Sharing is Caring in the Land of The Long Tail

Samy Bengio

Google

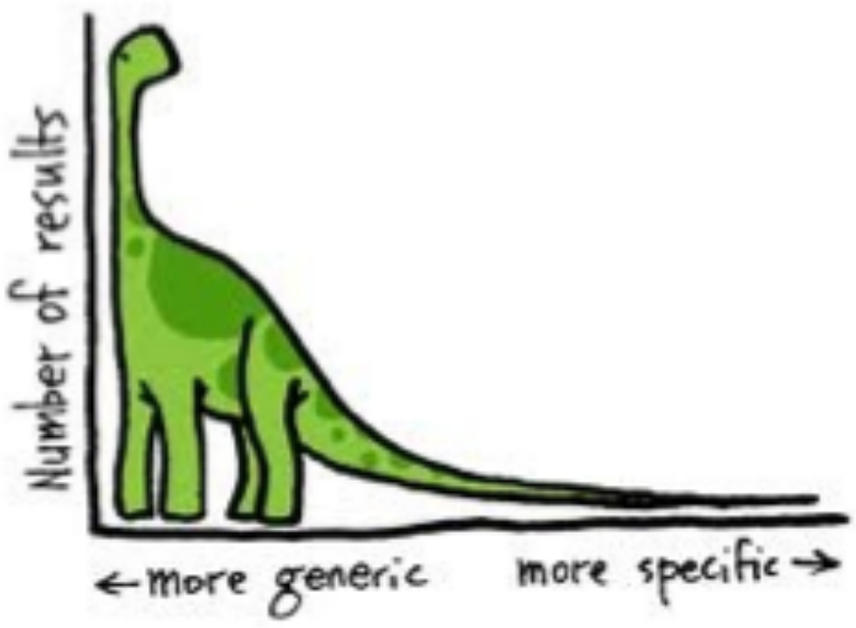
Real life setting

"Real problems rarely come packaged as 1M images uniformly belonging to a set of 1000 classes…"

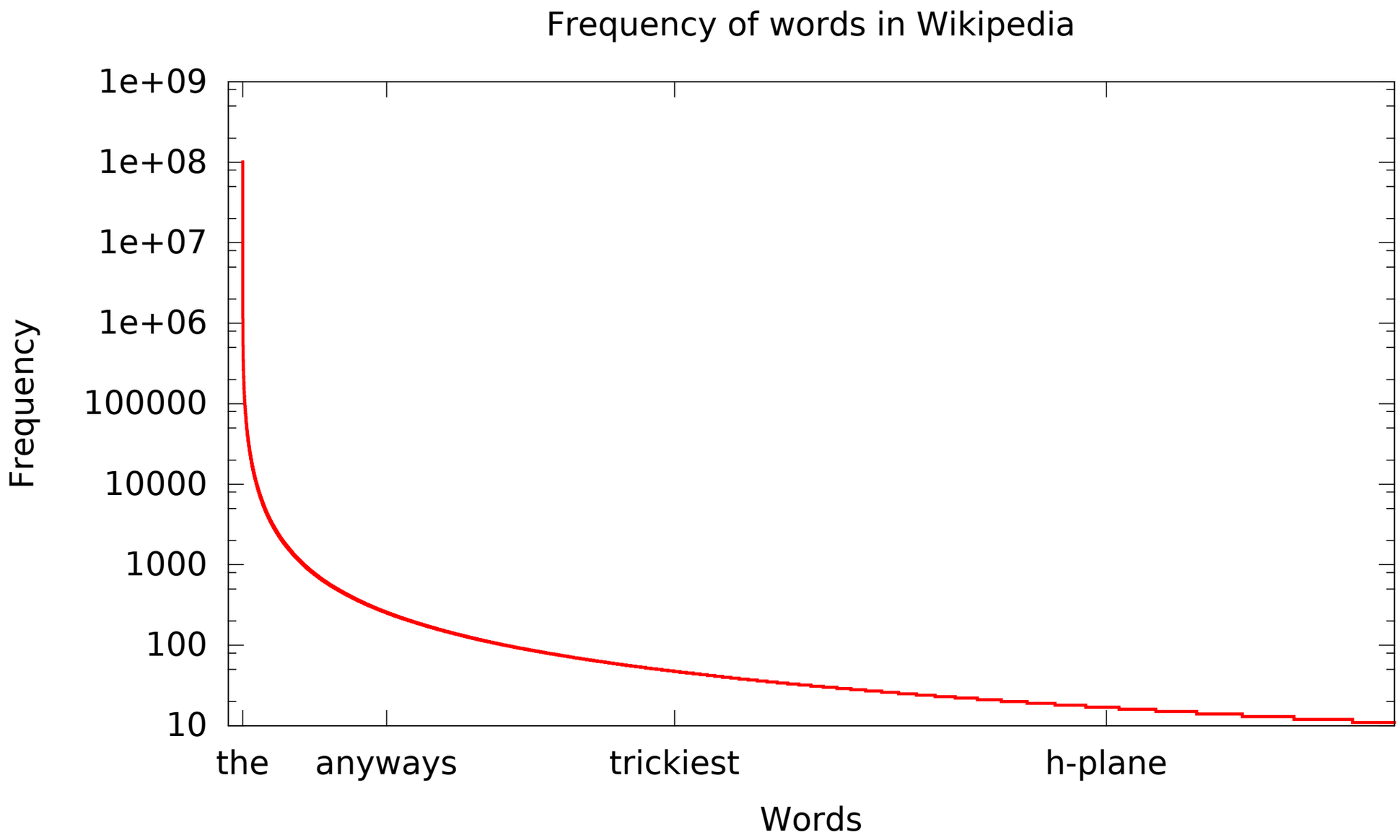
Google

The long tail

- Well known phenomena where a small number of generic objects/entities/words appear very often and most others appear more rarely.
- Also knows as Zipf or Power law, or Pareto distribution.
- The web is littered by this kind of distributions:
  - the frequency of each unique query on search engines,
  - the occurrences of each unique word in text documents,
  - etc.



Example of a long tail



Representation sharing

- How do we design a classifier or a ranker when data follows a long tail distribution?
- If we train one model per class, it is hard for poor classes to be well trained.
- How come we humans are able to recognize objects we have seen only once or even never?
- Most likely answer: representation sharing: all class models share/learn a joint representation.
- Poor classes can then benefit from knowledge learned from semantically similar but richer classes.
- Extreme case: zero-shot setting!

Google

Outline

In this talk, I will cover the following ideas:

- Wsabie: a joint embedding space of images and labels
- The many facets of text embeddings
- Zero-shot setting through embeddings
- Incorporate Knowledge Graph constraints
- Use of a language model

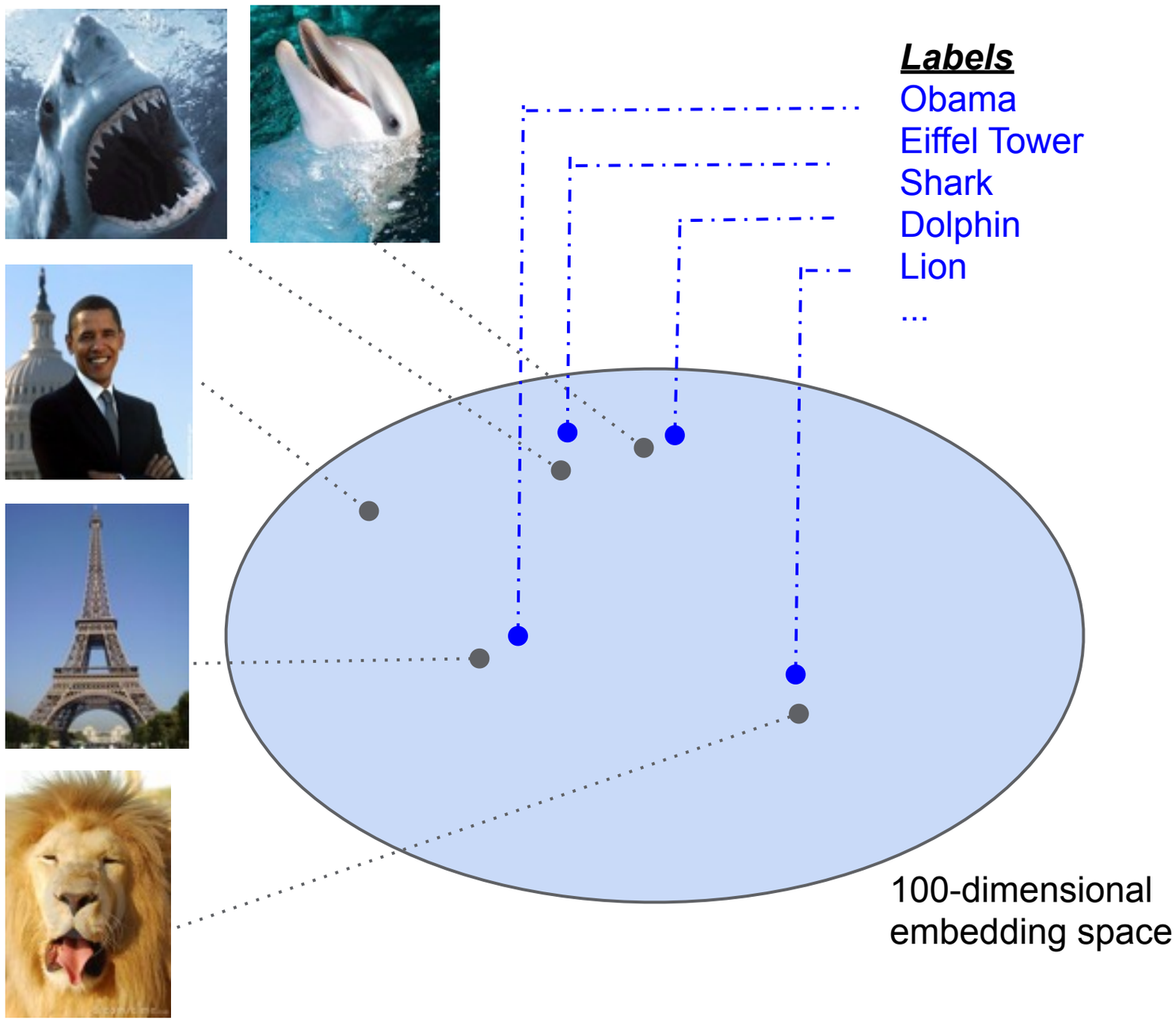
I will NOT cover the following important issues:

- Prediction time issues for extreme classification
- Memory issues

Google

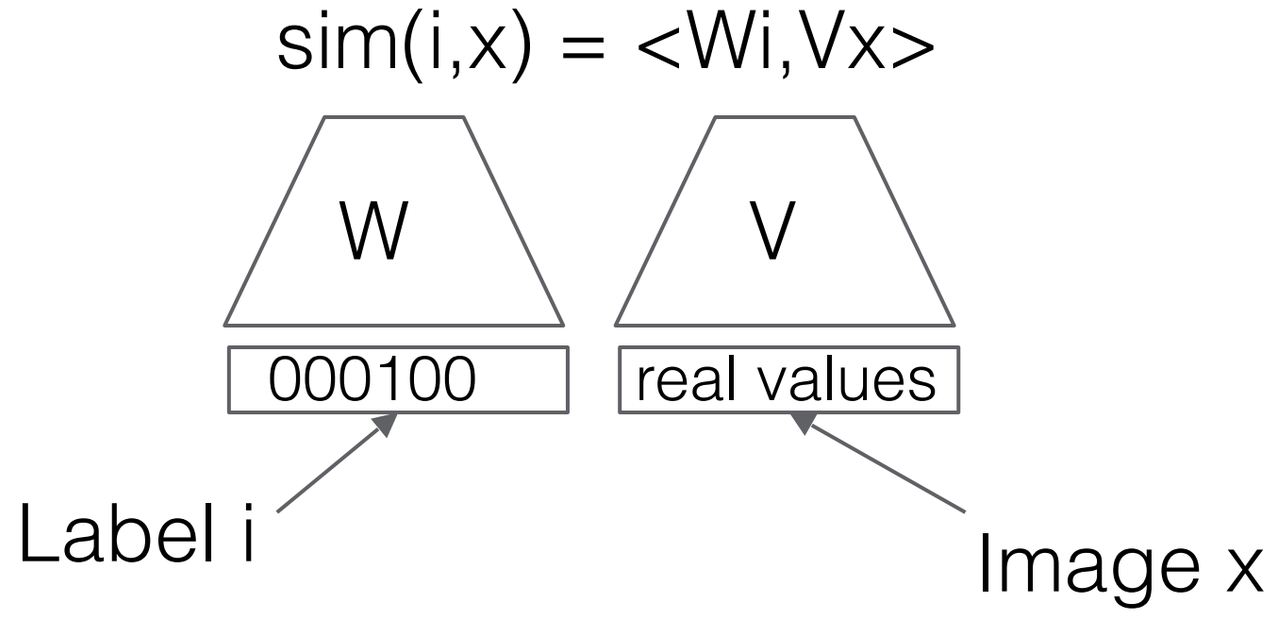
Wsabie

Learn to embed images & labels to optimize top-ranked items.



