Zero-Shot Learning by Convex Combination of Semantic Embeddings

---

Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer,
Jonathon Shlens, Andrea Frome, Greg S. Corrado, Jeffrey Dean

Presented by: Stefan Hosein

Problem

Zero Shot Learning

Echidna

Aye-aye

Patagonian Mara

Gerenuk

Dugong

Problem

Zero Shot Learning

Echidna

Aye-aye

Patagonian Mara

Gerenuk

Dugong

Supervised Learning

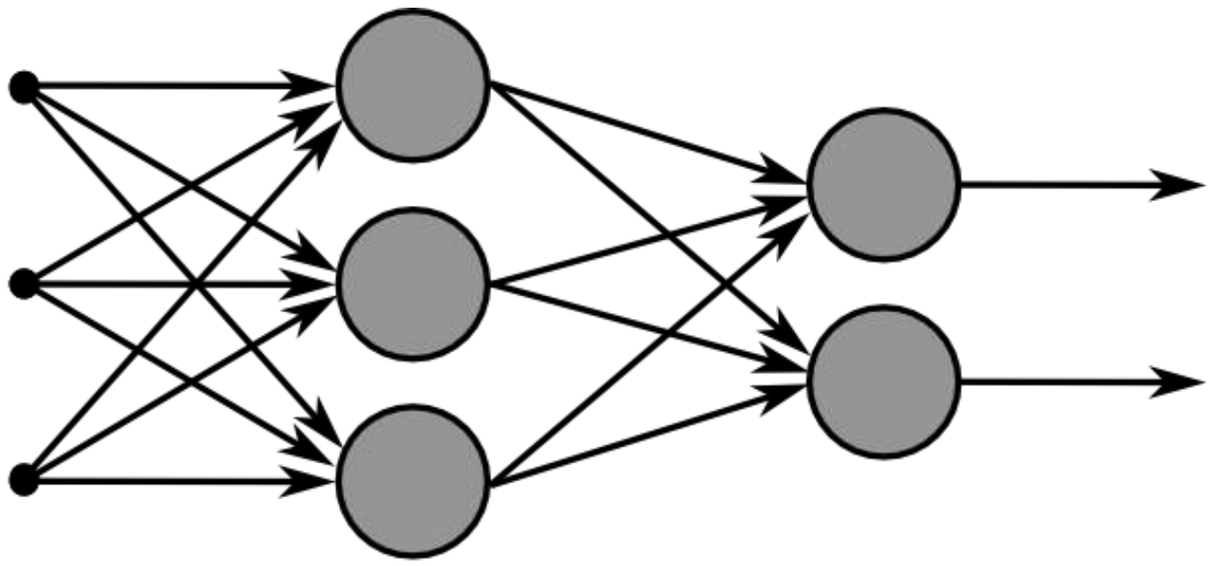
puppy

kitten

puppy

kitten

Supervised Learning



hidden layer output layer

puppy

Zero-Shot Learning

Generalize to Unseen Images

Training Data:

kitten

puppy

rabbit

Test Data:

deer

panda

Zero-Shot Learning: Supervised Attributes

- Amphibian, four legs, two eyes, moist and smooth skin, **not** dry and bumpy skin **-> TOAD [1,1,1,1,0]**

