards Fine-Grained Captioning



Microsoft
Common Objects in Context

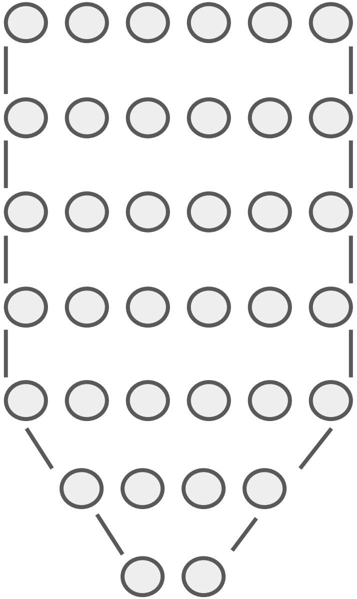
[Lin et al. 2014]

Better Loss Functions!

A few more experiments in the paper...

A few more experiments in the paper...

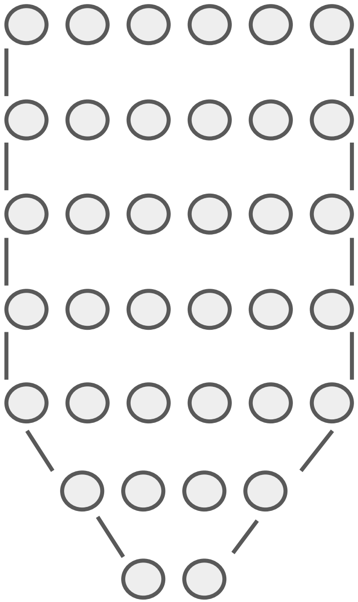
VGGNet



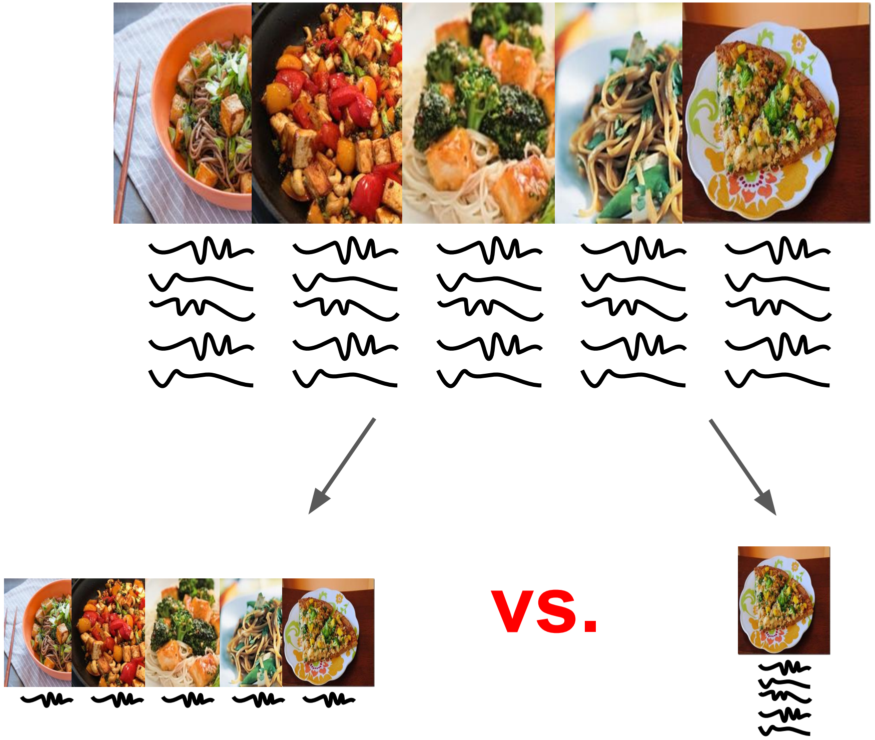
[Simonyan and Zisserman 2014]

A few more experiments in the paper...

