DeViSE: A Deep Visual-Semantic Embedding Model

dukim@dongguk.edu
Daeung Kim

Index

Introduction : Insufficient labeled training data

Koala?

Yes

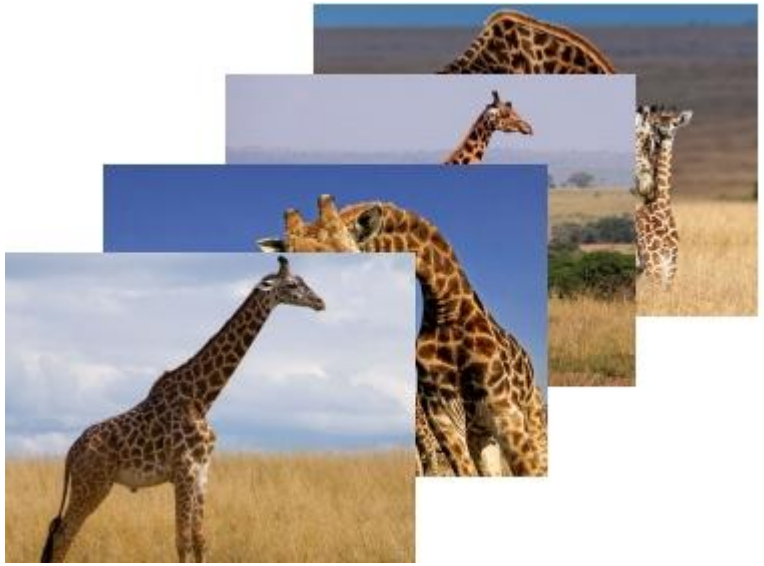
Cat?

Yes

Giraffe?

?

Introduction



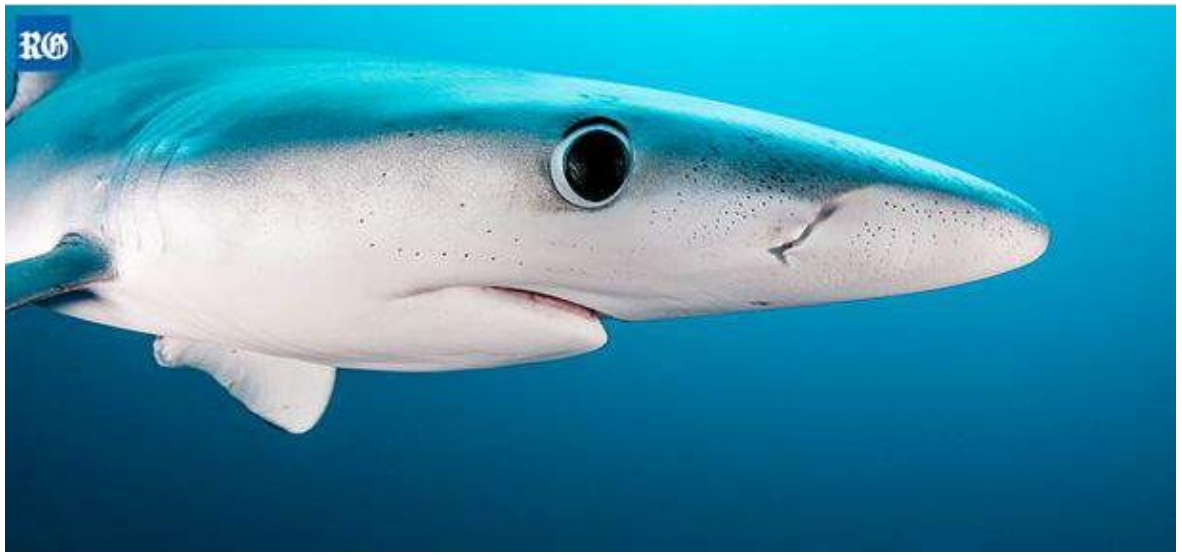
Collect more giraffe data

Horse?