Wsabie: J. Weston et al, ECML 2010, IJCAI 2011

Wsabie: summary



Triplet Loss: sim( , dolphin) > sim( , obama) + 1

Trained by stochastic gradient descent and smart sampling of negative examples

Google

Wsabie: experiments - results

Method	ImageNet 2010		Web	
	prec@1	prec@10	prec@1	prec@10
approx kNN	1.55%	0.41%	0.30%	0.34%
One-vs-Rest	2.27%	1.02%	0.52%	0.29%
Wsabie	4.03%	1.48%	1.03%	0.44%
Ensemble of 10 Wsabies	10.03%	3.02%		

ImageNet 2010: 16000 labels and 4M images
Web: 109000 labels and 16M images

Wsabie: embeddings

Label	Nearest Neighbors
barack obama	barak obama, obama, barack, barrack obama, bow wow
david beckham	beckham, david beckam, alessandro del piero, del piero
santa	santa claus, papa noel, pere noel, santa clause, joyeux noel
dolphin	delphin, dauphin, whale, delfin, delfini, baleine, blue whale
cows	cattle, shire, dairy cows, kuh, horse, cow, shire horse, kone
rose	rosen, hibiscus, rose flower, rosa, roze, pink rose, red rose
eiffel tower	eiffel, tour eiffel, la tour eiffel, big ben, paris, blue mosque
ipod	i pod, ipod nano, apple ipod, ipod apple, new ipod
f18	f 18, eurofighter, f14, fighter jet, tomcat, mig 21, f 16

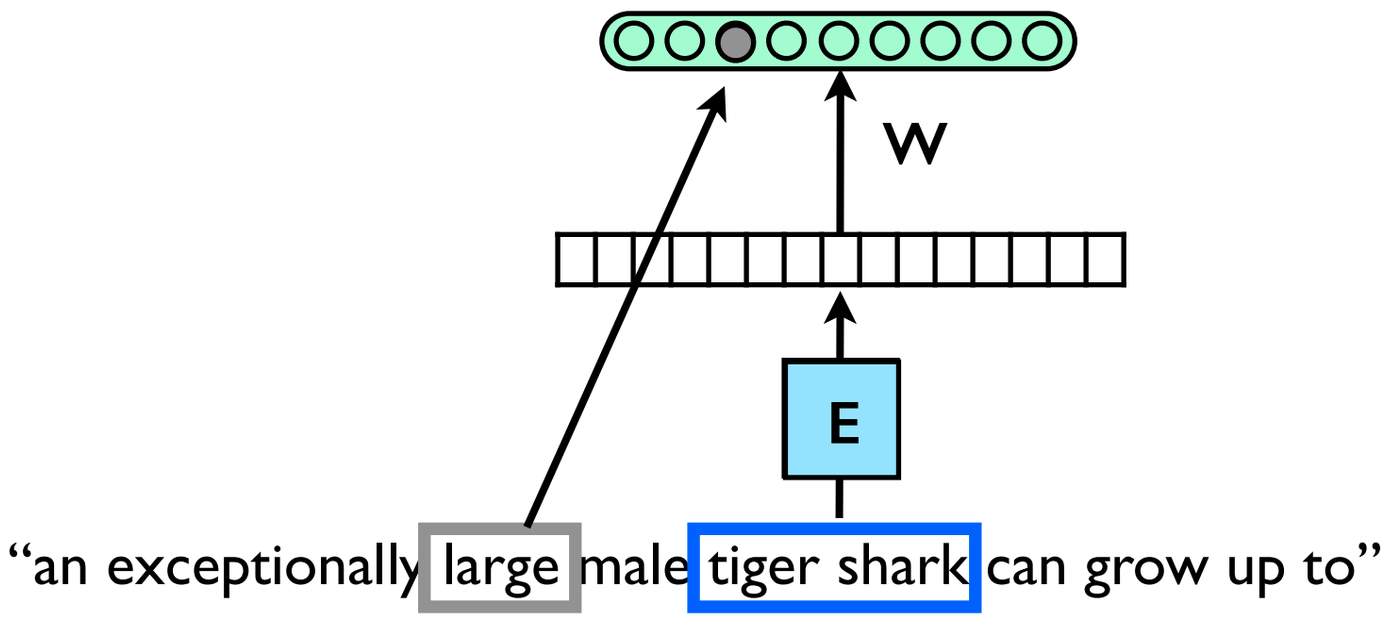
Wsabie: annotations

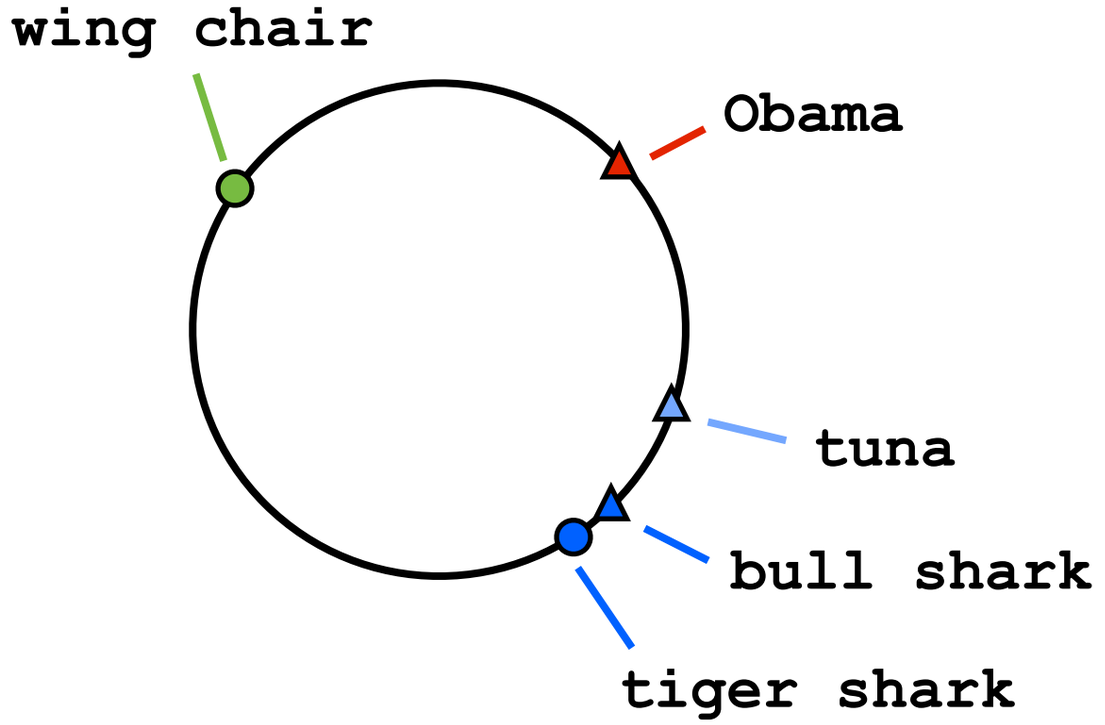
	delfini, orca, **dolphin**, mar, delfin, dauphin, whale, cancun, killer whale, sea world
	blue whale, whale shark, great white shark, underwater, white shark, shark, manta ray, **dolphin**, requin, blue shark, diving
	barrack obama, barak obama, barack hussein obama, **barack obama**, james marsden, jay z, obama, nelly, falco, barack
	eiffel, paris by night, la tour eiffel, tour eiffel, **eiffel tower**, las vegas strip, eifel, tokyo tower, eifel tower

“Why not an embedding of text only?”

Skip-Gram (Word2Vec)

Learn dense embedding vectors from an unannotated text corpus, e.g. Wikipedia





http://code.google.com/p/word2vec

Tomas Mikolov, Kai Chen, Greg Corrado, Jeff Dean (ICLR 2013)