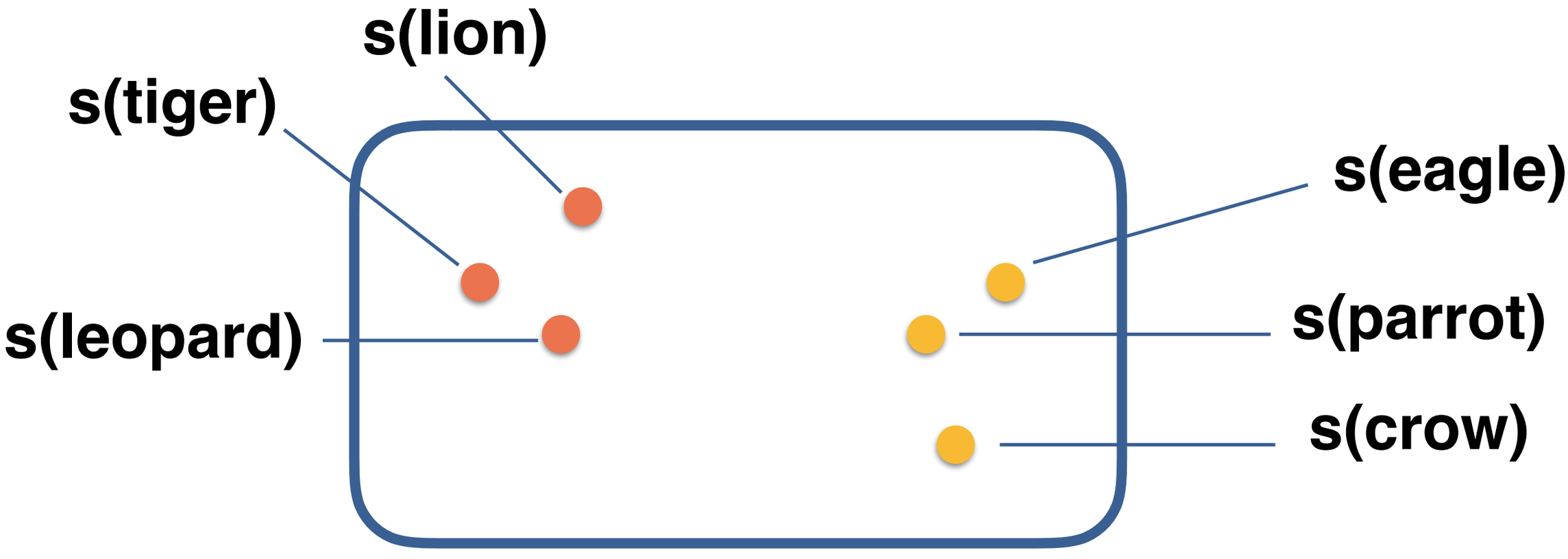
- Amphibian, four legs, two eyes, **not** moist and smooth, dry and bumpy skin **-> FROG [1,1,1,0,1]**

Zero-Shot Learning: Unsupervised Embeddings*

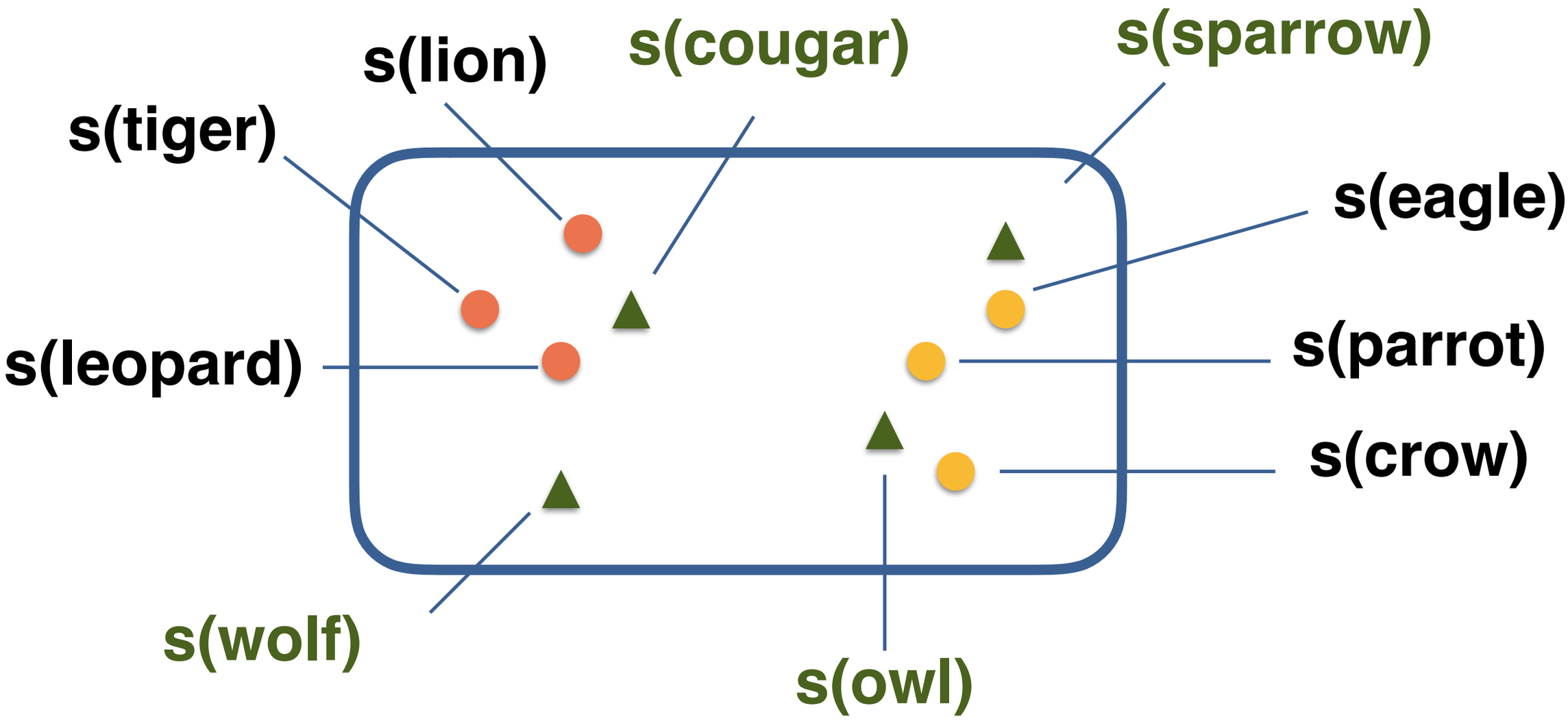
- Eagle, Parrot, Crow **: Owl**
- Lion, Tiger, Leopard **: Cougar**
- Shark, Dolphin, Whale **: Manatee**