VGGNet

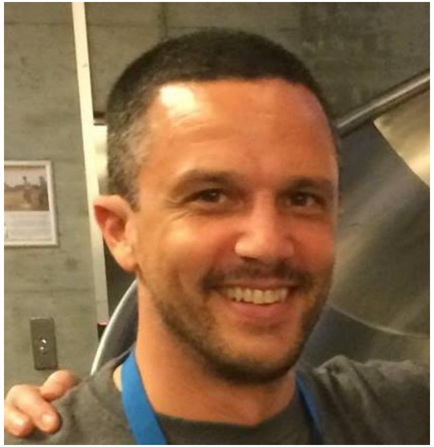


[Simonyan and Zisserman 2014]



Acknowledgements

Lillian Lee
David Mimno
Jason Yosinski
Serge Belongie
Xanda Schofield
Gregory Druck
Abby Lewis
CS 6670 Students



Serge

Xanda

David

Jason

Gregory

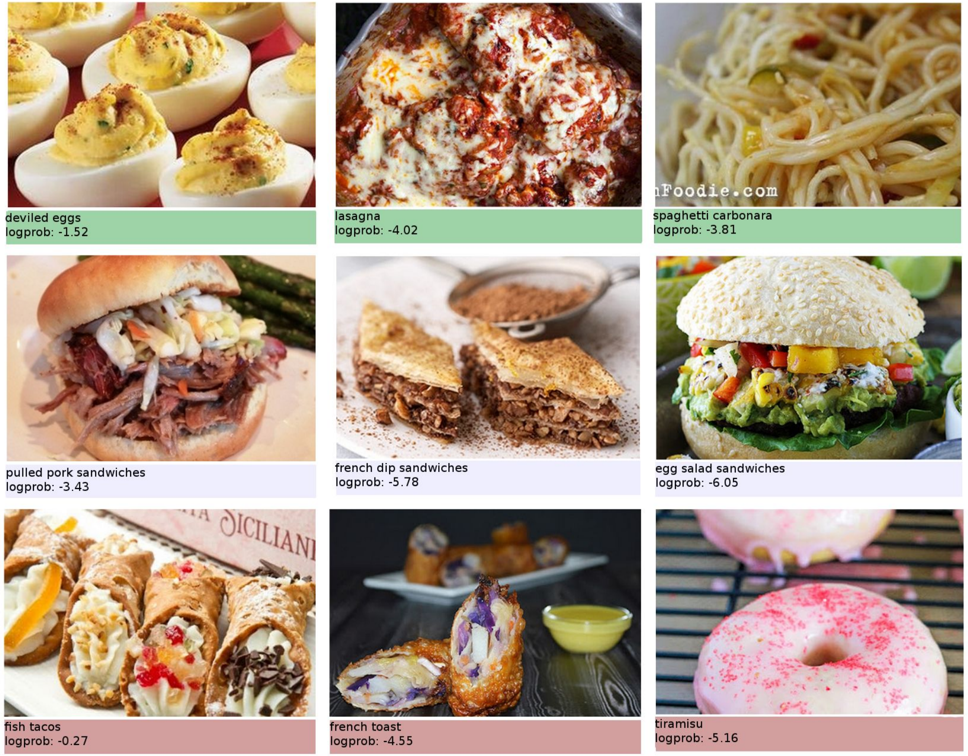
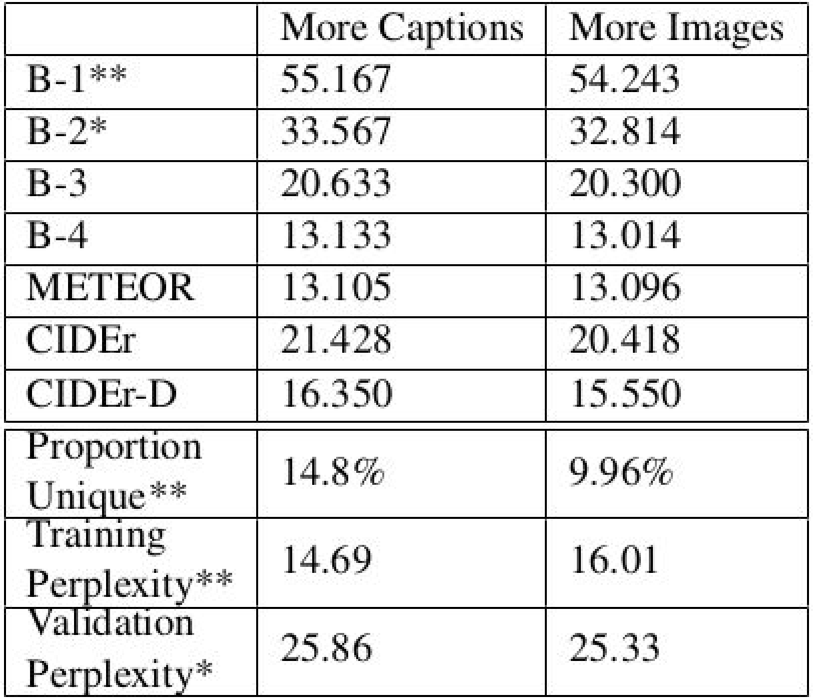


