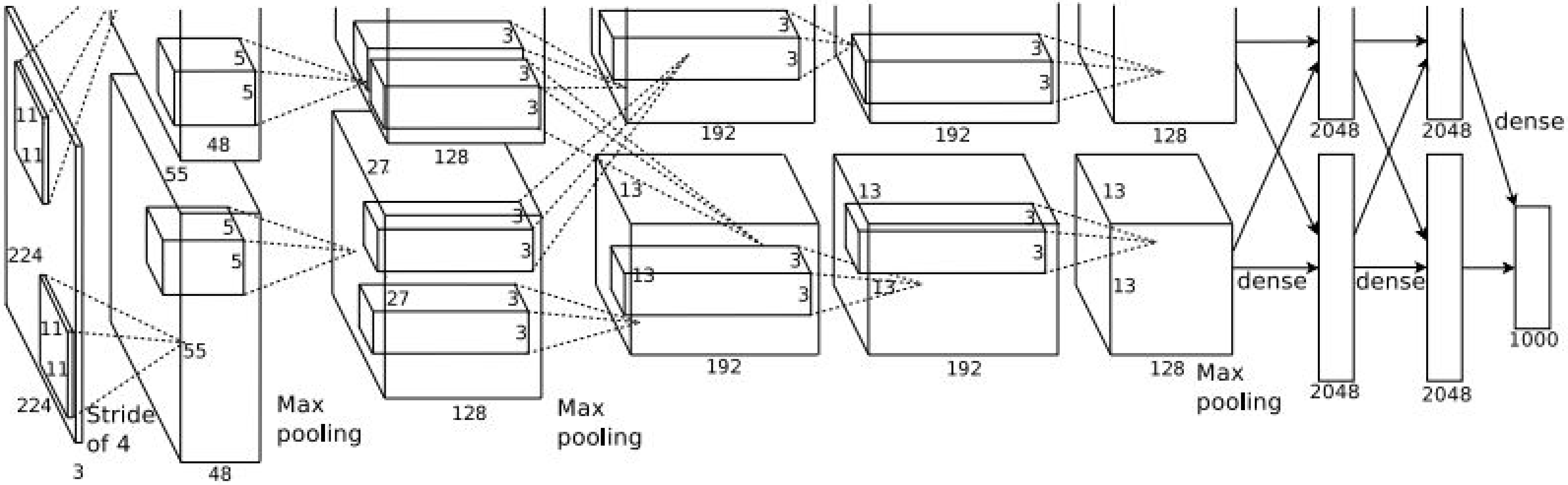
DeViSE: A Deep Visual-Semantic Embedding Model

Frome et al., Google Research

Presented by: Tushar Nagarajan

The year is 2012...



Krizhevsky et al. 2012

The year is 2012...

Koala?

Yes

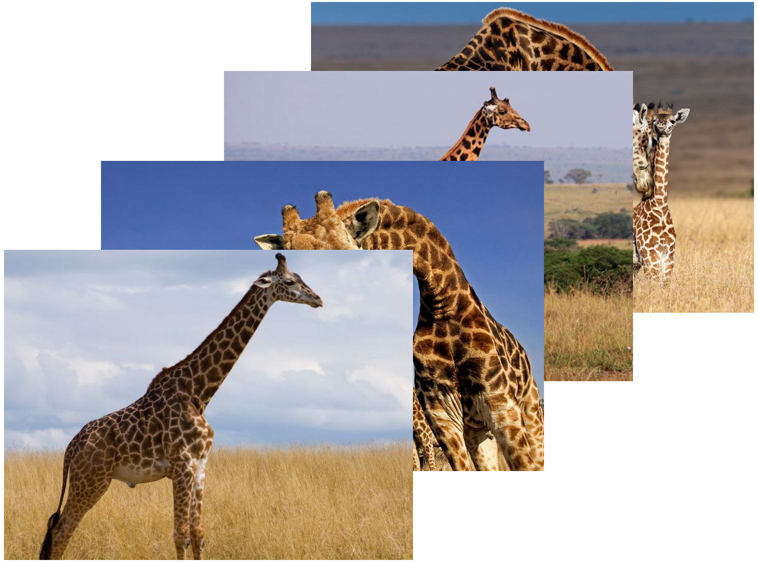
Cat?

Of course! Don’t be silly.

Giraffe?

What’s that?

The year is 2012...



Collect more giraffe data

Horse?