Introduction : Blurred distinction between classes

White shark

Blue shark



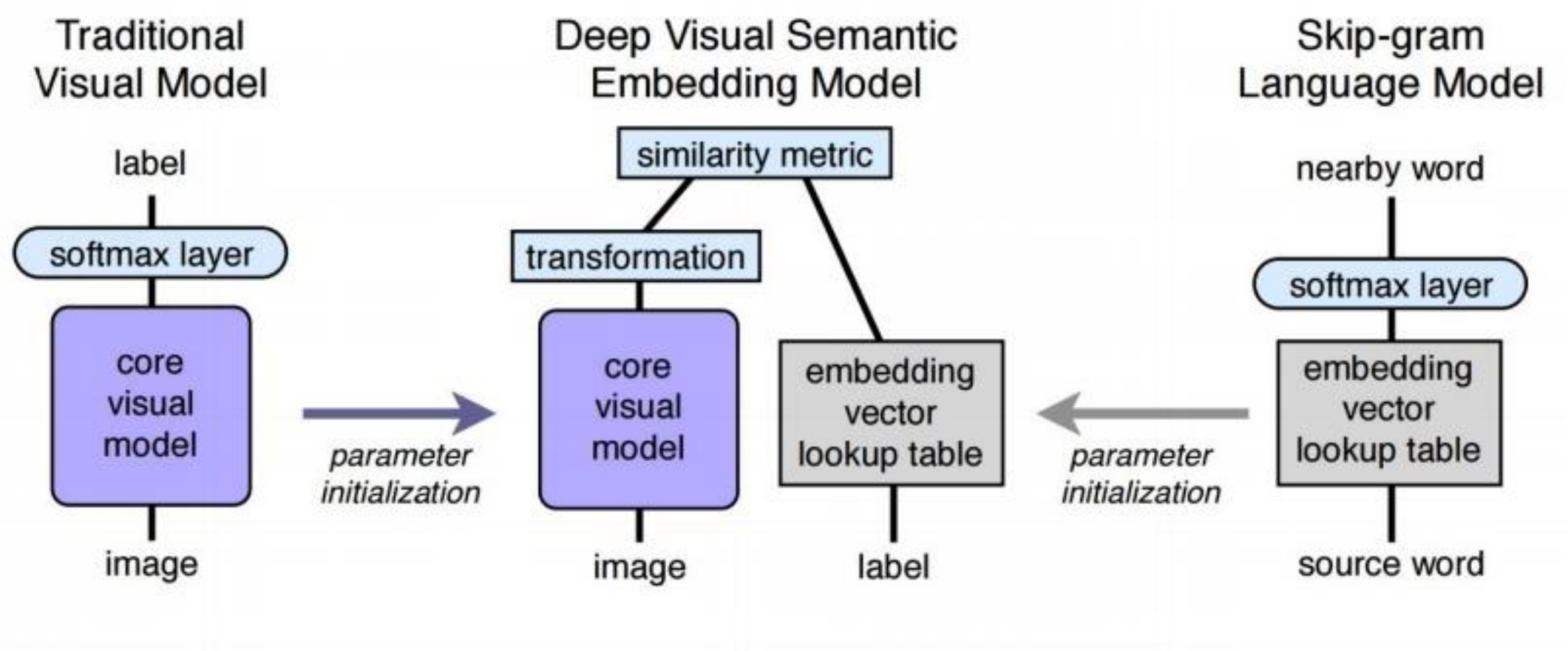
Introduction

How do we improve predictions of unknown categories?

Introduction : Related work

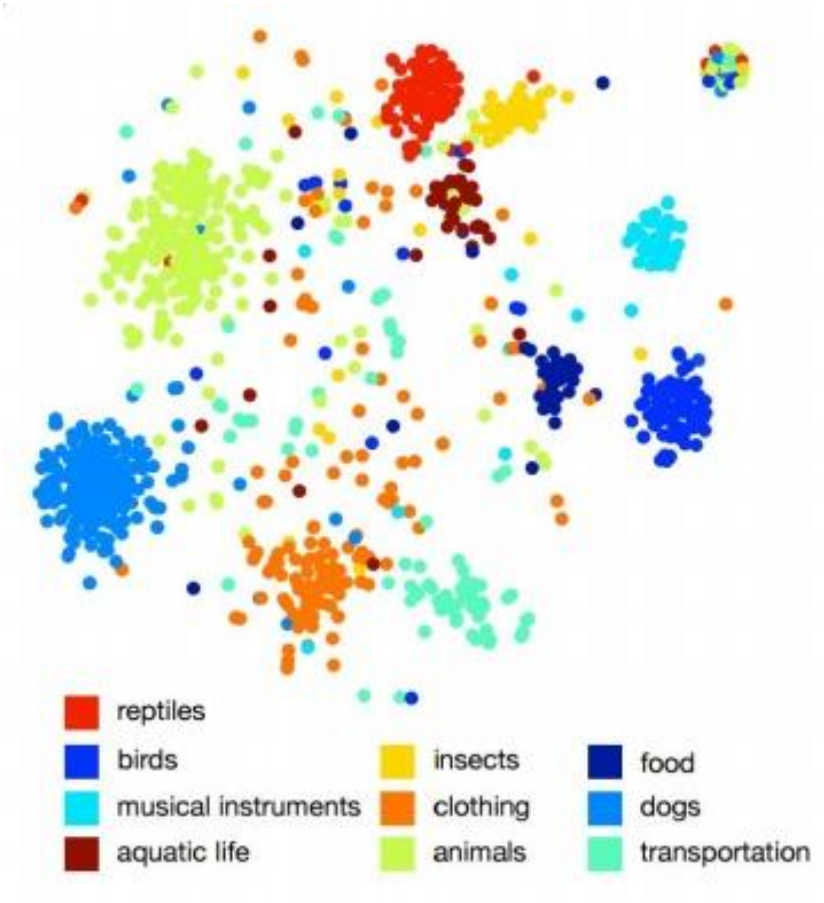
- WASABIE : Linear map from image features to embedding space. Only used training labels.
- Socher et al : Linear map from image features to embedding space. Only 8 known and 2 unknown classes.
- Other work that has shown zero-shot classification relies on curated information.