*Skip-gram on Wikipedia

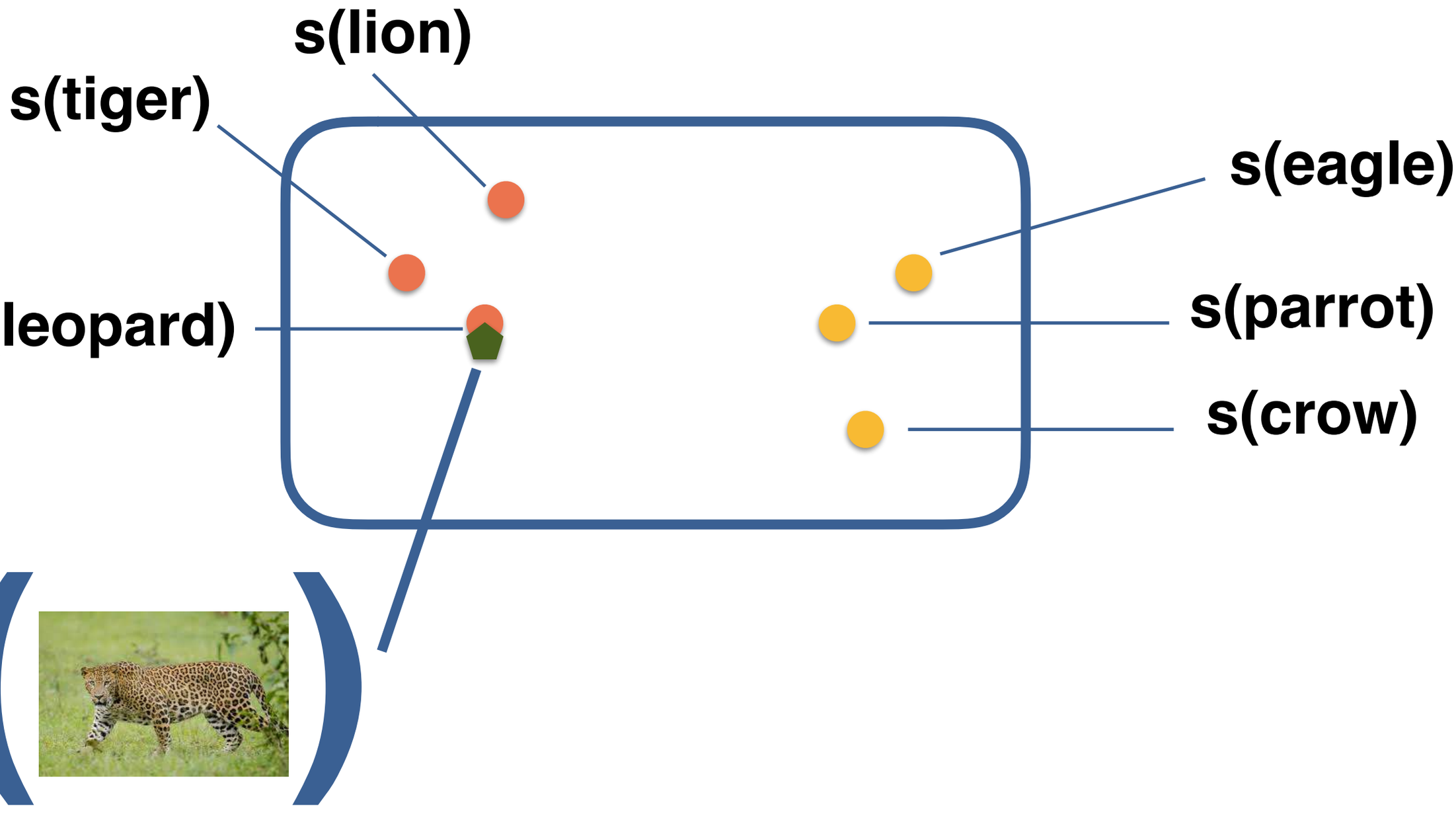
Embedding Labels



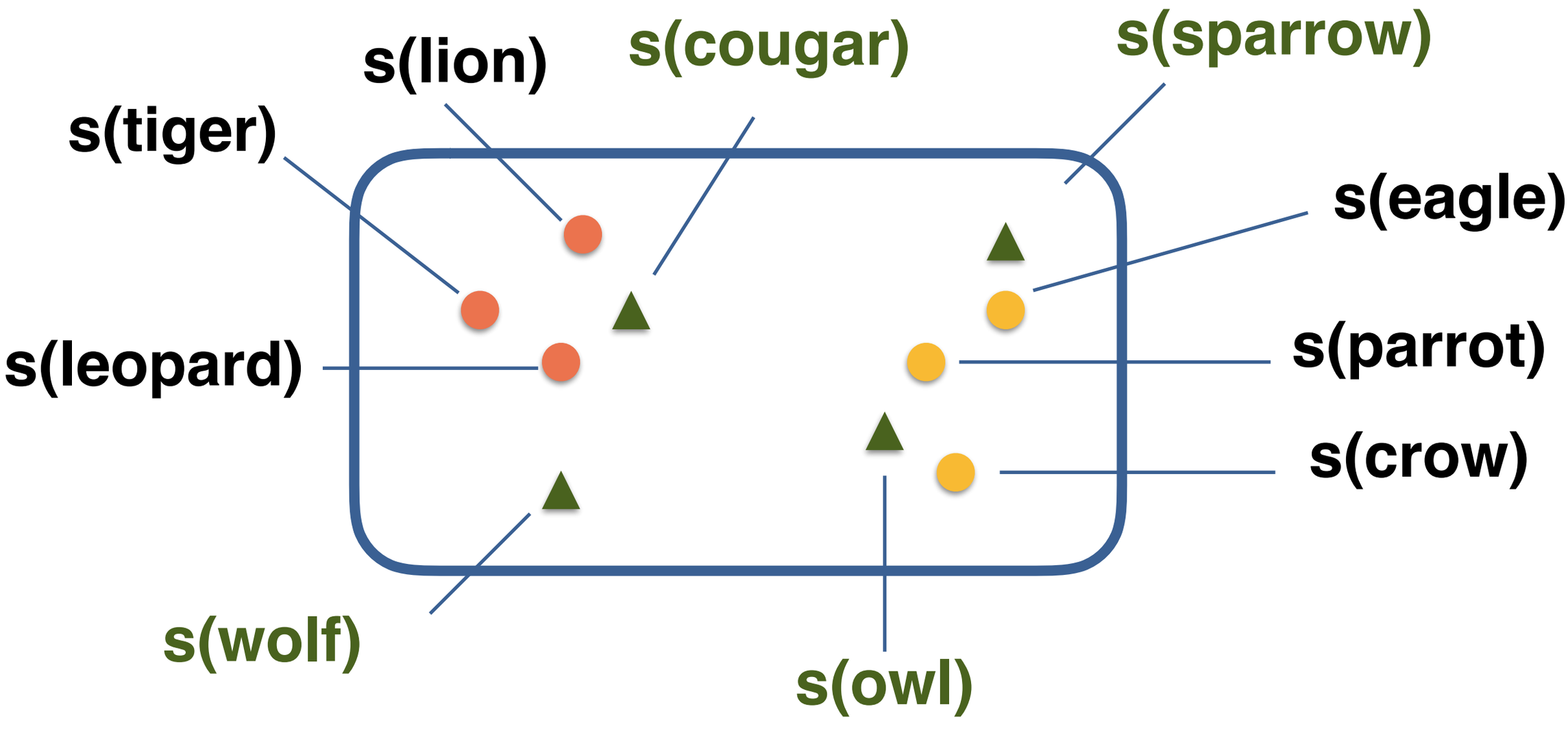
Embedding Labels



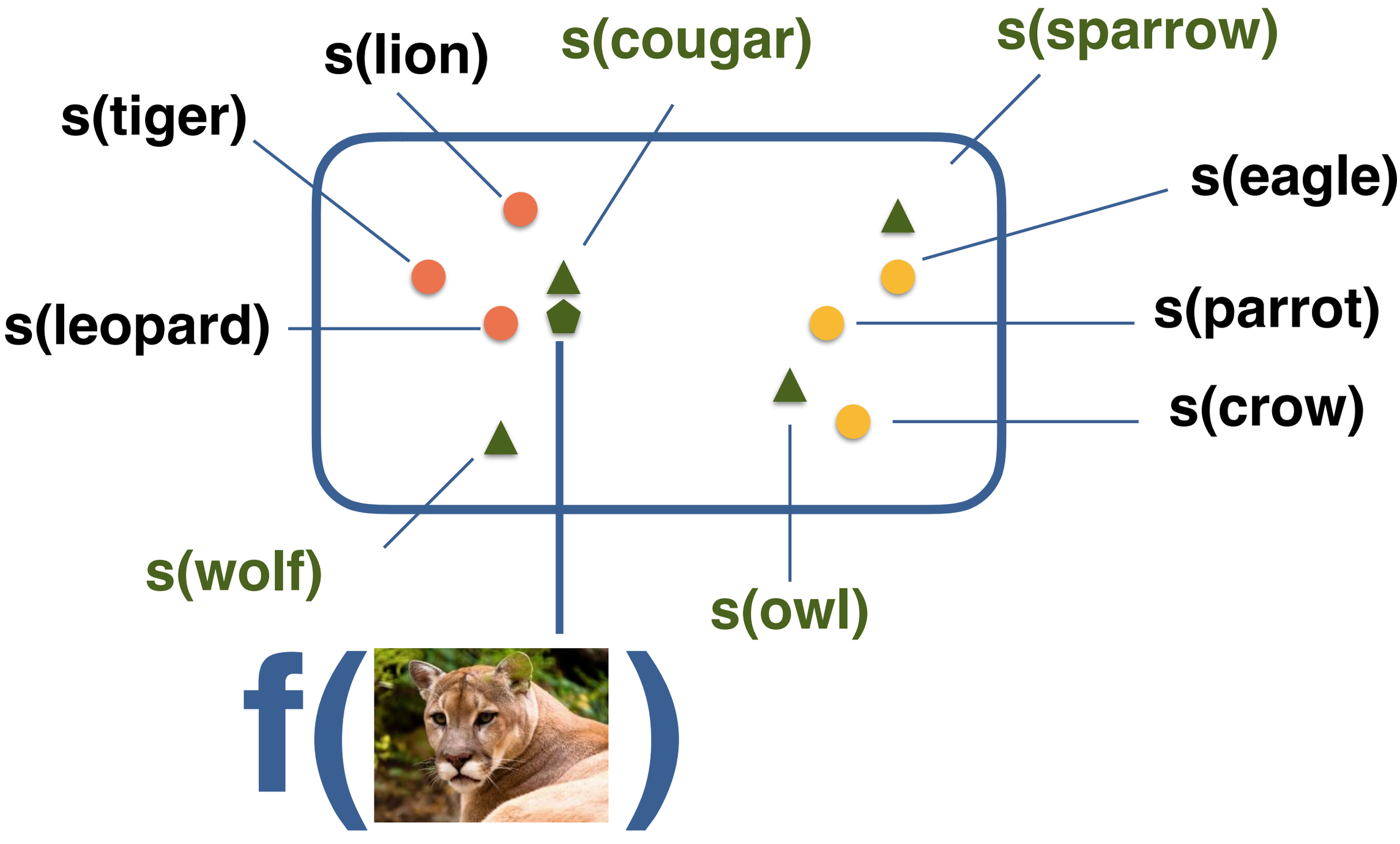
Embedding Images



Embedding Images



Embedding Images