:|

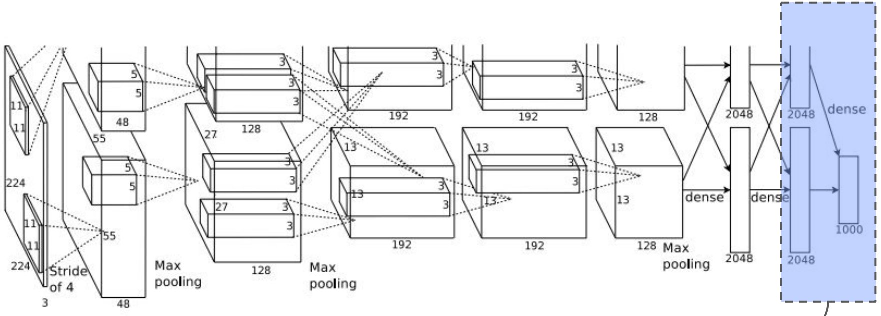
Imagenet 1k

- Only 1000 classes
- 3 year olds have a 1k word vocabulary

Getting data is hard

Label: “This thing”

Re-training networks is annoying



Doesn’t scale easily

Structure in Labels

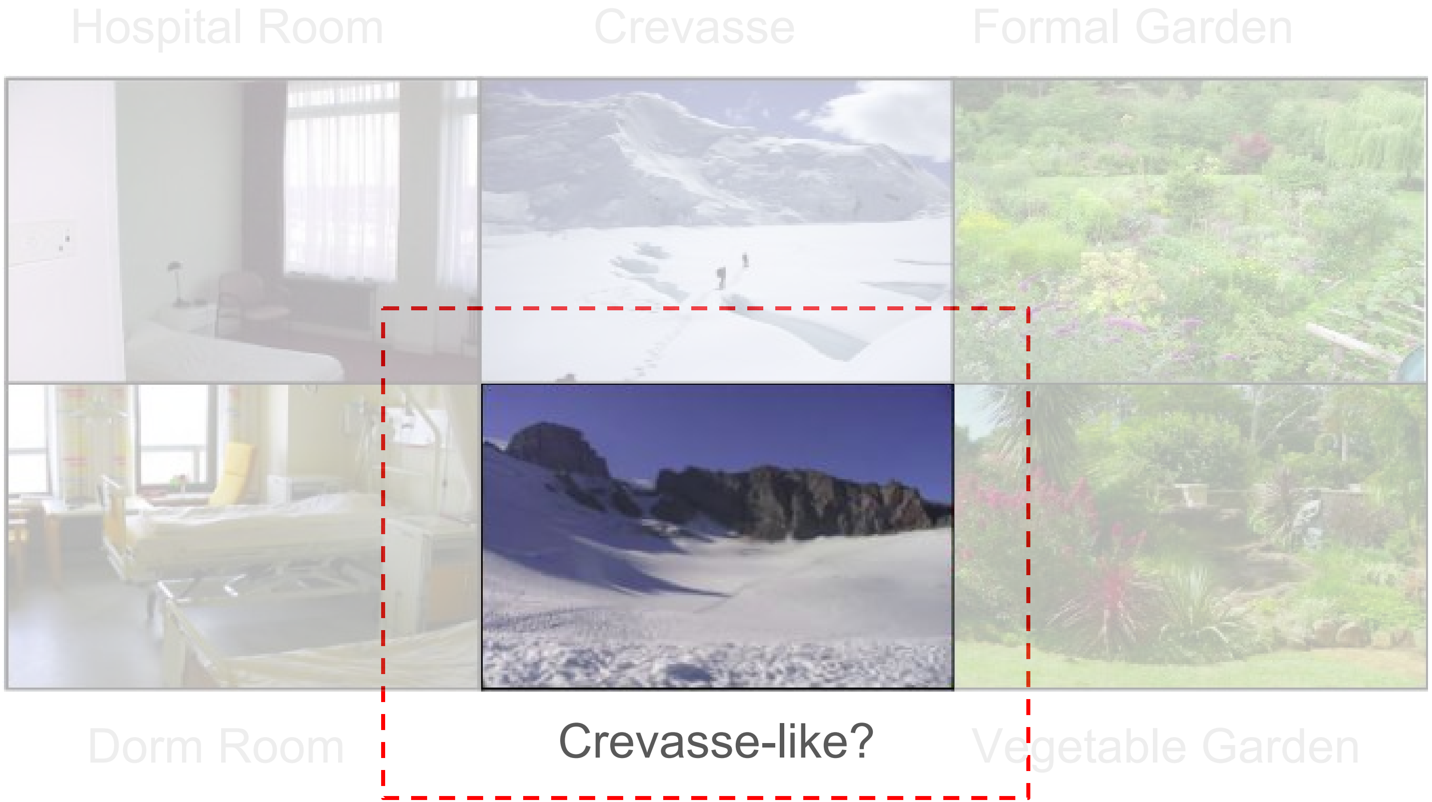
Label Structure - Similarity

Hospital Room | Crevasse | Formal Garden

Dorm Room | Snowfield | Vegetable Garden

SUN dataset, Xiao et al.

Label Structure - Similarity



SUN dataset, Xiao et al.