Proposed Method



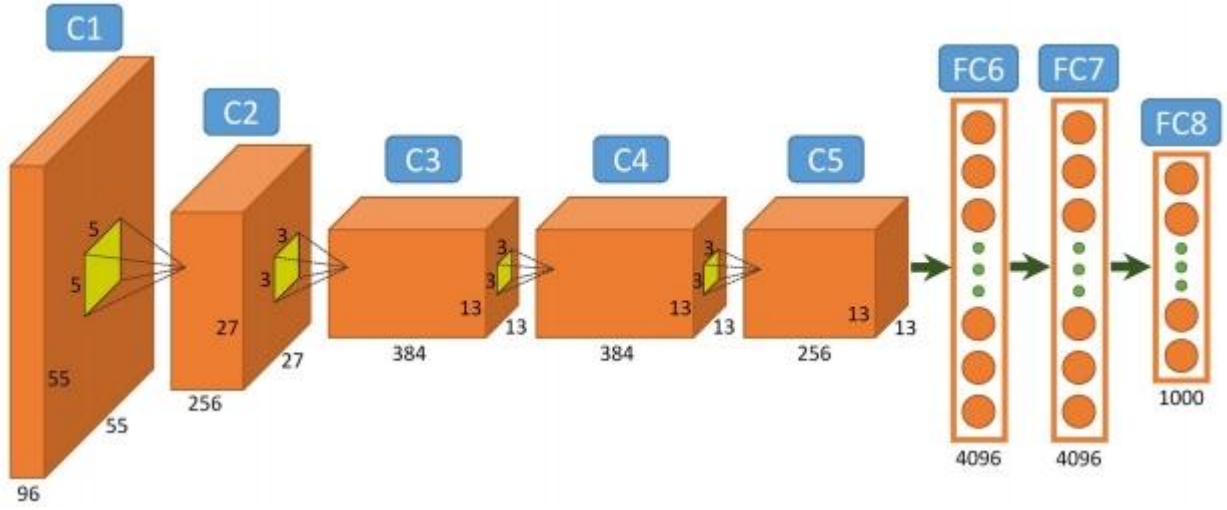
Proposed Method : Skip-gram LM

- 5.7M document of wikipedia
- Hierarchical softmax
- Window size : 20
- Embedding dim : 500, 1000



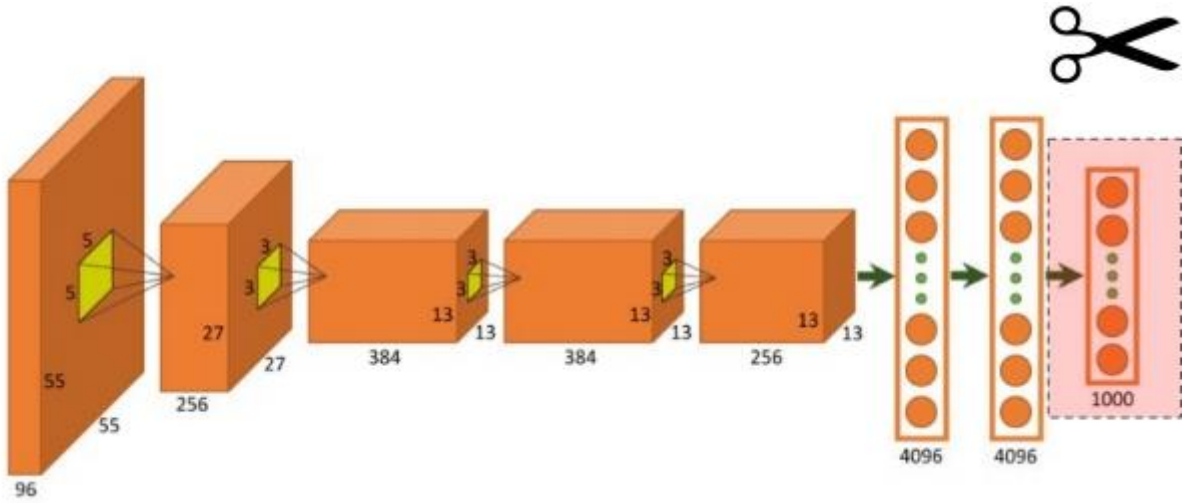
Proposed Method : Traditional Visual Model

- AlexNet
- 1,000 class from ILSVRC 2012 1K dataset
- Softmax baseline & initialization of DeViSE's visual model