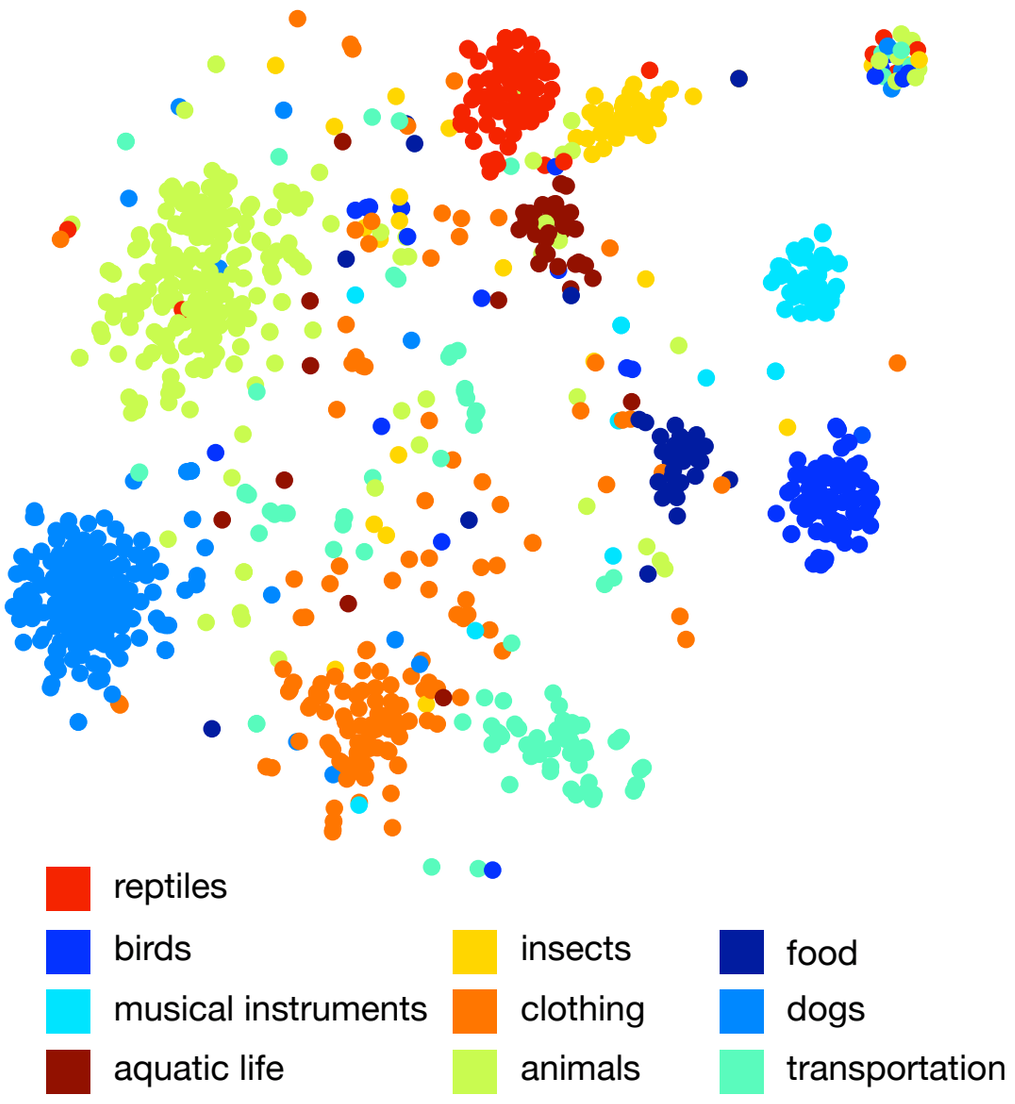
Google

Skip-Gram Wikipedia

Skip-gram trained on Wikipedia, 155K terms

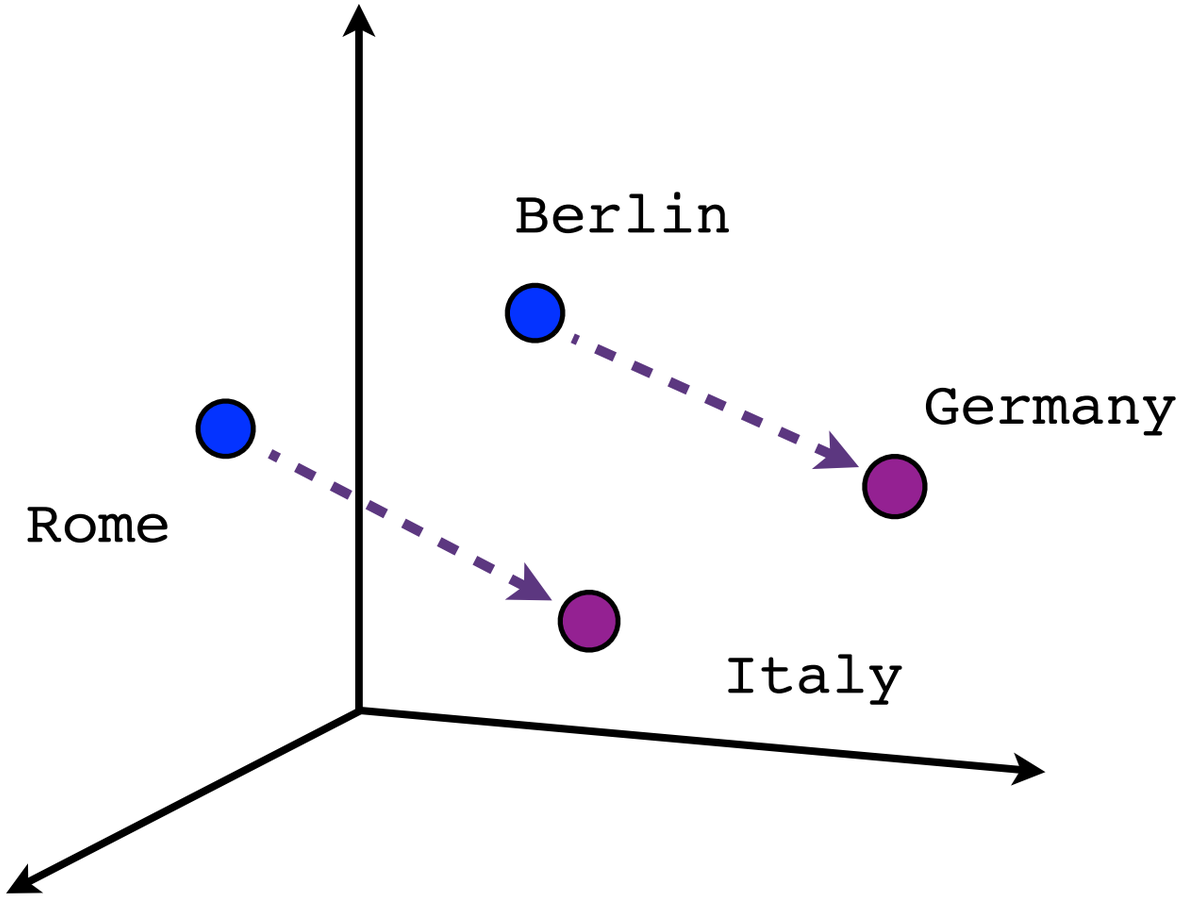
tiger shark	**car**
bull shark	cars
blacktip shark	muscle car
shark	sports car
oceanic whitetip shark	compact car
sandbar shark	autocar
dusky shark	automobile
blue shark	pickup truck
requiem shark	racing car
great white shark	passenger car
lemon shark	dealership

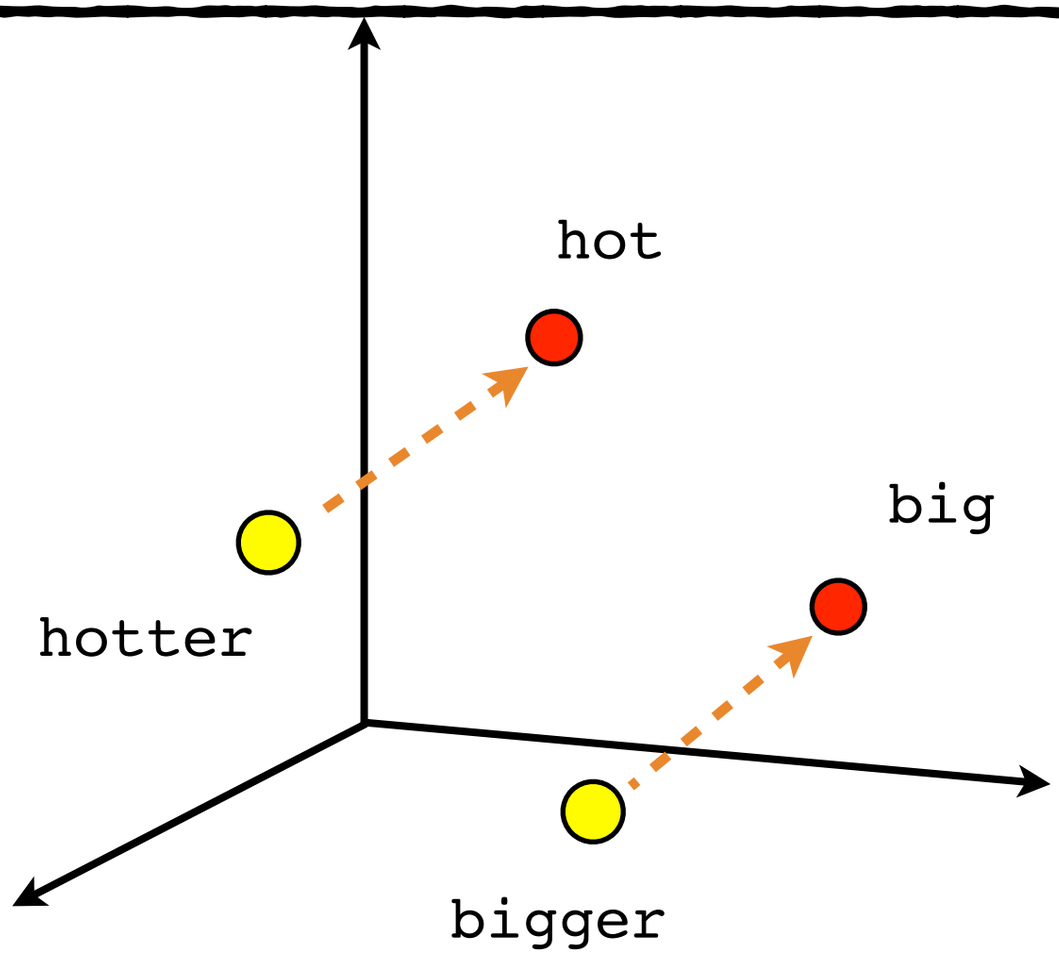
t-SNE visualization of ImageNet labels



Google

Embeddings are powerful

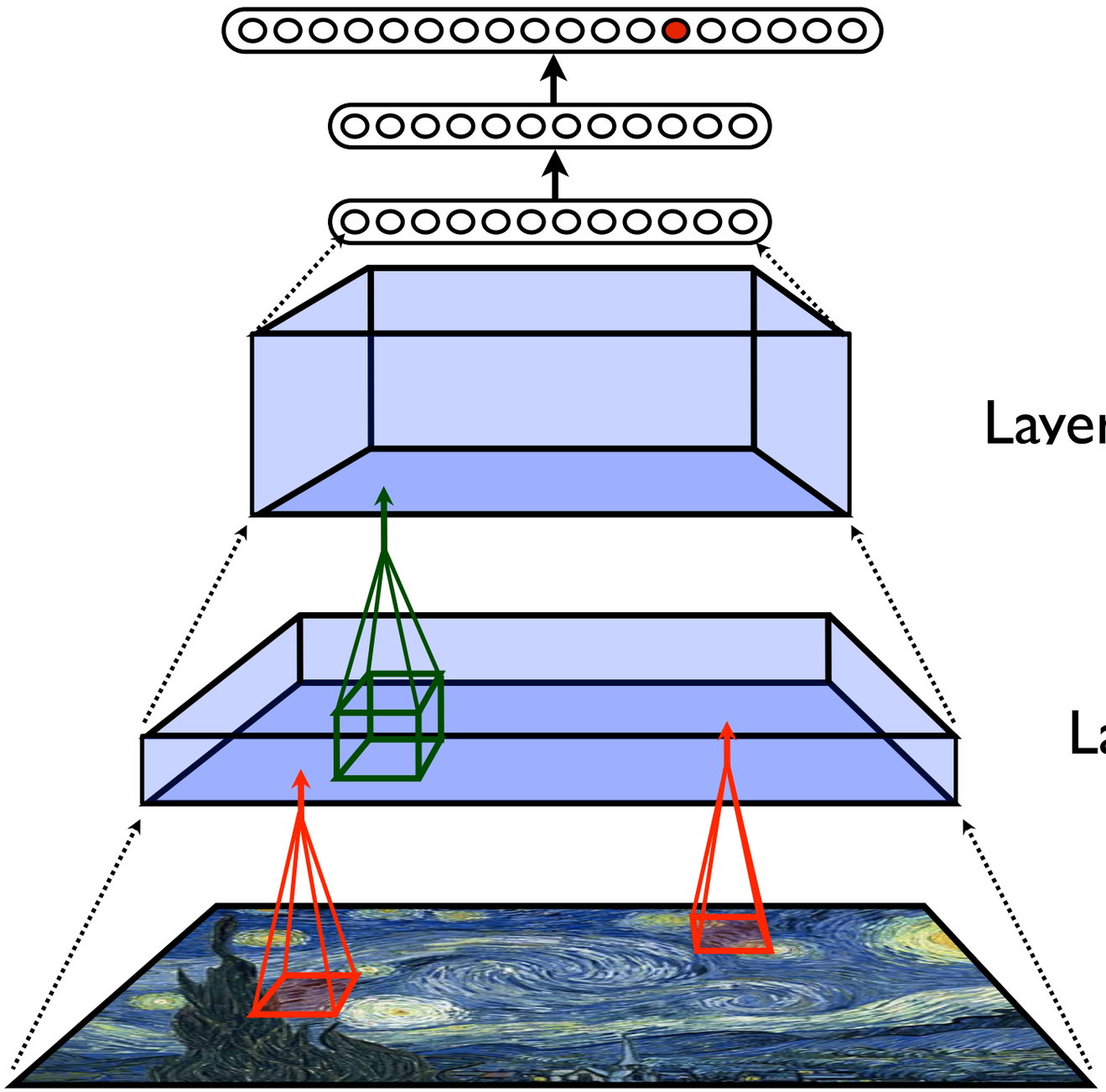




E(*Rome*) - E(*Italy*) + E(*Germany*) ≈ E(*Berlin*)
E(*hotter*) - E(*hot*) + E(*big*) ≈ E(*bigger*)

Let’s go back to images!