Label Structure - Similarity

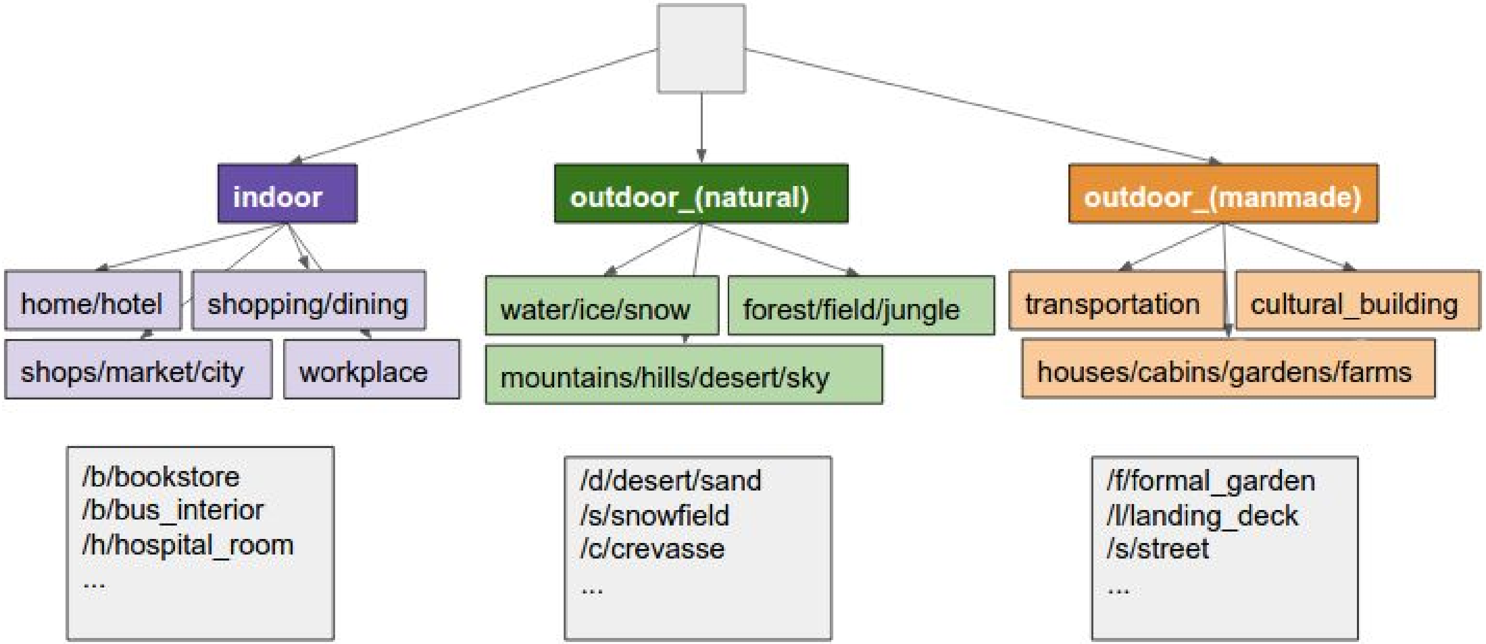
similar(Crevasse, Snowfield)

Visual

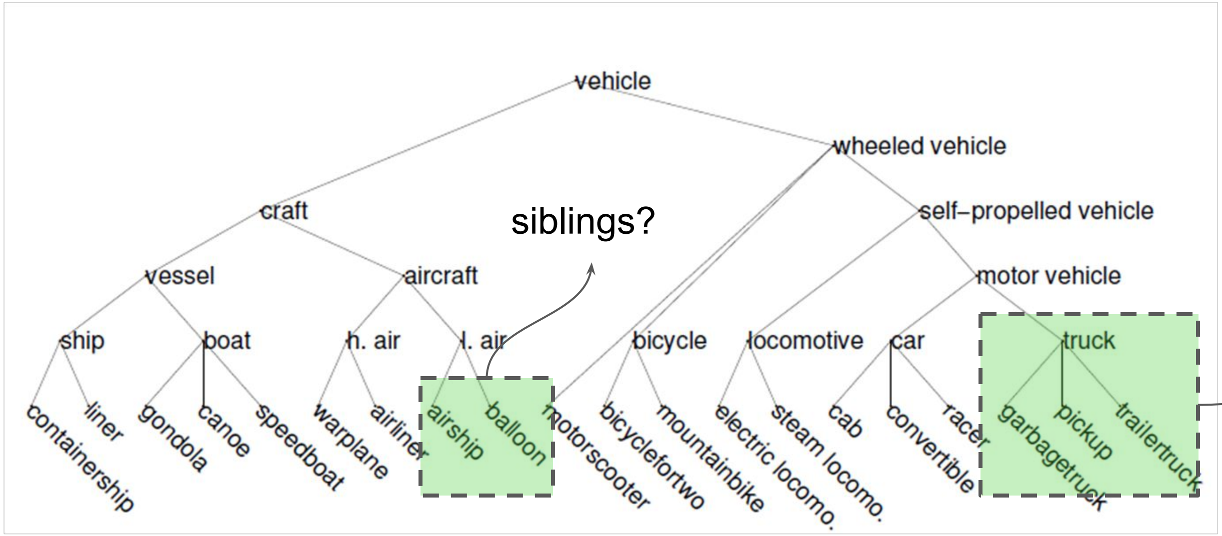
similar(Guitar, Harp)