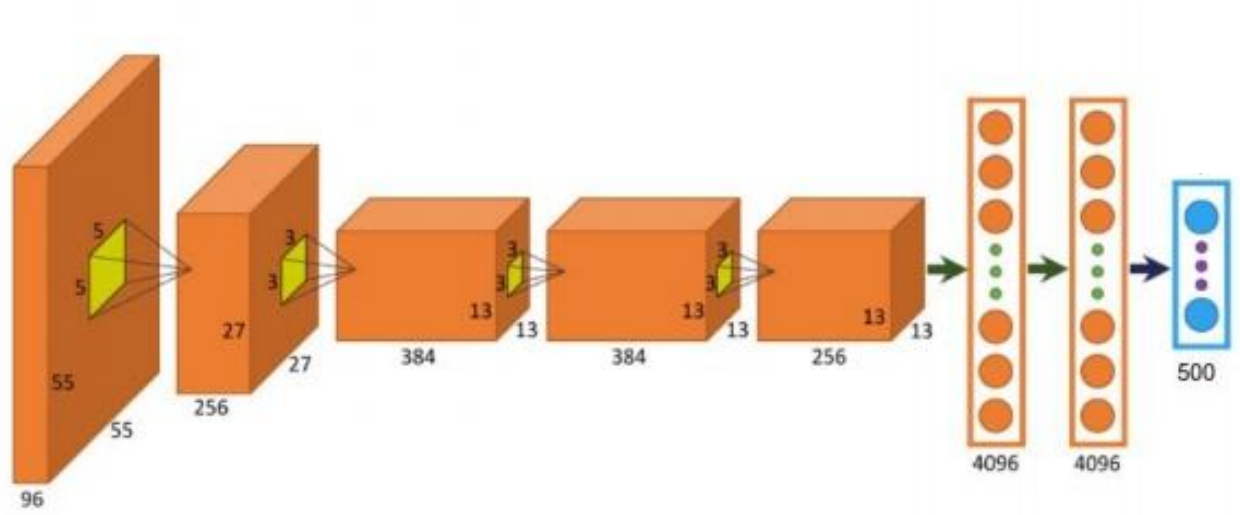
Proposed Method : Combined Model

- Abandon softmax layer, and add projection layer.
- linear transformation 4096-D to 500-D or 1000-D