Deep convolutional models for images

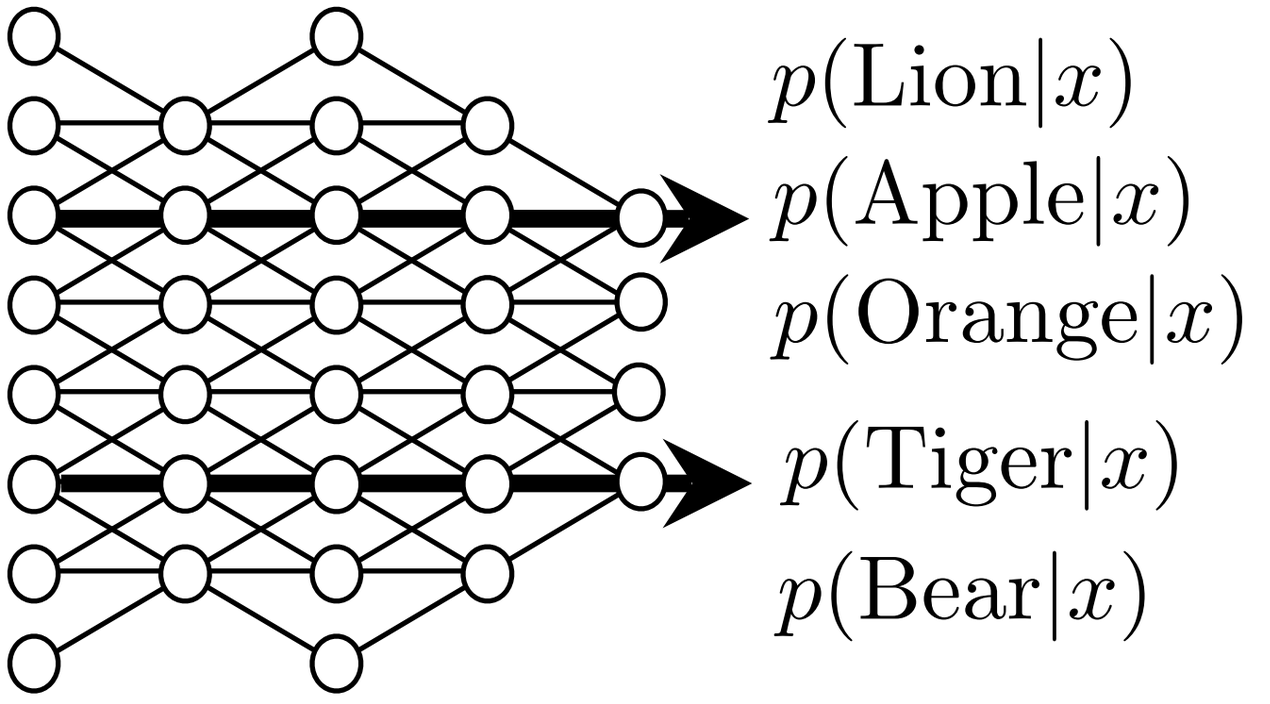


But what about the long tail of classes?

What about using our semantic embeddings for that?

Google

ConSE: Convex Combination of Semantic Embeddings [Norouzi et al, ICLR'2014]



Google

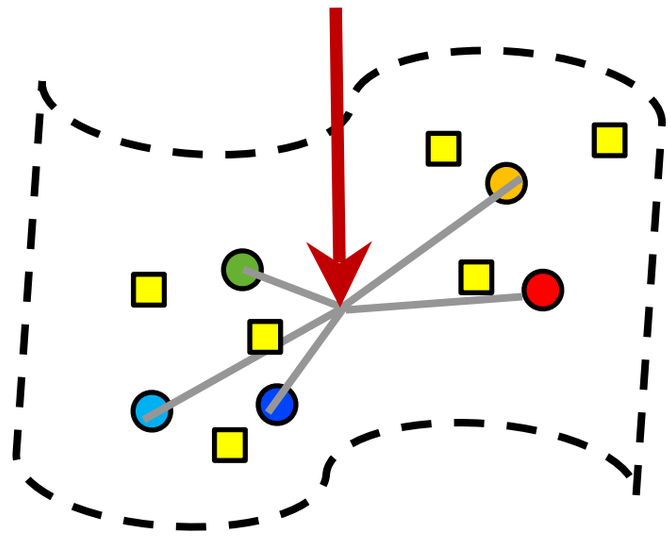
ConSE: Convex Combination of Semantic Embeddings

from Skip-Gram for instance:

$s(\mathrm{y}) = \text{embedding position of } y$

$$f(x) = \sum_i p(y_i|x)s(y_i)$$

$$\begin{aligned} f(x) = {} & p(\text{Lion}|x)s(\text{Lion}) + \\ & p(\text{Apple}|x)s(\text{Apple}) + \\ & p(\text{Orange}|x)s(\text{Orange}) + \\ & p(\text{Tiger}|x)s(\text{Tiger}) + \\ & p(\text{Bear}|x)s(\text{Bear}) \end{aligned}$$



Do a nearest neighbor search around $f(x)$ to find the corresponding label

ConSE(T): Convex Combination of Semantic Embeddings

In practice, consider the
average of only a few labels:

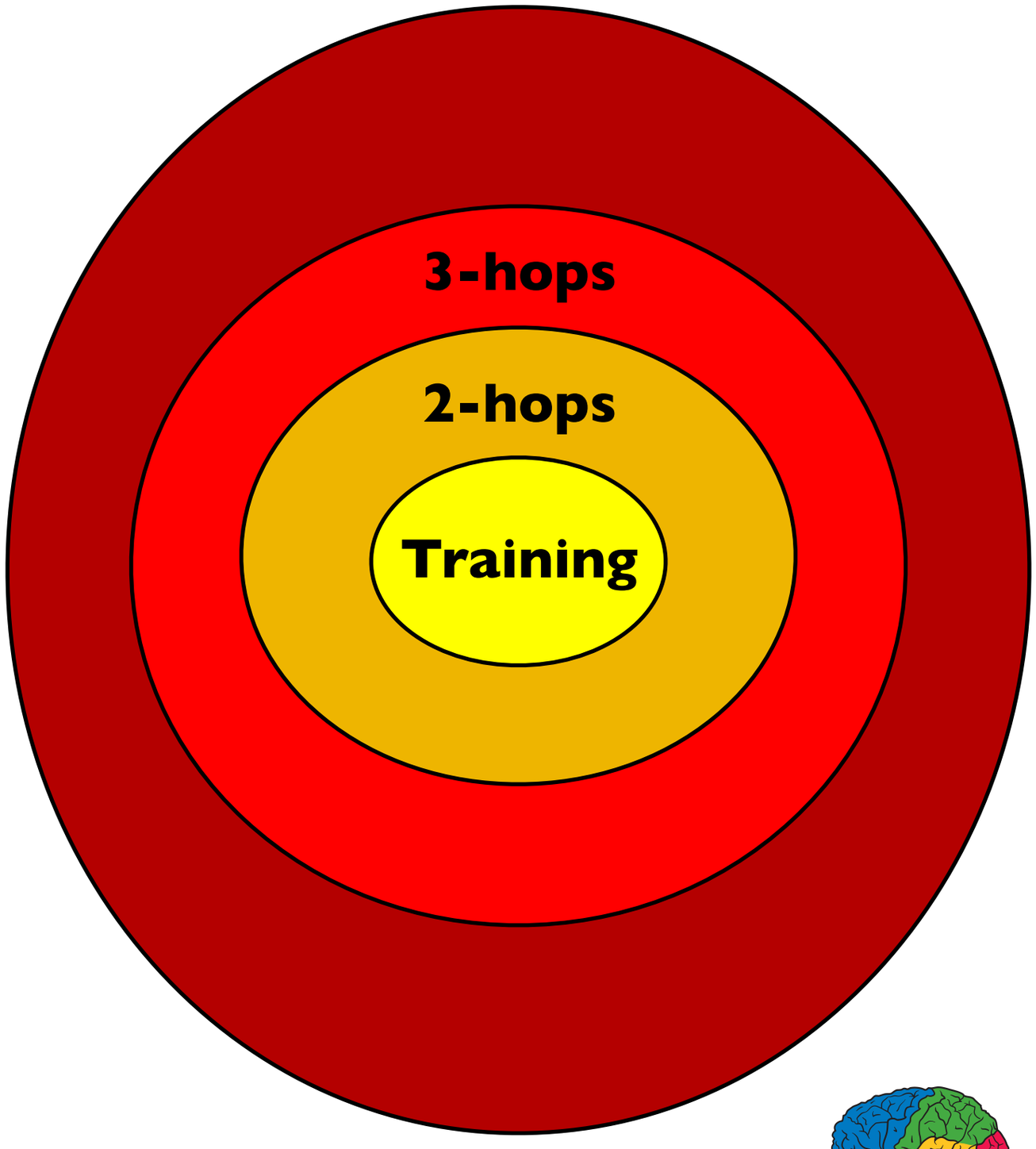
$$top(T) = \{i \mid p(y_i|x) \text{ is among top } T \text{ probabilities}\}$$

$$f(x) = \frac{1}{Z} \sum_{i \in top(T)} p(y_i|x) s(y_i)$$

Google

ConSE(T): experiments on ImageNet

- Model trained with 1.2M ILSVRC 2012 images from 1,000 classes
- Evaluated on images from same classes.
- Results are measured as hit@*k*.