Semantic

Label Structure - Hierarchy



Label Structure - Hierarchy

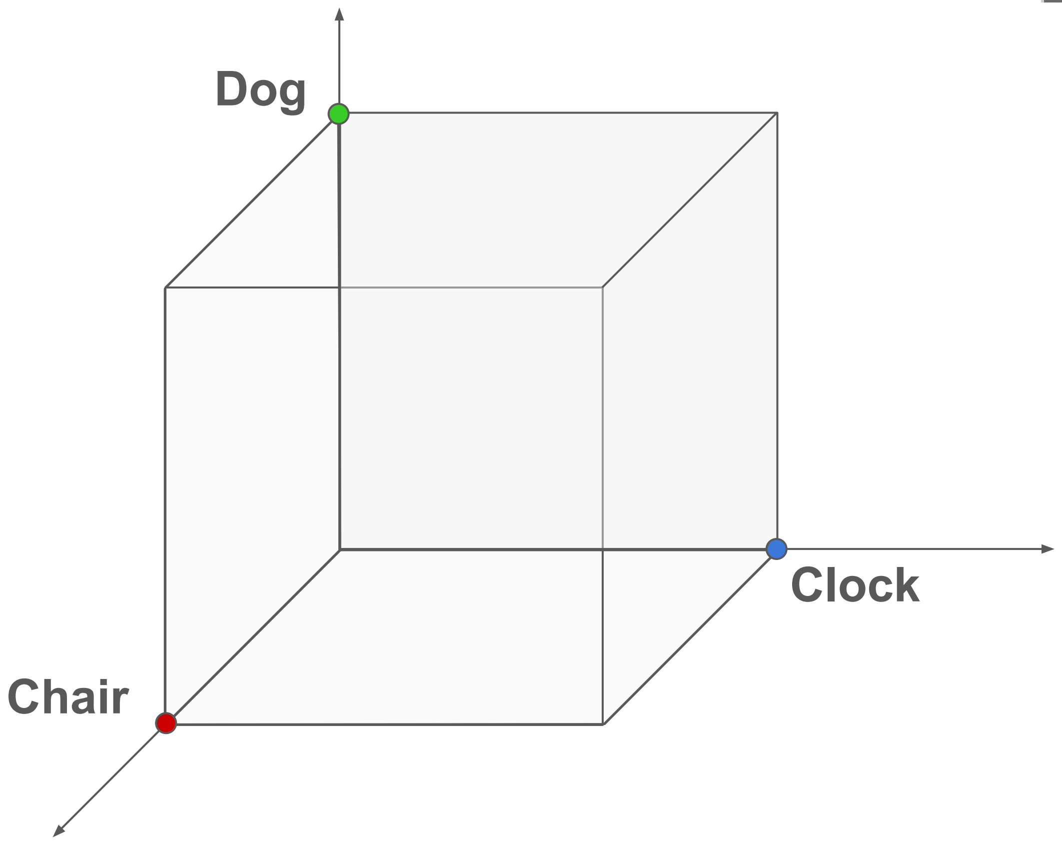


parents?