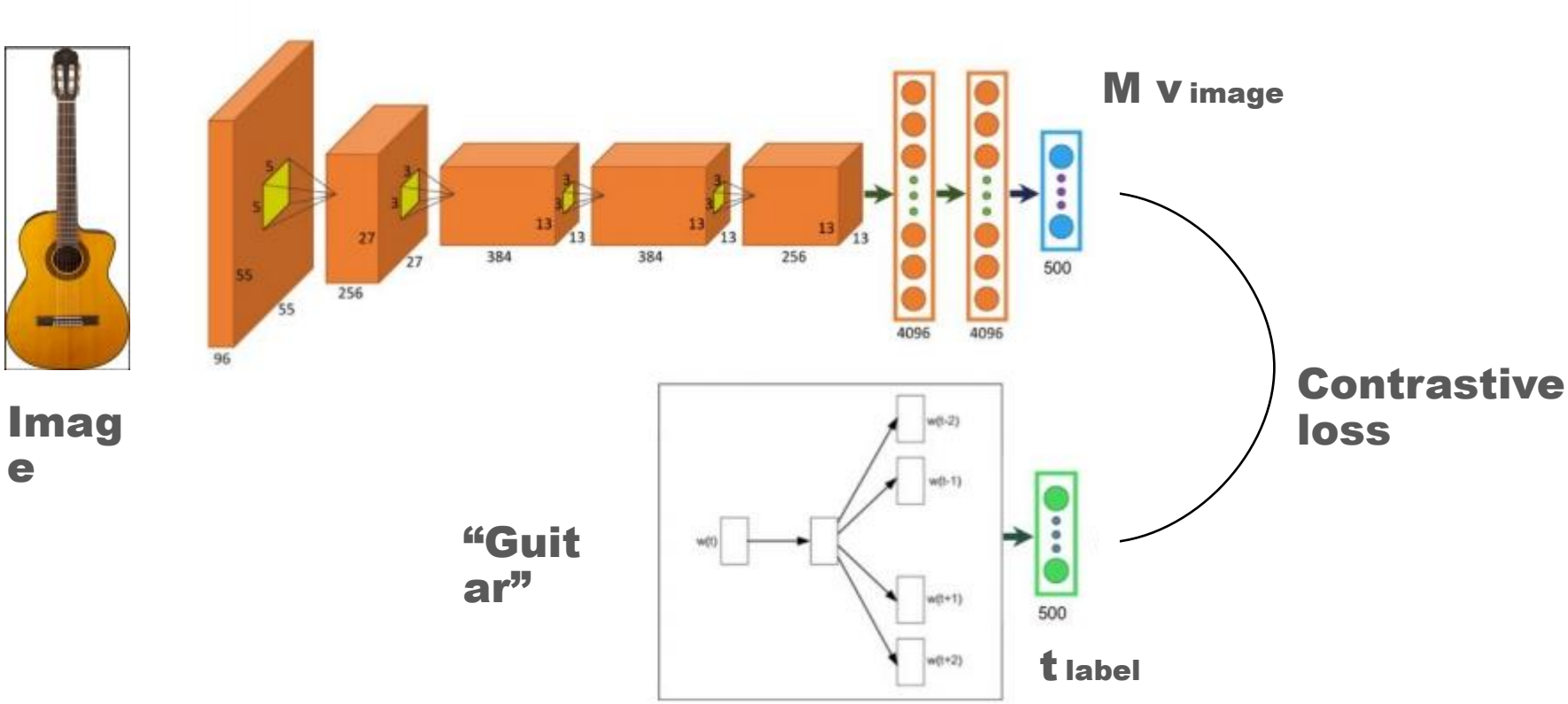


Proposed Method : Combined Model

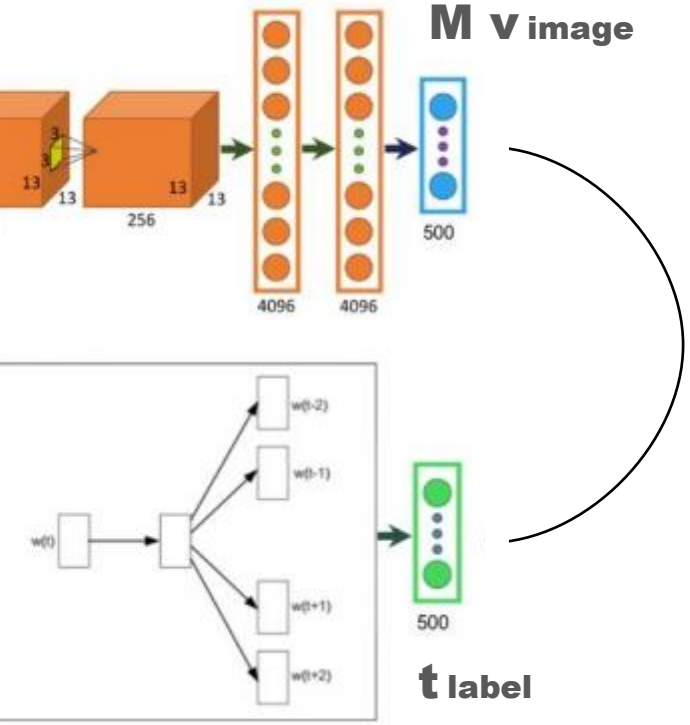
- abandon softmax layer, and add projection layer.
- linear transformation 4096-D to 500-D or 1000-D



Proposed Method : Combined model



Proposed Method : Combined model



Contrastive loss

$$loss(image, label) = max\begin{cases} 0 \\ margin\ -\vec{t}_{label}M\vec{v}(image) + \vec{t}_j M\vec{v}(image) \end{cases}$$

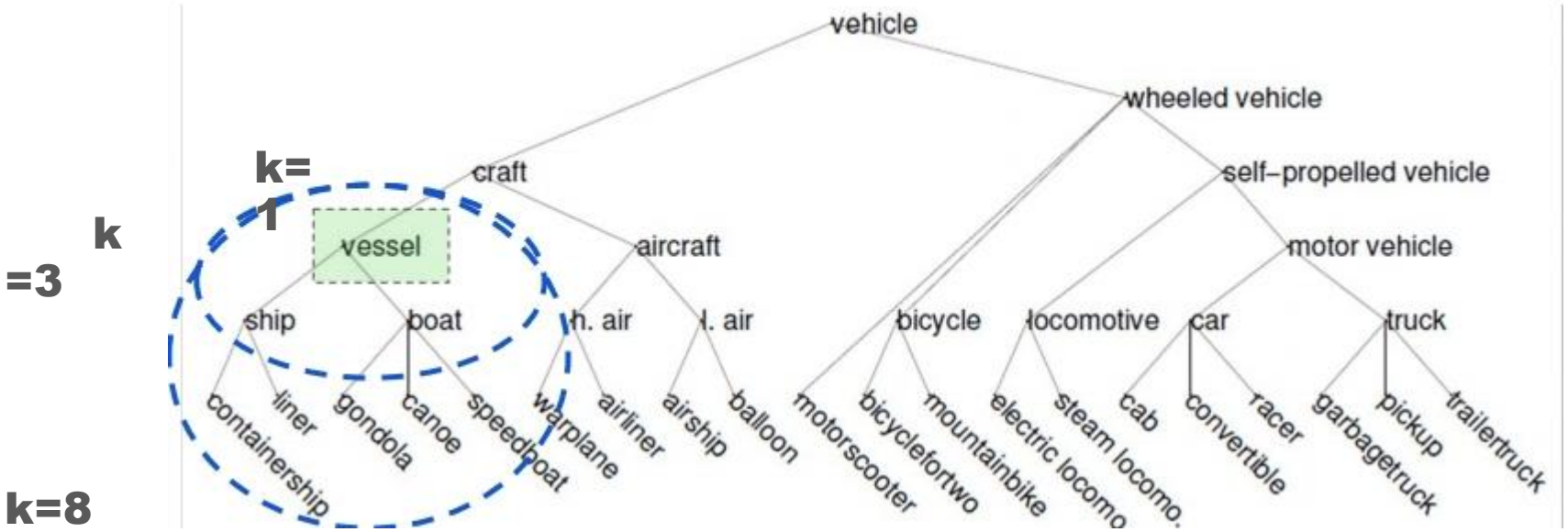
$\vec{t}_{label}$: unit-normalized (Z-distribution)
Margin : 0.1

Results

1. Evaluation Metrics
2. ImageNet (ILSVRC) 2012 1K Results
3. Generalization and Zero-Shot Learning

Results : Evaluation Metrics

- Flat hit@k
- Hierarchical precision @ k



Results : Evaluation Metrics

- Cf. precision @ k
  - Proportion of top-K document that are relevant
  - Assume 20 docs (red is relevant docs)

K	P@K	K	P@K
1	(1/1)=1.0	6	(5/6)=0.83
2	(1/2)=0.5	7	(6/7)=0.86
3	(2/3)=0.67	8	(6/8)=0.75
4	(3/4)=0.75	9	(7/9)=0.78
5	(4/5)=0.80	10	(7/10)=0.70

Results : ImageNet Results

Model type	dim	Flat hit@k (%)				Hierarchical precision@k			
		1	2	5	10	2	5	10	20
Softmax baseline	N/A	**55.6**	**67.4**	**78.5**	**85.0**	0.452	0.342	0.313	0.319
DeViSE	500	53.2	65.2	76.7	83.3	0.447	**0.352**	**0.331**	**0.341**
	1000	54.9	66.9	78.4	**85.0**	**0.454**	0.351	0.325	0.331
Random embeddings	500	52.4	63.9	74.8	80.6	0.428	0.315	0.271	0.248
	1000	50.5	62.2	74.2	81.5	0.418	0.318	0.290	0.292
Chance	N/A	0.1	0.2	0.5	1.0	0.007	0.013	0.022	0.042

- DeViSE gets pretty close with a Softmax baseline.