Convex Combination of Semantic Embeddings (ConSE)

$$\widehat{y}_0(\mathbf{x}, 1) \equiv \underset{y \in \mathcal{Y}_0}{\operatorname{argmax}}\, p_0(y \mid \mathbf{x})$$

best training label for an image **x**

image

label

classifier

Training

ConSE(T) - Top T Predictions

$$f(\mathbf{x}) = \frac{1}{Z} \sum_{t=1}^{T} p(\widehat{y}_0(\mathbf{x}, t) \mid \mathbf{x}) \cdot s(\widehat{y}_0(\mathbf{x}, t))$$

input **x**

weighted
by probabilities

semantic
embeddings

Training

ConSE

$$\widehat{y}_1(\mathbf{x}, 1) \equiv \underset{y' \in \mathcal{Y}_1}{\operatorname{argmax}}\, cos(f(\mathbf{x}), s(y'))$$

top prediction of image

cosine similarity

test label set

Testing

DeVise(baseline)

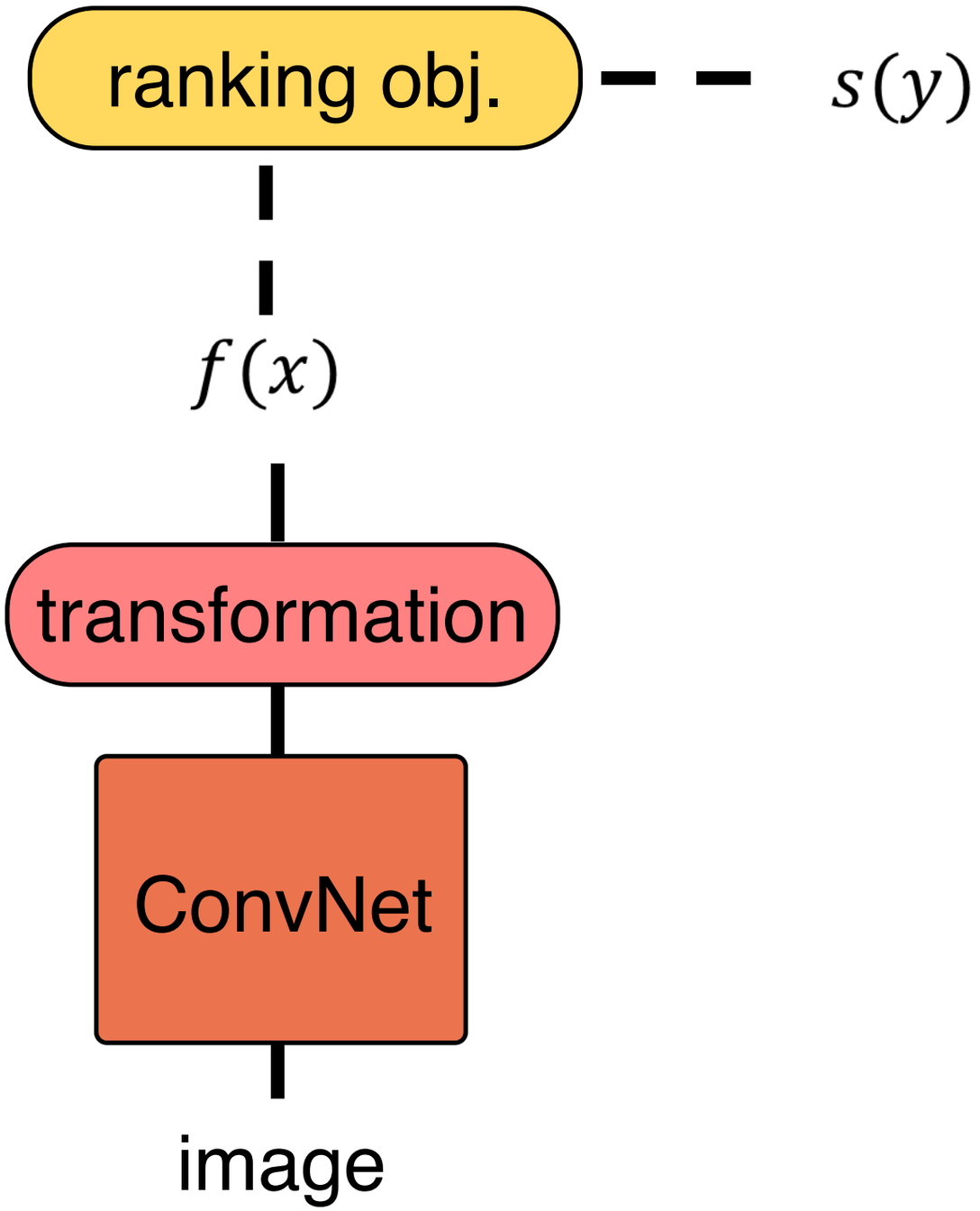
- Train a ranking model on the training set so that $f(x;\theta)$ is closer to $s(y)$ than $s(y^-)$