Image Representations and New Domains in Neural Image Captioning

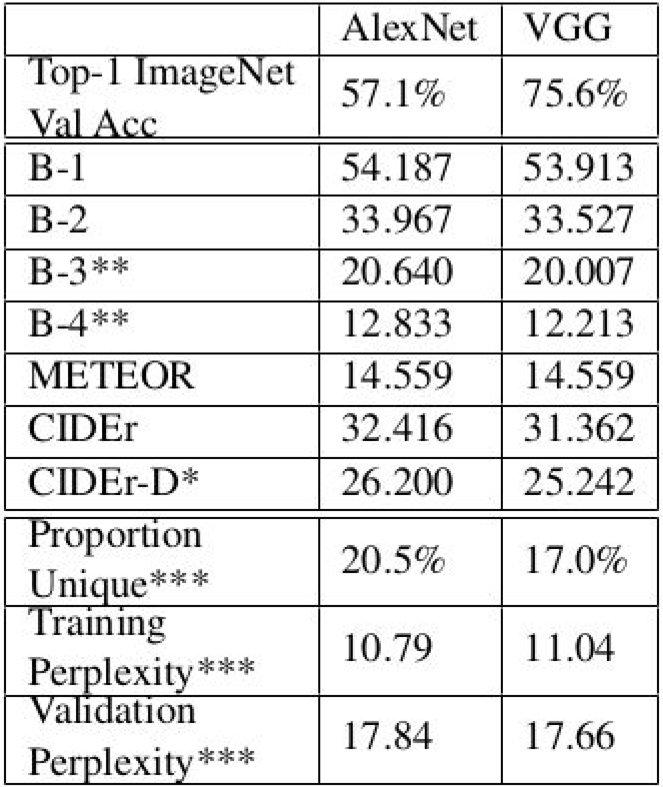
Jack Hessel, Nic Saava, Mike Wilber

New domain advice: one caption per image?



	More Captions	More Images
B-1**	55.167	54.243
B-2*	33.567	32.814
B-3	20.633	20.300
B-4	13.133	13.014
METEOR	13.105	13.096
CIDEr	21.428	20.418
CIDEr-D	16.350	15.550
Proportion Unique**	14.8%	9.96%
Training Perplexity**	14.69	16.01
Validation Perplexity*	25.86	25.33

AlexNet vs VGG-Net



	AlexNet	VGG
Top-1 ImageNet Val Acc	57.1%	75.6%
B-1	54.187	53.913
B-2	33.967	33.527
B-3**	20.640	20.007
B-4**	12.833	12.213
METEOR	14.559	14.559
CIDEr	32.416	31.362
CIDEr-D*	26.200	25.242
Proportion Unique***	20.5%	17.0%
Training Perplexity***	10.79	11.04
Validation Perplexity***	17.84	17.66