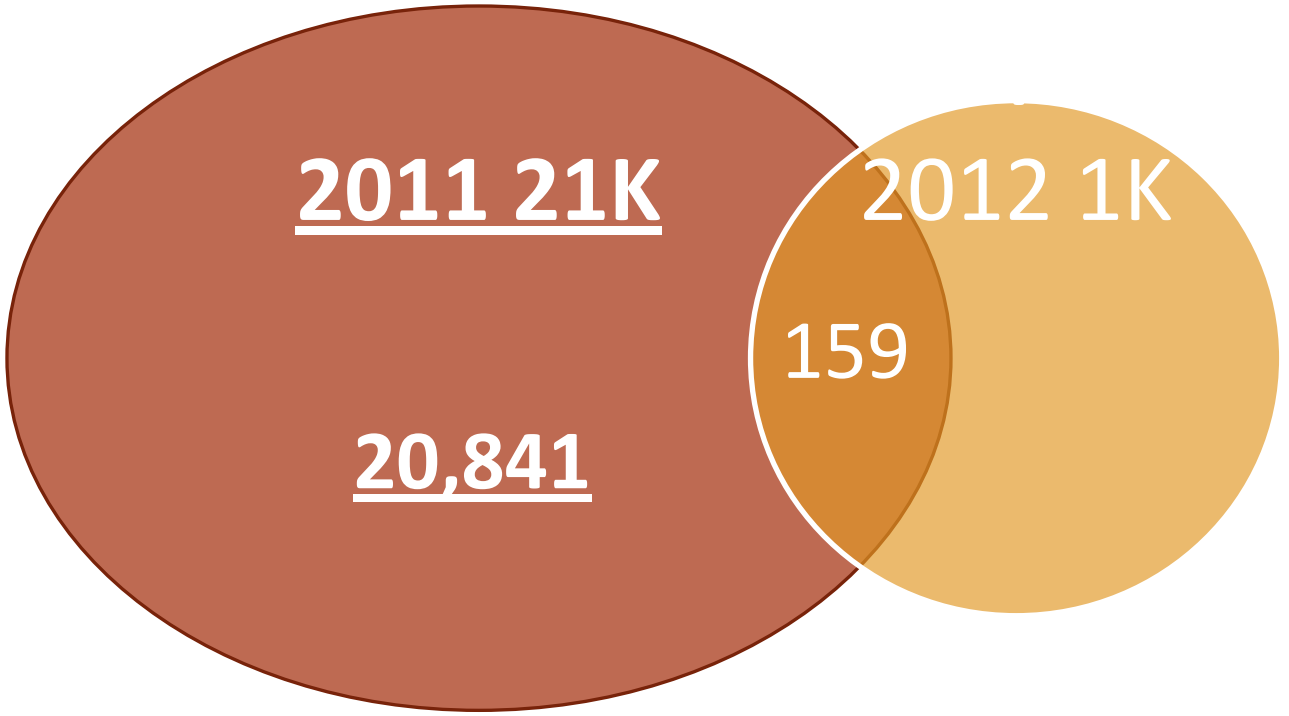
Results : ImageNet Results

Model type	dim	Flat hit@k (%)				Hierarchical precision@k			
		1	2	5	10	2	5	10	20
Softmax baseline	N/A	**55.6**	**67.4**	**78.5**	**85.0**	0.452	0.342	0.313	0.319
DeViSE	500	53.2	65.2	76.7	83.3	0.447	**0.352**	**0.331**	**0.341**
	1000	54.9	66.9	78.4	**85.0**	**0.454**	0.351	0.325	0.331
Random embeddings	500	52.4	63.9	74.8	80.6	0.428	0.315	0.271	0.248
	1000	50.5	62.2	74.2	81.5	0.418	0.318	0.290	0.292
Chance	N/A	0.1	0.2	0.5	1.0	0.007	0.013	0.022	0.042

- the gap of Hp@k between softmax baseline and DeViSE model reflects the benefit of semantic information

Results : Generalization & ZSL

- Dataset



Results : Generalization & ZSL

- Dataset

A

Our model	Softmax over ImageNet 1K
eyepiece, ocular	typewriter keyboard
Polaroid	tape player
compound lens	reflex camera
telephoto lens, zoom lens	CD player
rangefinder, range finder	space bar

D

fruit	pineapple, ananas
pineapple	coral fungus
pineapple plant, Ananas ..	artichoke, globe artichoke
sweet orange	sea anemone, anemone
sweet orange tree, ...	cardoon