  Triplet ranking hinge loss

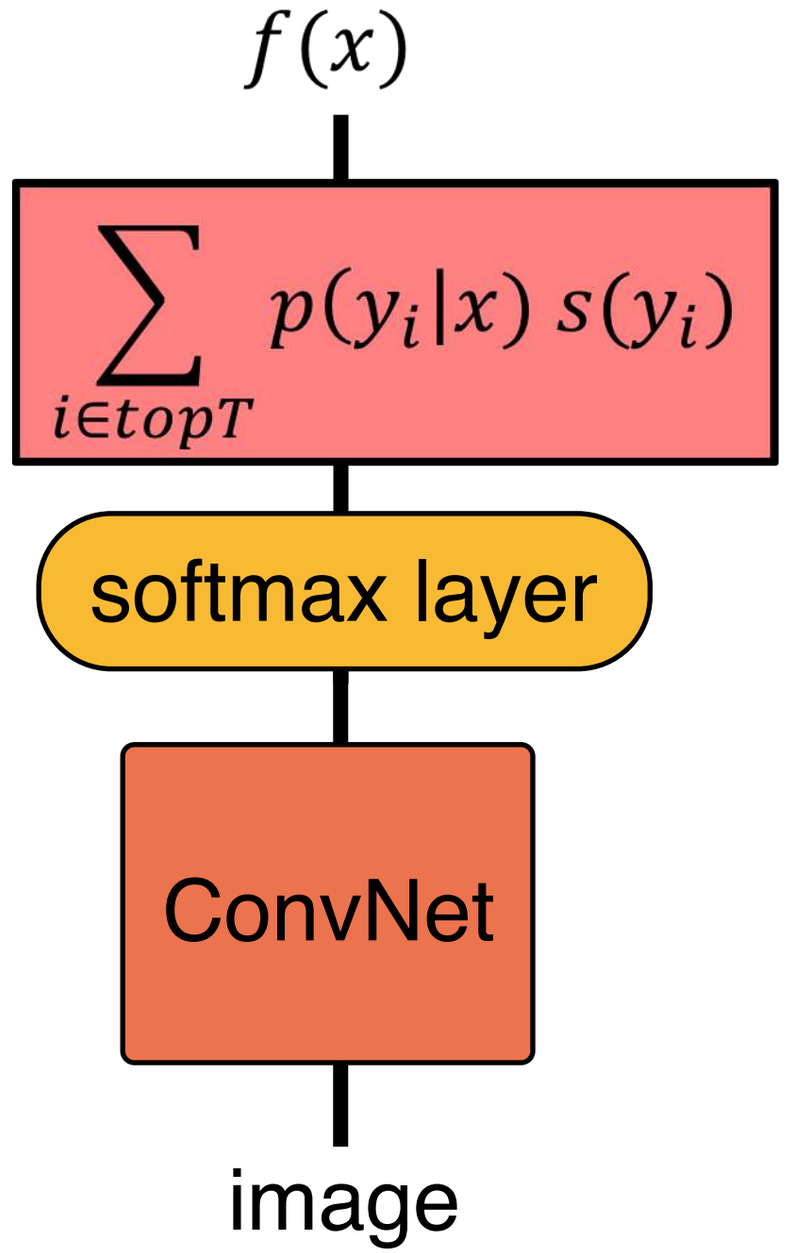
$$\sum_{x,y,y^-} [\|f(x;\theta) - s(y)\| - \|f(x;\theta) - s(y^-)\|]_+$$