Hwang et al., 2011

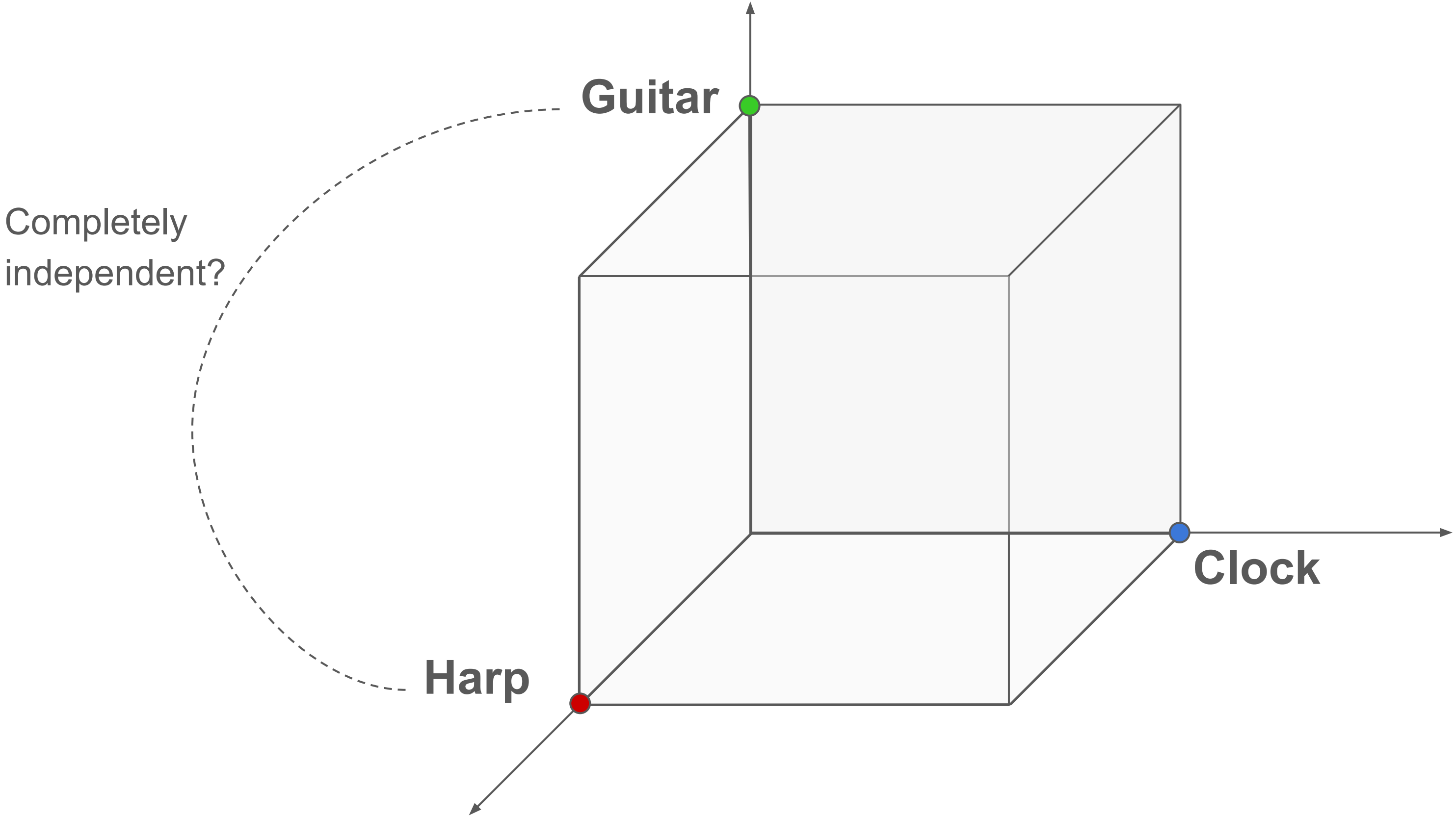
Does Softmax Care?

$$P(y = j|\mathbf{x}) = \frac{e^{\mathbf{x}^\top \mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\top \mathbf{w}_k}}$$



Does Softmax Care?

$$P(y=j|\mathbf{x}) = \frac{e^{\mathbf{x}^\mathsf{T}\mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\mathsf{T}\mathbf{w}_k}}$$

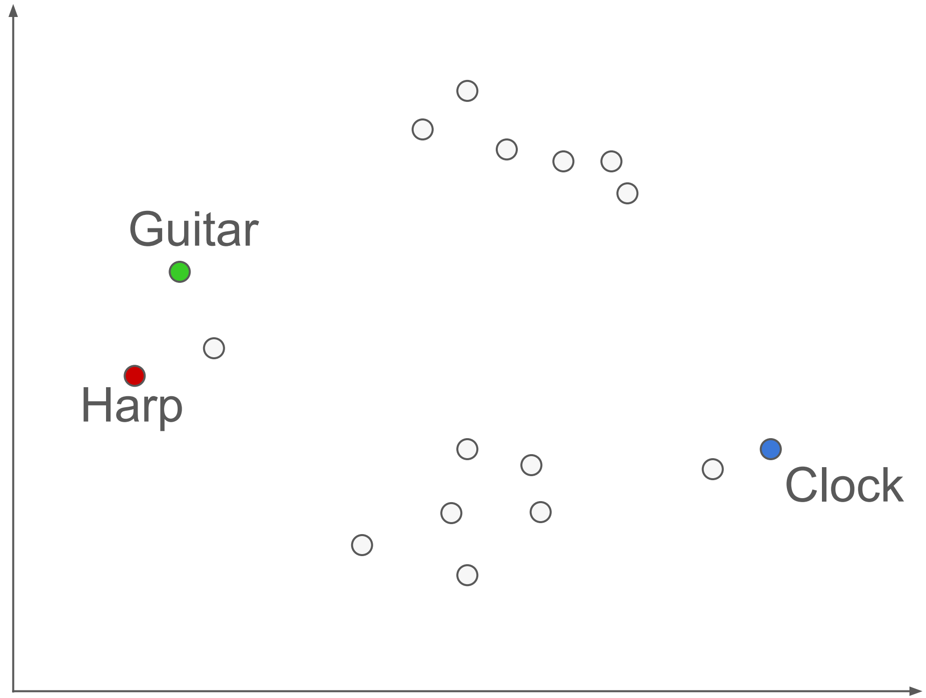


Does Softmax Care?