Results : Generalization & ZSL

- Dataset

E

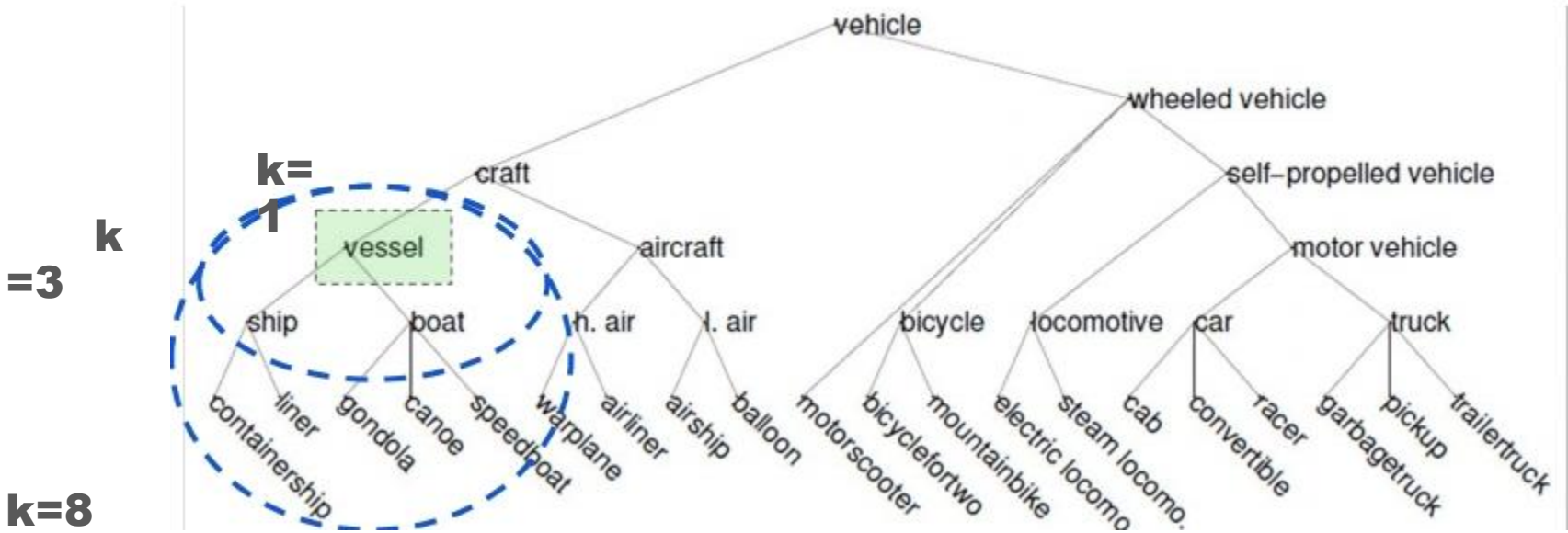
Our model	Softmax over ImageNet 1K
comestible, edible, ...	pot, flowerpot
dressing, salad dressing	cauliflower
Sicilian pizza	guacamole
vegetable, veggie, veg	cucumber, cuke
fruit	broccoli

F

dune buggy, beach buggy	warplane, military plane
searcher beetle, ...	missile
seeker, searcher, quester	projectile, missile
Tragelaphus eurycerus, ...	sports car, sport car
bongo, bongo drum	submarine, pigboat, sub, ...

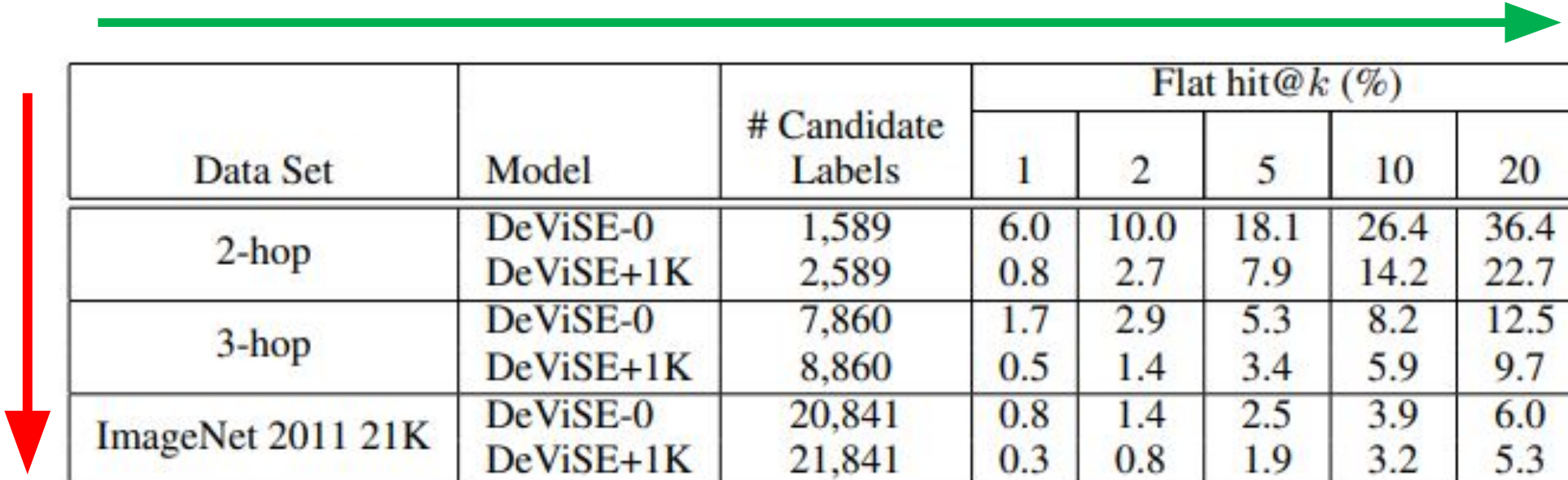
Results : Generalization & ZSL

- “2-hop”, “3-hop”, “ImageNet 2011 21K”