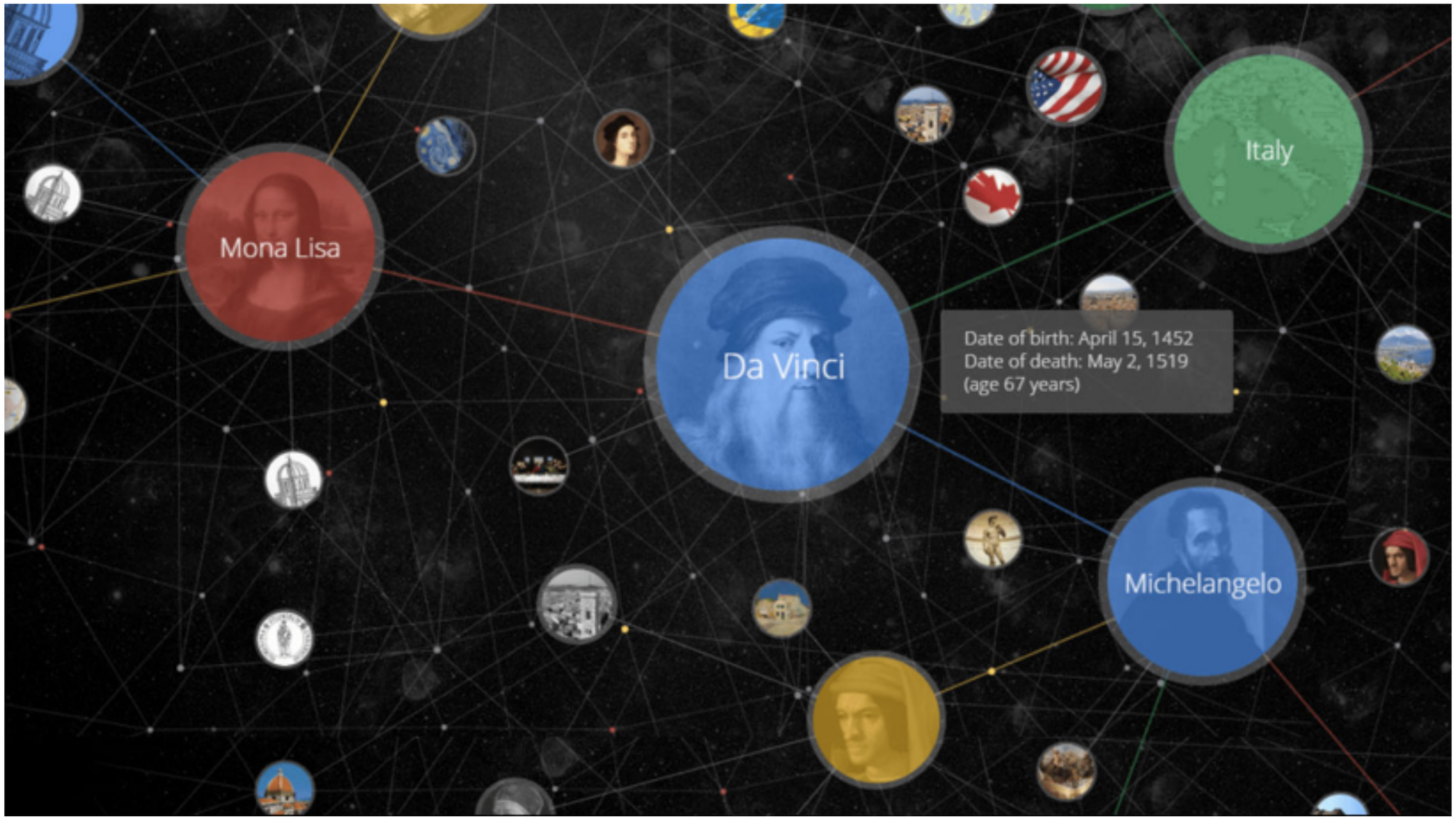
Google

ConSe(T) experiments

Test Label Set	# Candidate Labels	Model	Flat hit@k (%)				
			1	2	5	10	20
2-hops	1,589	DeViSE	6.0	10.0	18.1	26.4	36.4
		ConSE(1)	9.3	14.4	23.7	30.8	38.7
		ConSE(10)	**9.4**	**15.1**	**24.7**	**32.7**	**41.8**
		ConSE(1000)	9.2	14.8	24.1	32.1	41.1
2-hops (+1K)	1,589 +1000	DeViSE	**0.8**	2.7	7.9	14.2	22.7
		ConSE(1)	0.2	**7.1**	**17.2**	24.0	31.8
		ConSE(10)	0.3	6.2	17.0	**24.9**	**33.5**
		ConSE(1000)	0.3	6.2	16.7	24.5	32.9
3-hops	7,860	DeViSE	1.7	2.9	5.3	8.2	12.5
		ConSE(1)	2.6	4.2	7.3	10.8	14.8
		ConSE(10)	**2.7**	**4.4**	**7.8**	**11.5**	**16.1**
		ConSE(1000)	2.6	4.3	7.6	11.3	15.7
3-hops (+1K)	7,860 +1000	DeViSE	**0.5**	1.4	3.4	5.9	9.7
		ConSE(1)	0.2	**2.4**	**5.9**	9.3	13.4
		ConSE(10)	0.2	2.2	**5.9**	**9.7**	**14.3**
		ConSE(1000)	0.2	2.2	5.8	9.5	14.0
ImageNet 2011 21K	20,841	DeViSE	0.8	1.4	2.5	3.9	6.0
		ConSE(1)	1.3	2.1	3.6	5.4	7.6
		ConSE(10)	**1.4**	**2.2**	**3.9**	**5.8**	**8.3**
		ConSE(1000)	1.3	2.1	3.8	5.6	8.1
ImageNet 2011 21K (+1K)	20,841 +1000	DeViSE	**0.3**	0.8	1.9	3.2	5.3
		ConSE(1)	0.1	1.2	3.0	4.8	7.0
		ConSE(10)	0.2	1.2	3.0	**5.0**	**7.5**
		ConSE(1000)	0.2	1.2	3.0	4.9	7.3