Are labels independent?

Not really - guitar and harp are more closely related than guitar and clock.

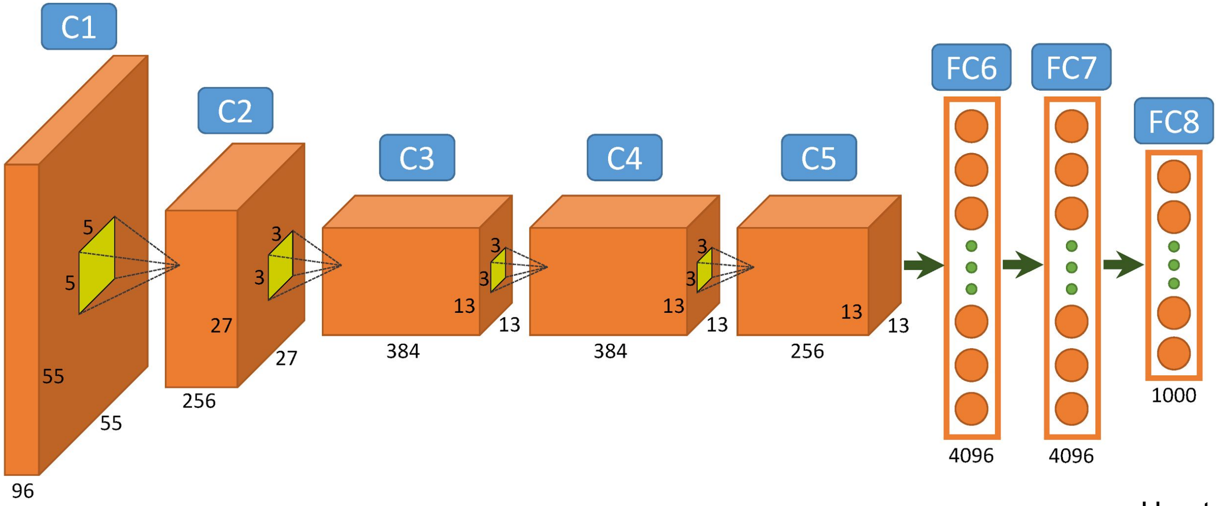
Abandon softmax - move to **label space**



Regress to Label Space

Step 1: Train a CNN for classification

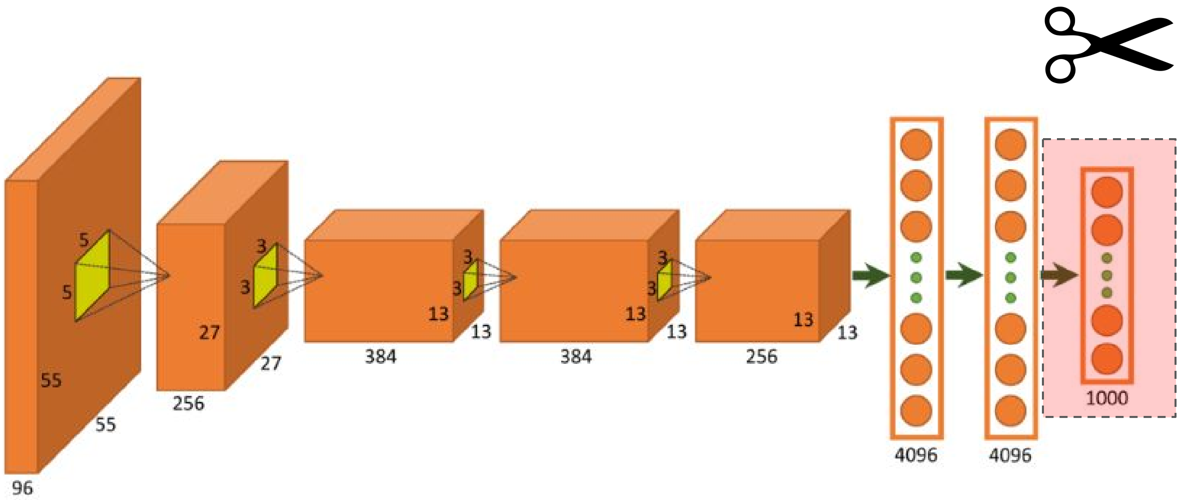
- Regular CNN for object classification
- **1000 way softmax** output