Results : Generalization & ZSL

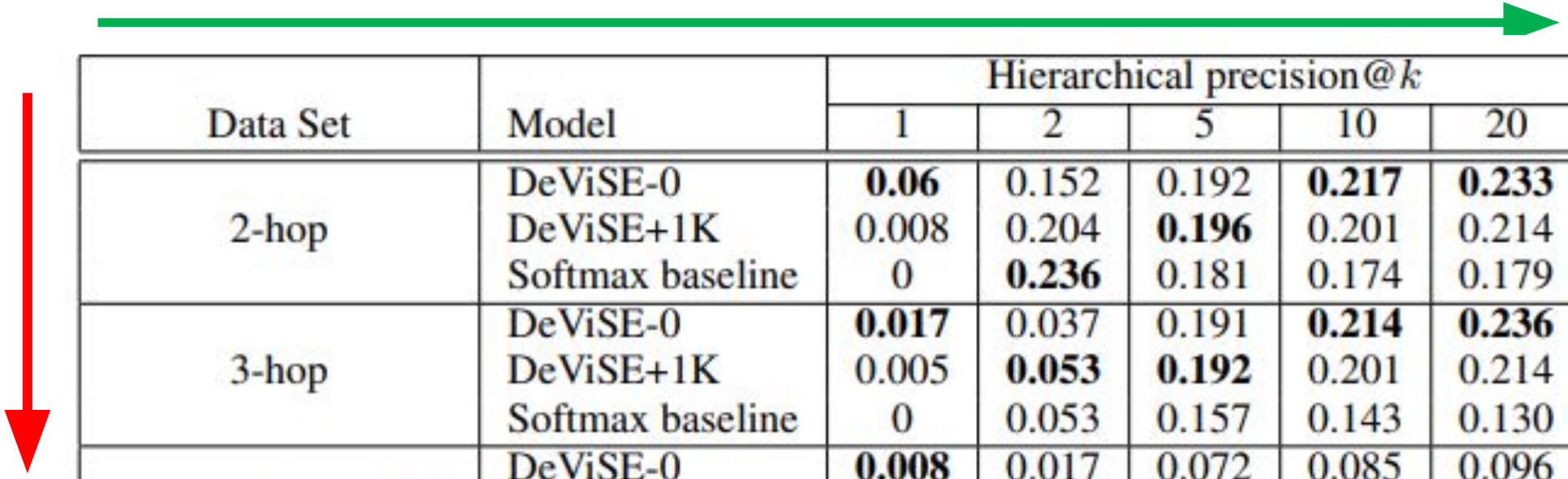
- Flat hit@k



Data Set	Model	# Candidate Labels	Flat hit@k (%)				
			1	2	5	10	20
2-hop	DeViSE-0	1,589	6.0	10.0	18.1	26.4	36.4
	DeViSE+1K	2,589	0.8	2.7	7.9	14.2	22.7
3-hop	DeViSE-0	7,860	1.7	2.9	5.3	8.2	12.5
	DeViSE+1K	8,860	0.5	1.4	3.4	5.9	9.7
ImageNet 2011 21K	DeViSE-0	20,841	0.8	1.4	2.5	3.9	6.0
	DeViSE+1K	21,841	0.3	0.8	1.9	3.2	5.3

Results : Generalization & ZSL

- Hierarchical precision @k



Data Set	Model	Hierarchical precision@k				
		1	2	5	10	20
2-hop	DeViSE-0	**0.06**	0.152	0.192	**0.217**	**0.233**
	DeViSE+1K	0.008	0.204	**0.196**	0.201	0.214
	Softmax baseline	0	**0.236**	0.181	0.174	0.179
3-hop	DeViSE-0	**0.017**	0.037	0.191	**0.214**	**0.236**
	DeViSE+1K	0.005	**0.053**	**0.192**	0.201	0.214
	Softmax baseline	0	0.053	0.157	0.143	0.130
ImageNet 2011 21K	DeViSE-0	**0.008**	0.017	0.072	0.085	0.096
	DeViSE+1K	0.003	**0.025**	**0.083**	**0.092**	**0.101**
	Softmax baseline	0	0.023	0.071	0.069	0.065

Conclusion

- By leveraging the semantic structure imparted by the language model, DeViSE make reasonable inferences about candidate labels it has never observed.

Reference

- ILSVRC 2012 1K dataset labels :
  https://gist.github.com/xkumiyu/dd200f3f51986888c9151df4f2a9ef30
- Evaluation Metric :
  https://ils.unc.edu/courses/2013_spring/inls509_001/lectures/10-EvaluationMetrics.pdf

- Images :
- https://pdfs.semanticscholar.org/ab64/16690dfdd4e255724a20160848ea4095afd3.pdf
- http://www.cs.virginia.edu/~vicente/vislang/slides/devise.pdf
- https://www.wired.com/story/dun-dun-duun-duuun-the-great-white-shark-genome-is-here/
- http://www.royalgazette.com/news/article/20180222/blue-shark-merlin-now-off-coast-of-bermuda

Q & A