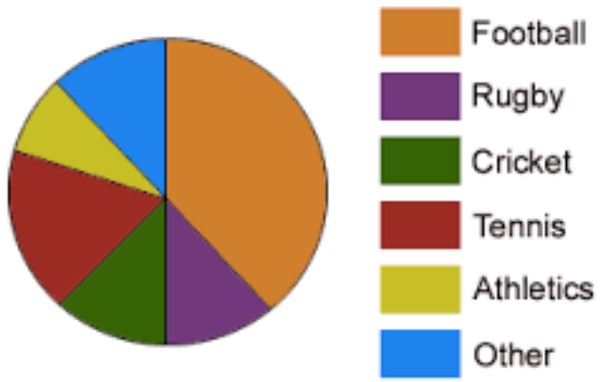
Google

Knowledge Graph

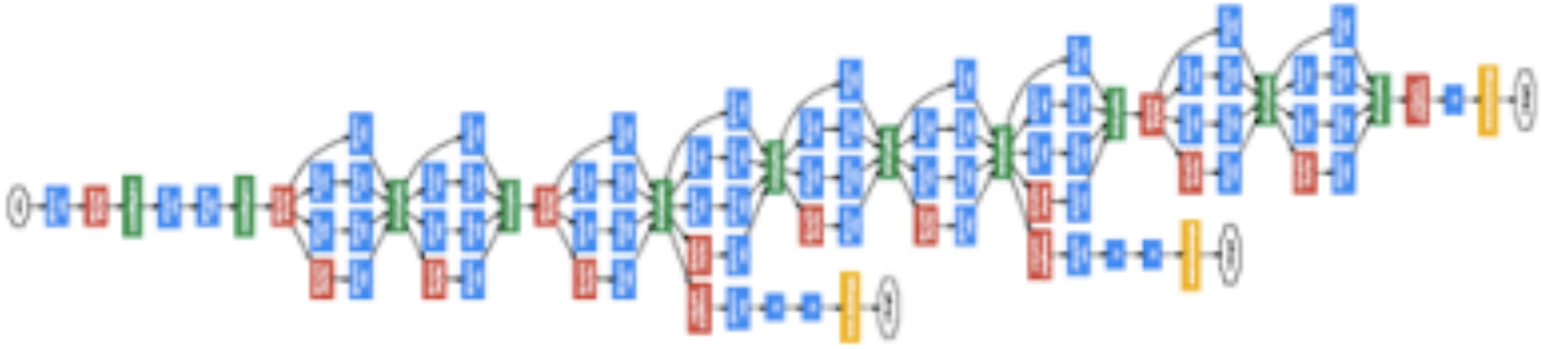


Multiclass Classifiers

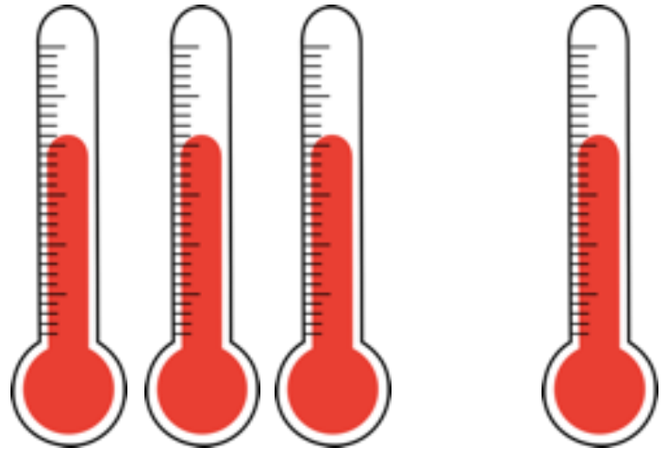
Softmax



GoogleLeNet model

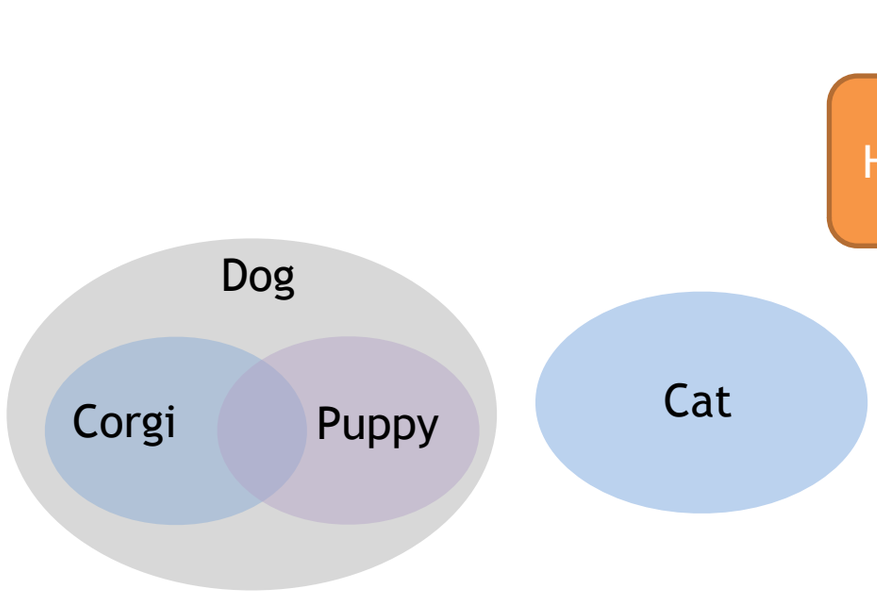


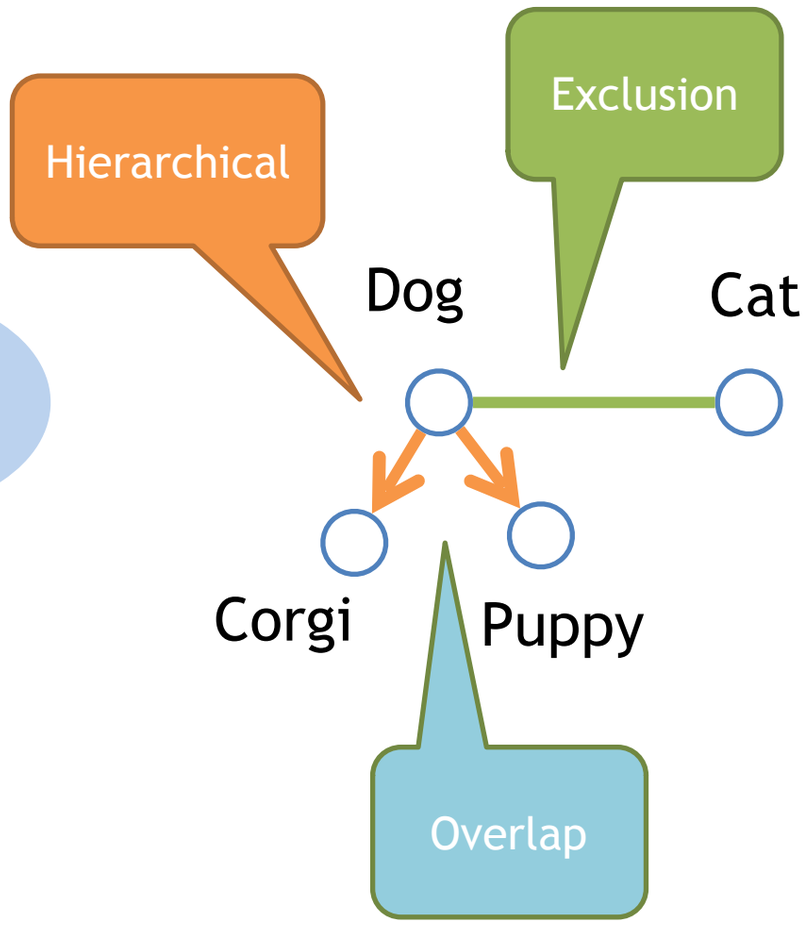
Logistic



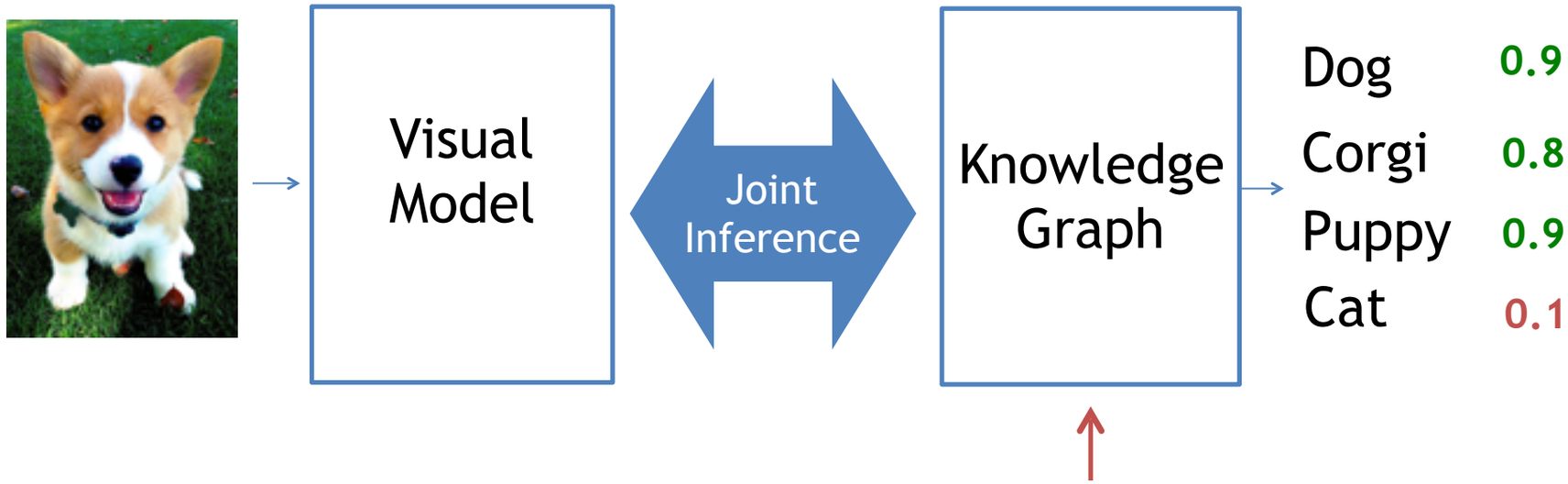
Google

Object labels have rich relations

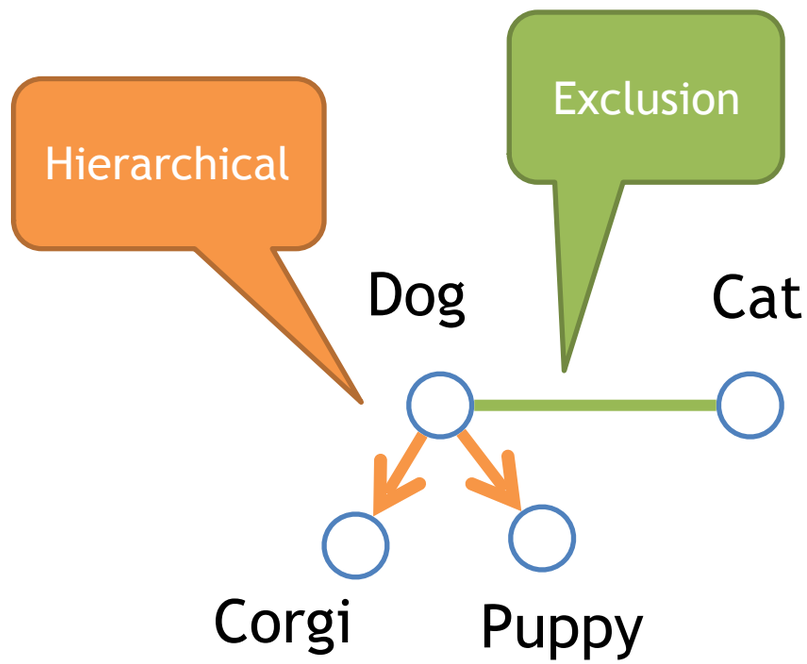




Visual Model + Knowledge Graph



Hierarchy and Exclusion (HEX) Graph

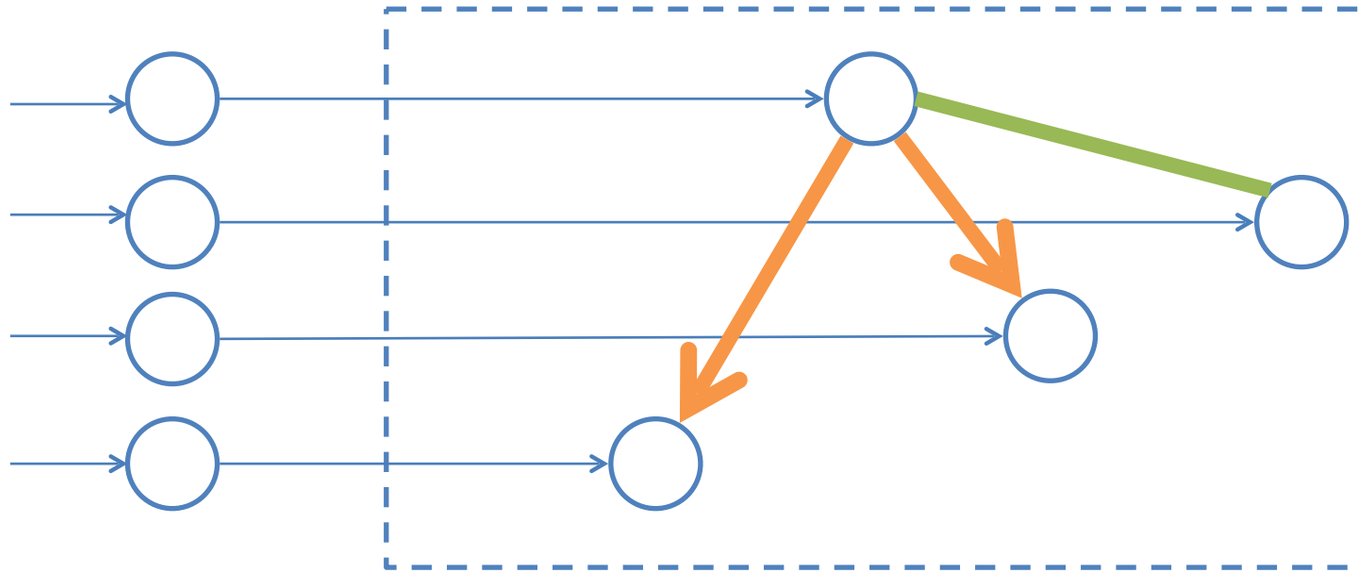


[Deng et al, ECCV 2014]

Google

HEX Classification Model

$x \in R^n$
Input scores



$y \in \{0,1\}^n$
Binary Label vector

$$\Pr(y \mid x) = \frac{1}{Z(x)} \prod_i \phi_i(x_i, y_i) \prod_{i,j} \psi_{i,j}(y_i, y_j)$$

$$\phi_i(x_i, y_i) = \begin{cases} sigmoid(x_i) & if \ y_i = 1 \\ 1 - sigmoid(x_i) & if \ y_i = 0 \end{cases}$$

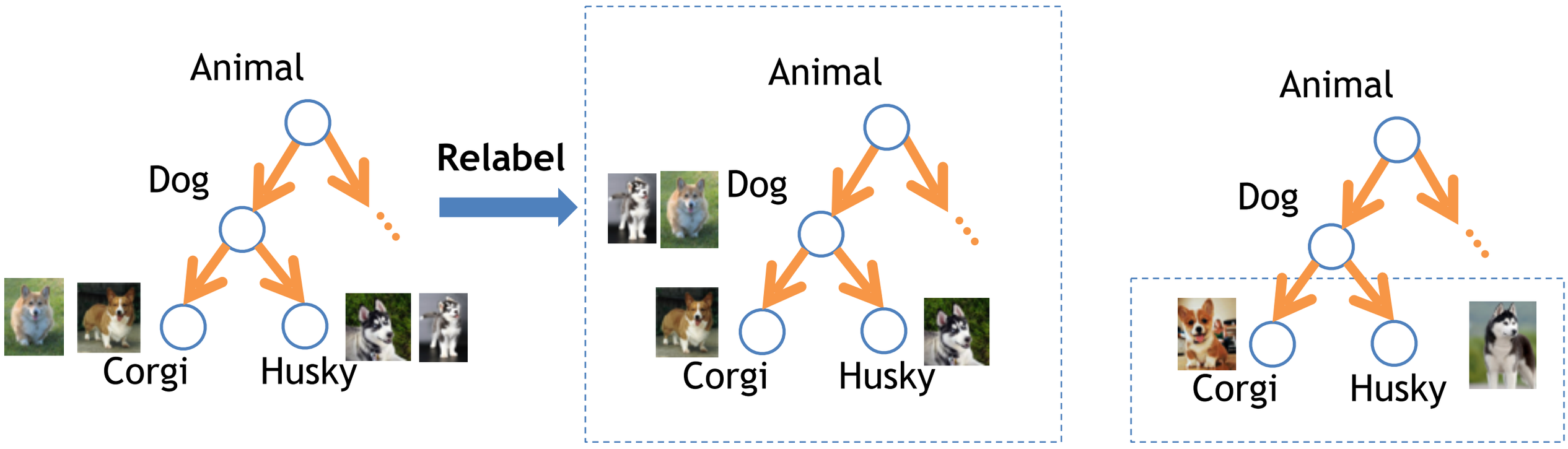
$$\psi_{i,j}(y_i, y_j) = \begin{cases} 0 & \text{If violates constraints} \\ 1 & \text{Otherwise} \end{cases}$$

Unary: same as logistic regression Pairwise: set illegal configuration to zero

Google → All illegal configurations have probability zero.