Hu et al., Remote Sens. 2015

Regress to Label Space

Step 2: Abandon Softmax

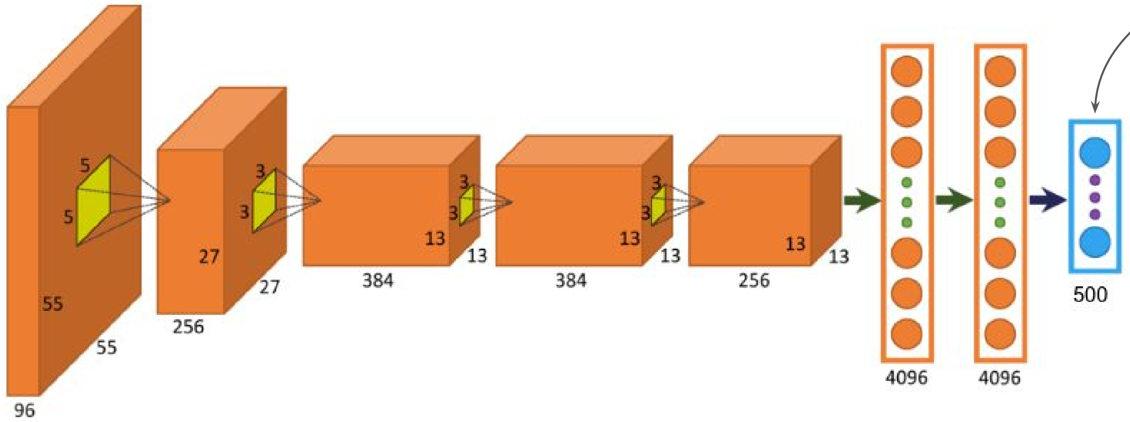


Hu et al., Remote Sens. 2015

Regress to Label Space

Step 2: Abandon Softmax

What regression labels?



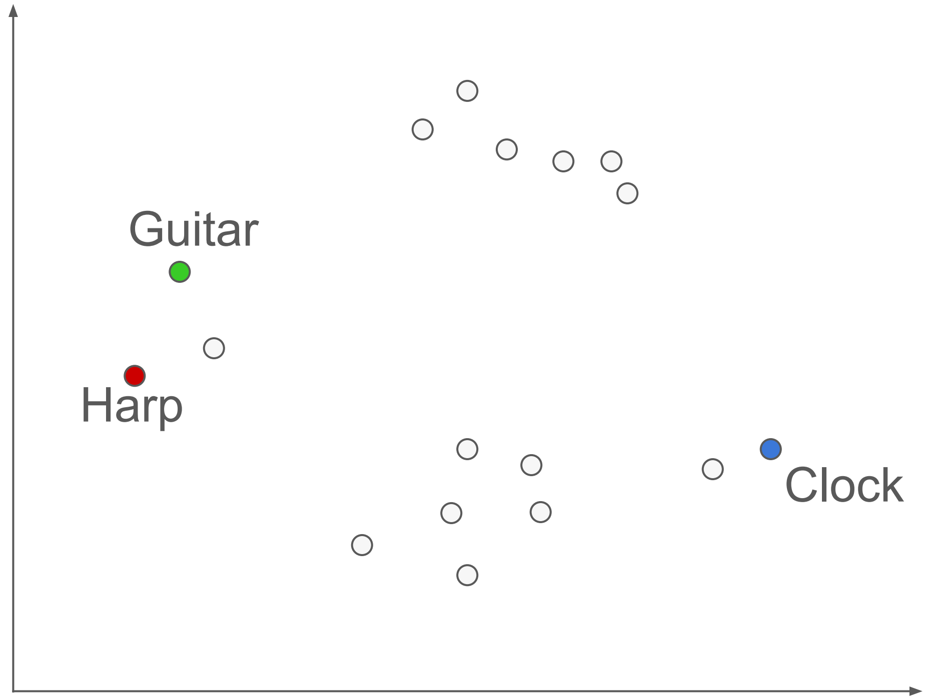
Hu et al., Remote Sens. 2015

Label Space

We didn't think this through…

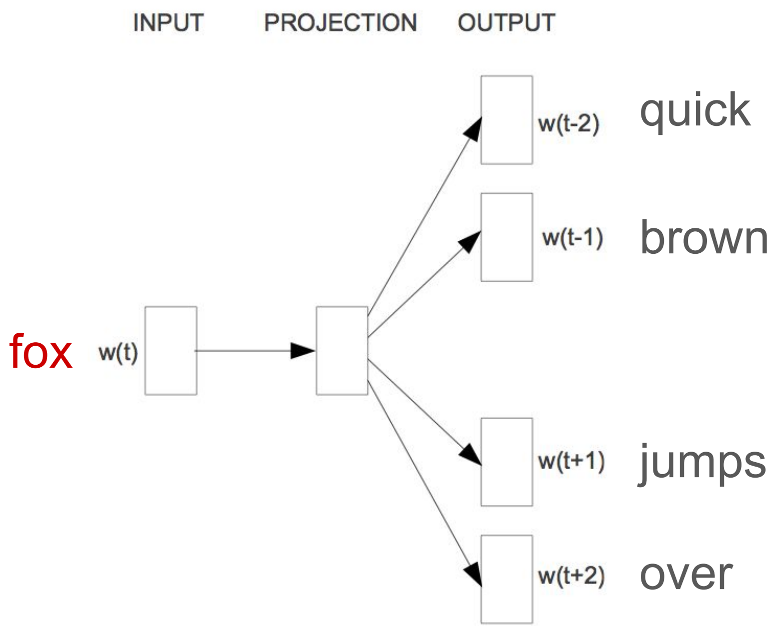
Where do we get this space from?