[Andrea Frome, Greg S. Corrado, Jon Shlens *et al.* DeViSE: A Deep Visual-Semantic Embedding Model, 13]

Norouzi et.al. presentation



DeViSE



ConSE(T)

Norouzi et.al. presentation

Dataset

Training:

- ImageNet **1000** labels

Testing (no overlap with training labels):

- **2-hops** (visually and semantically similar to training) - 1,549 labels
- **3-hops** - 7,860 labels
- **All** - 20,184 labels

Flat hit@k Performance

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

Flat hit@k Performance

Data Set	#Candidate Labels	Model	Flat hit@k (%)				
			1	2	5	10	20
2-hop	1,549	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-hop **(+1K)**	1,549 +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

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	5.5
		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

Hierarchical precision@k Performance

Test Label Set	Model	Hierarchical precision@k				
		1	2	5	10	20
2-hops	DeViSE	0.06	0.152	0.192	0.217	0.233
	ConSE(10)	**0.094**	**0.214**	**0.247**	**0.269**	**0.284**
2-hops (+1K)	Softmax baseline	0	**0.236**	0.181	0.174	0.179
	DeViSE	**0.008**	0.204	0.196	0.201	0.214
	ConSE(10)	0.003	0.234	**0.254**	**0.260**	**0.271**
3-hops	DeViSE	0.017	0.037	0.191	0.214	0.236
	ConSE(10)	**0.027**	**0.053**	**0.202**	**0.224**	**0.247**
3-hops (+1K)	Softmax baseline	0	0.053	0.157	0.143	0.130
	DeViSE	**0.005**	0.053	0.192	0.201	0.214
	ConSE(10)	0.002	**0.061**	**0.211**	**0.225**	**0.240**
ImageNet 2011 21K	DeViSE	0.008	0.017	0.072	0.085	0.096
	ConSE(10)	**0.014**	**0.025**	**0.078**	**0.092**	**0.104**
ImageNet 2011 21K (+1K)	Softmax baseline	0	0.023	0.071	0.069	0.065
	DeViSE	**0.003**	0.025	0.083	0.092	0.101
	ConSE(10)	0.002	**0.029**	**0.086**	**0.097**	**0.105**

Test Image	ConvNet	DeViSE	ConSE(10)
 **(dress, frock)**	• wig • fur coat • Saluki • Afghan hound	• water spaniel • tea gown • bridal gown • spaniel • tights, leotards	• business suit • **dress, frock** • hairpiece, false hair • swimsuit • kit, outfit
 **(flightless bird)**	• Ostrich • black stork • vulture • crane • peacock	• heron • owl • hawk • raptor • finch	• **flightless bird, ratite** • Peafowl • common spoonbill • New World vulture • Greek partridge

Norouzi et.al. presentation

Test Image	ConvNet	DeViSE	ConSE(10)
 **(farm machine)**	• thresher • tractor • harvester • half-track • snowplow	• truck • skidder • tank car • automatic rifle • house trailer	• flatcar • truck • tracked vehicle • bulldozer • wheeled vehicle
 **(Lama pacos)**	• Tibetan mastiff • titi monkey • Koala • llama • chow-chow	• kernel • littoral zone • carillon • Cabernet • poodle dog	• domestic dog • domestic cat • schnauzer • Belgian sheepdog • domestic llama

Norouzi et.al. presentation

Comments

Pros:

- Showed significant improvement over state-of-the-art
- Built a zero-shot learning algorithm without regression
- Clear and self-contained paper

Cons:

- No running time
- Was it necessary to have experiments with labelled data