Exp: Learning with weak labels

- ILSVRC 2012: “relabel” or “weaken” a portion of fine-grained leaf labels to basic level labels.
- Evaluate on fine-grained recognition



Original ILSVRC 2012 (leaf labels)

Training (“weakened” labels)

Test

Google

Exp: Learning with weak labels

- ILSVRC 2012: “relabel” or “weaken” a portion of fine-grained leaf labels to basic level labels.
- Evaluate on fine-grained recognition.
- **Consistently outperforms baselines.**

relabeling	softmax-leaf	softmax-all	logistic	**ours**
50%	50.5(74.7)	56.4(79.6)	21.0(45.2)	**58.2(80.8)**
90%	26.2(47.3)	52.9(77.2)	9.3(27.2)	**55.3(79.4)**
95%	16.0(32.2)	50.8(76.0)	5.6(17.2)	**52.4(77.2)**
99%	2.5 (7.2)	**41.5**(68.1)	1.0(3.8)	**41.5(68.5)**

Top 1 accuracy (top 5 accuracy)

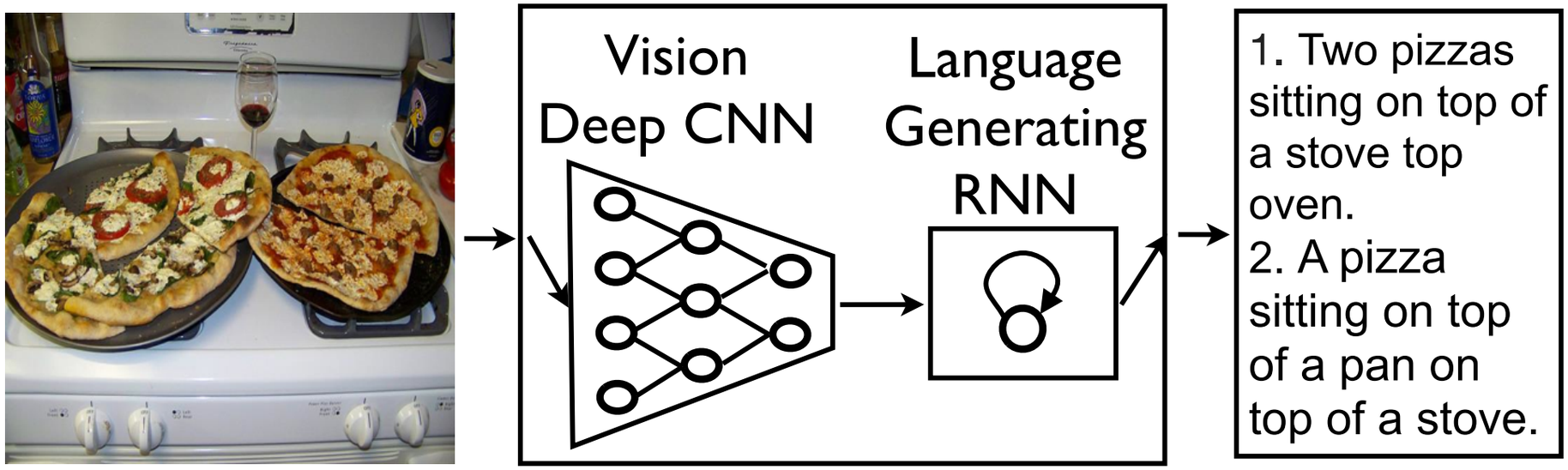
Google

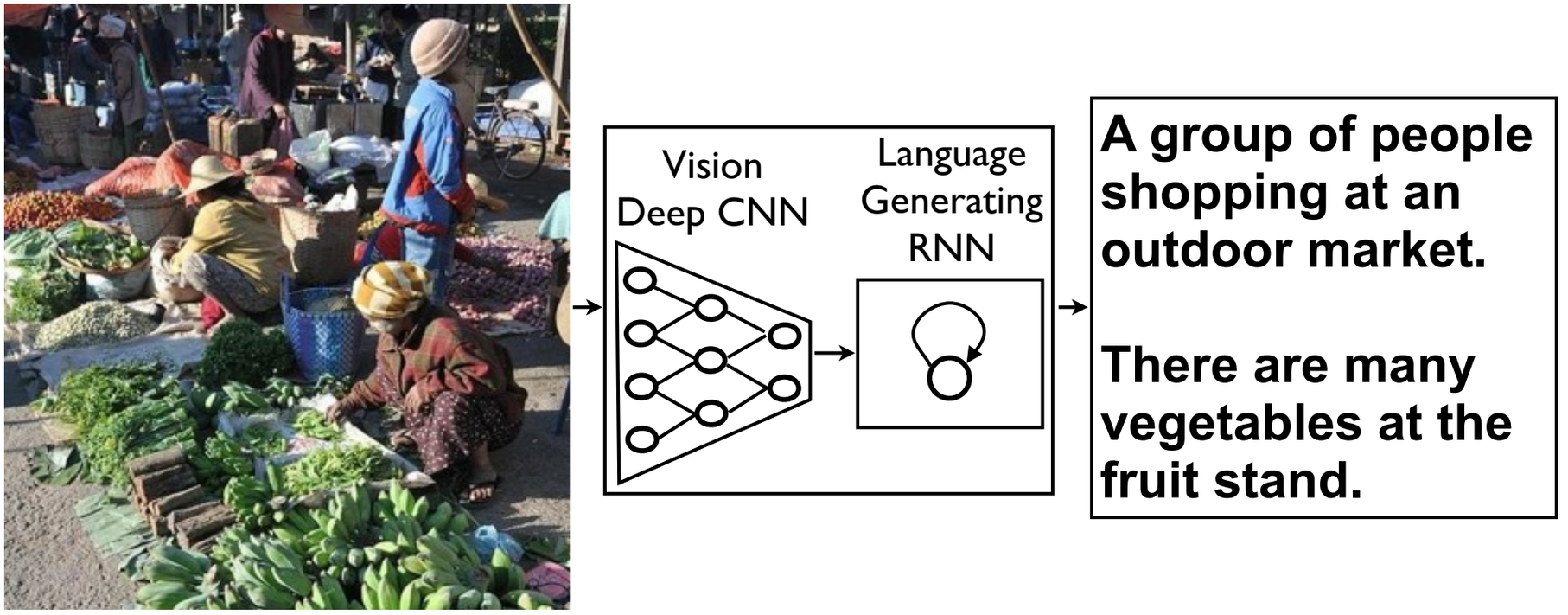
What about textual descriptions?

- We have considered the long tail of objets.
- What about more complex descriptions, involving multi-word descriptions, or captions?
- We can use language models to help.

Neural Image Caption Generator

[Vinyals et al, CVPR 2015]





Google

NIC: objective

- Let I be an image (pixels).
- Let S be the corresponding sentence (sequence of words).
- Likelihood of producing the right sentence given the image:

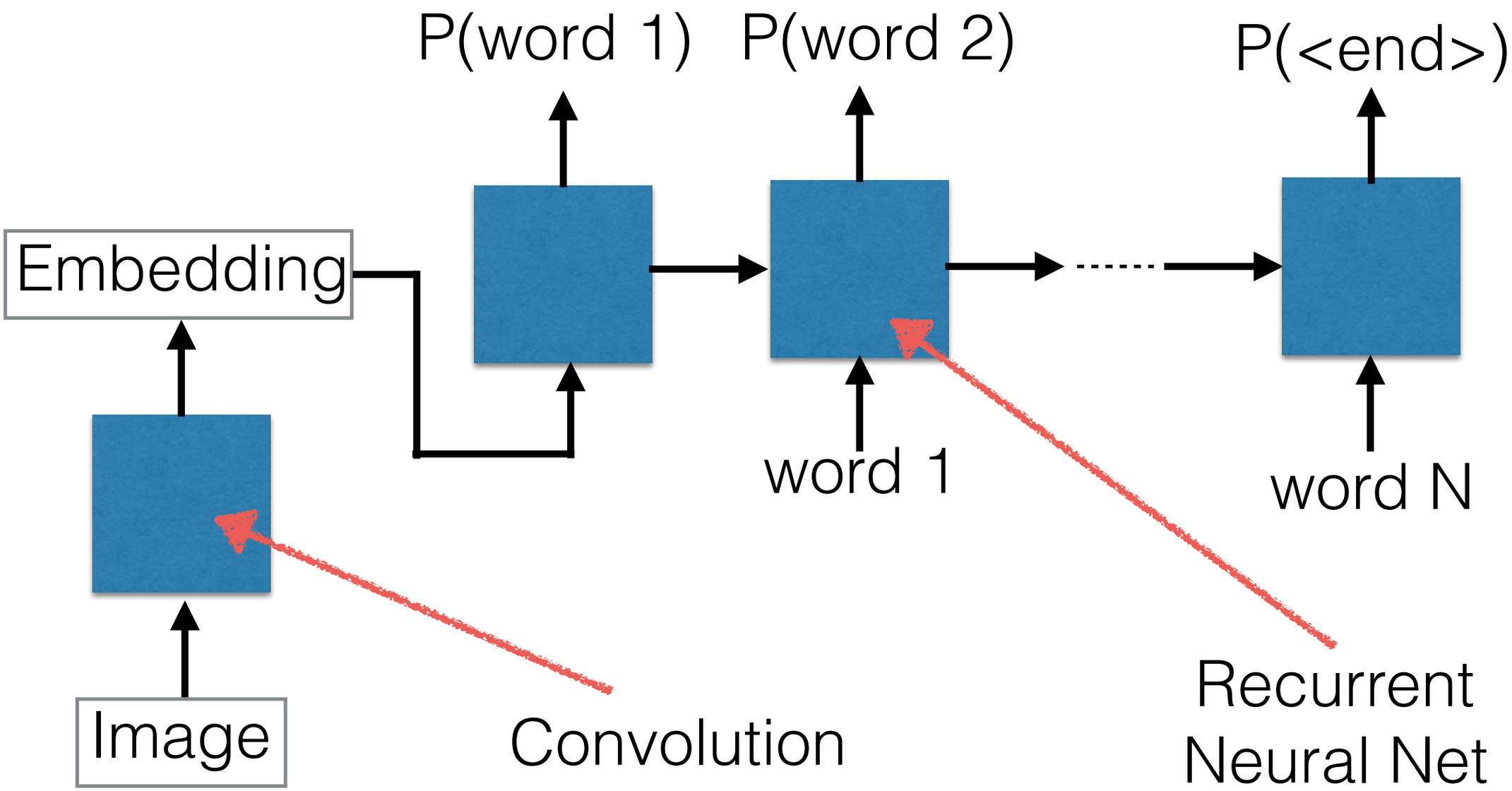
$$\log p(S|I) = \sum_{t=0}^{N} \log p(S_t|I, S_0, \ldots, S_{t-1})$$

- We maximize the likelihood of producing the right sentence given the image:

$$\theta^\star = \arg\max_\theta \sum_{(I,S)} \log p(S|I;\theta)$$

Google

NIC: model


P(word 1)
P(word 2)
P(<end>)
Embedding
word 1
word N
Image
Convolution
Neural Net
Recurrent
Neural Net

Examples

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A skateboarder does a trick on a ramp.

A dog is jumping to catch a frisbee.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A little girl in a pink hat is blowing bubbles.

A refrigerator filled with lots of food and drinks.

A herd of elephants walking across a dry grass field.

A close up of a cat laying on a couch.

A red motorcycle parked on the side of the road.

A yellow school bus parked in a parking lot.

Describes without errors	Describes with minor errors	Somewhat related to the image	Unrelated to the image

Google

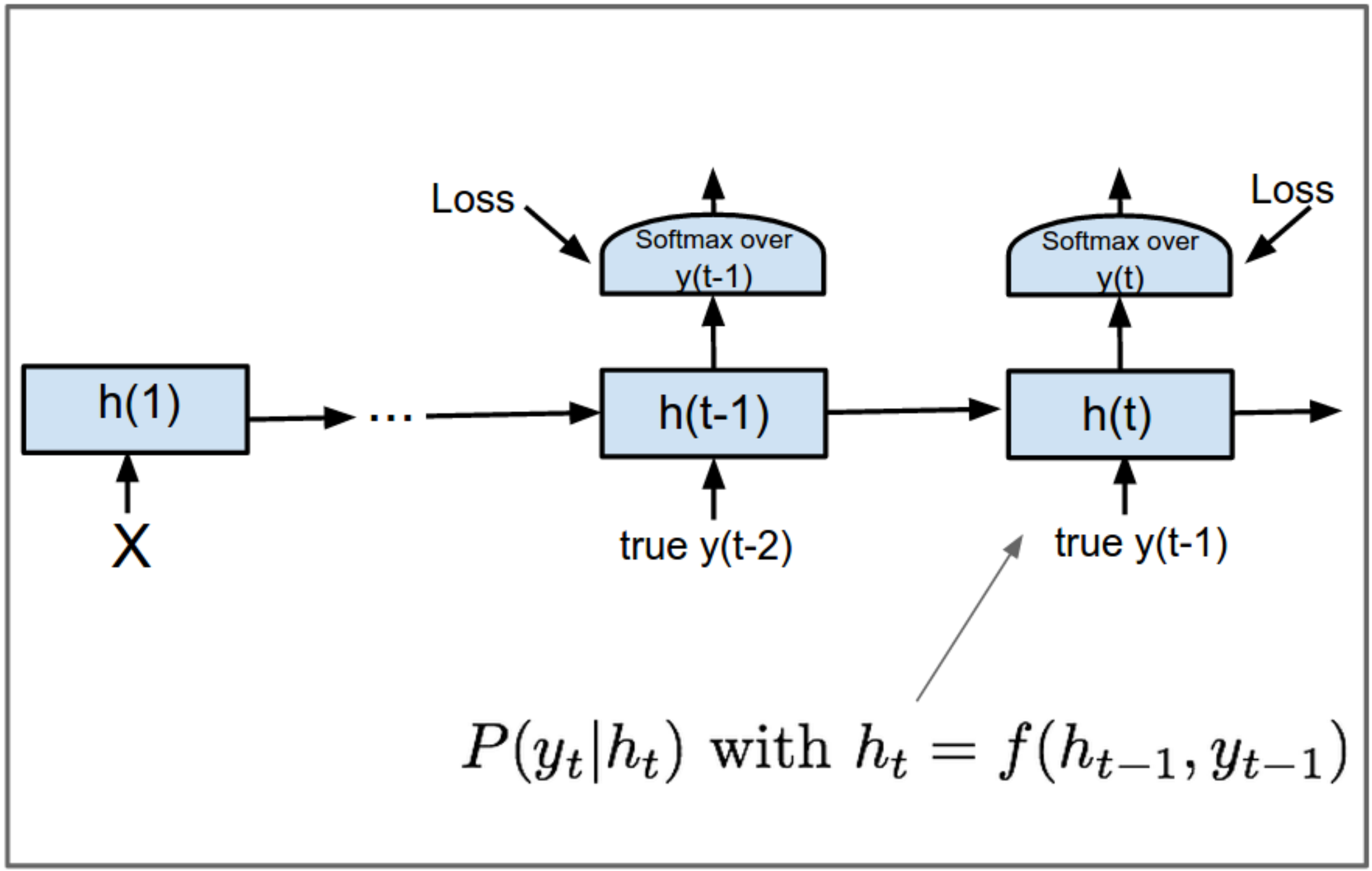
It doesn't always work…

Human: A blue and black dress ... No! I see white and gold!

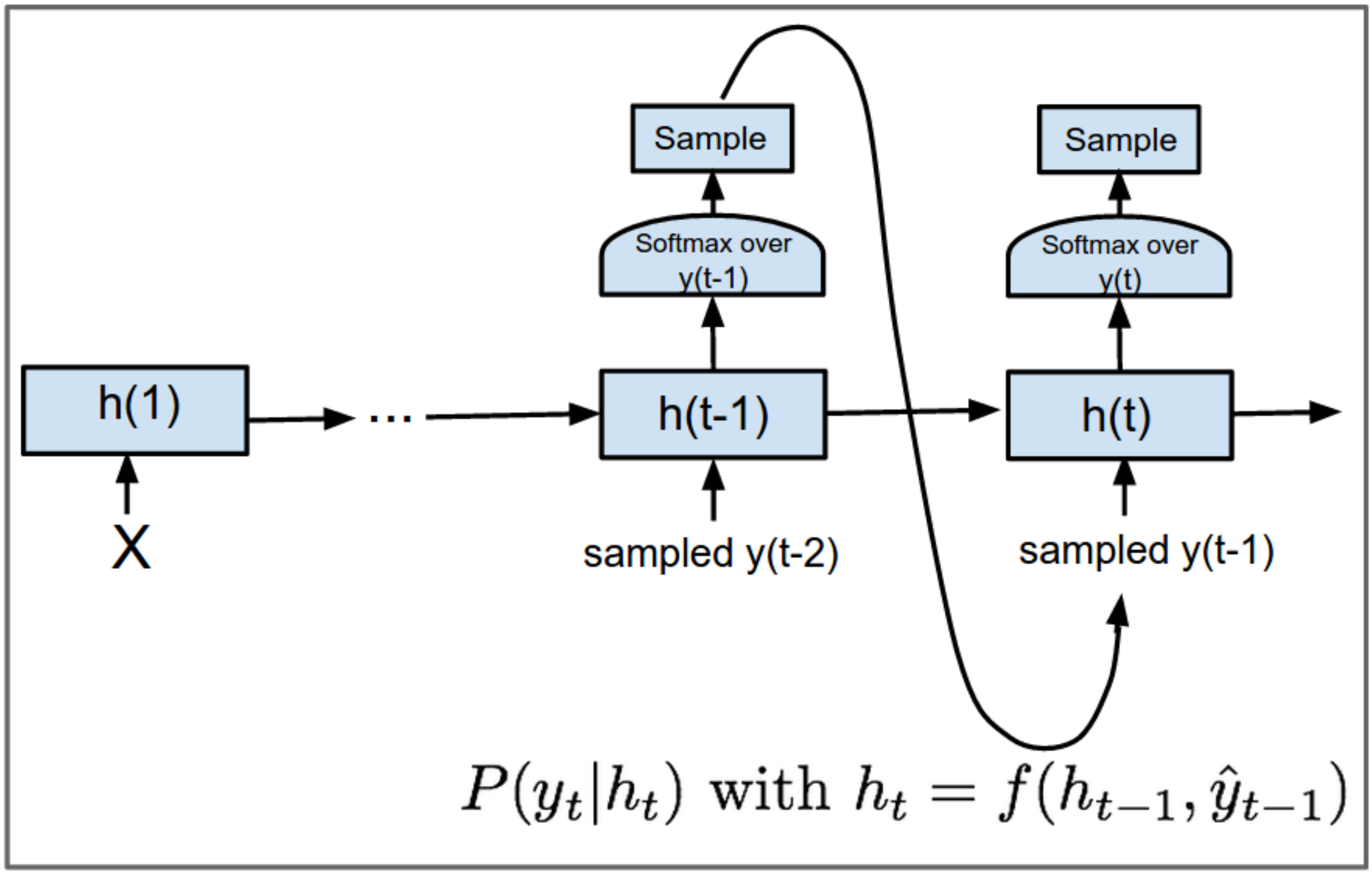
Our model: A close up of a vase with flowers.

Google

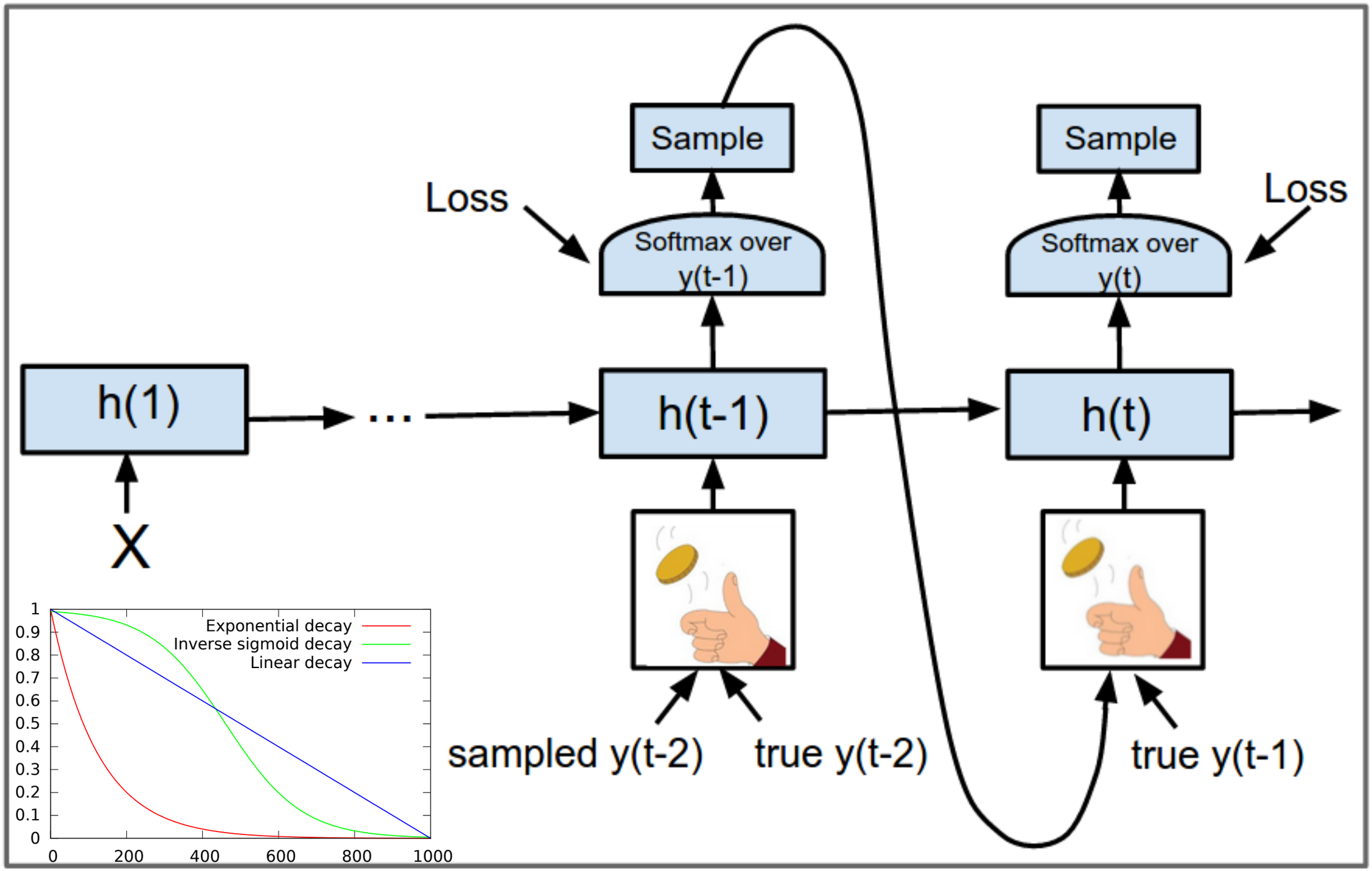
Scheduled Sampling [NIPS 2015]



Scheduled Sampling



Scheduled Sampling



Conclusions

- The long tail problem happens in most of the interesting tasks.
- Sharing approaches can help “poor” classes to generalize thanks to “rich” classes.
- At the extreme, semantic embeddings can represent classes with zero training examples.
- Sharing approaches are interesting not only for text and images, but also for complete sentences.
- Other recent approaches: represent text using characters; good for long tail words.
- … but never under-estimate the long tail …

Google

TensorFlow™

Open source machine learning

tensorflow.org

Google

Google Brain Residency Programme

New one year immersion program in deep learning research

- Learn to conduct deep learning research w/experts in our team
- Fixed one-year employment with salary, benefits, ...
- Goal after one year is to have conducted several research projects
- Interesting problems, TensorFlow, and access to computational resources
- Apply before January 15, 2016.

- For more information: g.co/brainresidency
- Contact us: brain-residency@google.com

Google