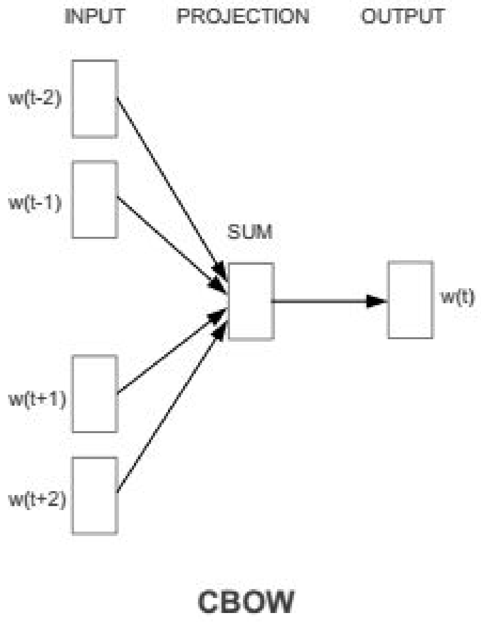
Hint: Imagenet classes are words!



Word Embeddings - Skip-gram

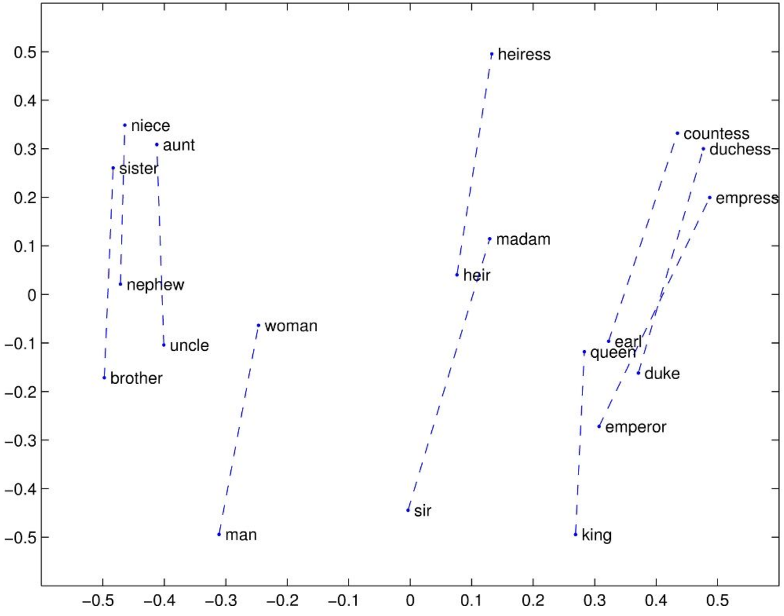
The **quick brown fox jumps over** the lazy dog.



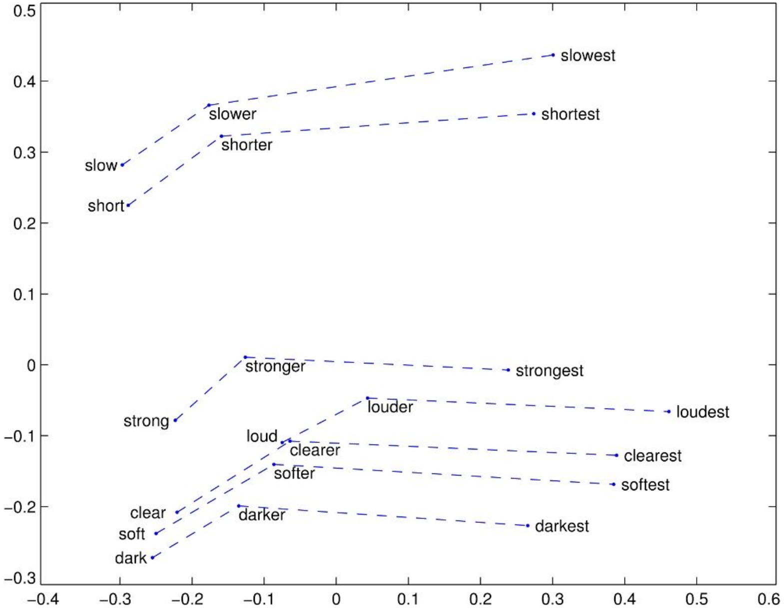


Mikolov et al., 2013

Word Embeddings - Skip-gram



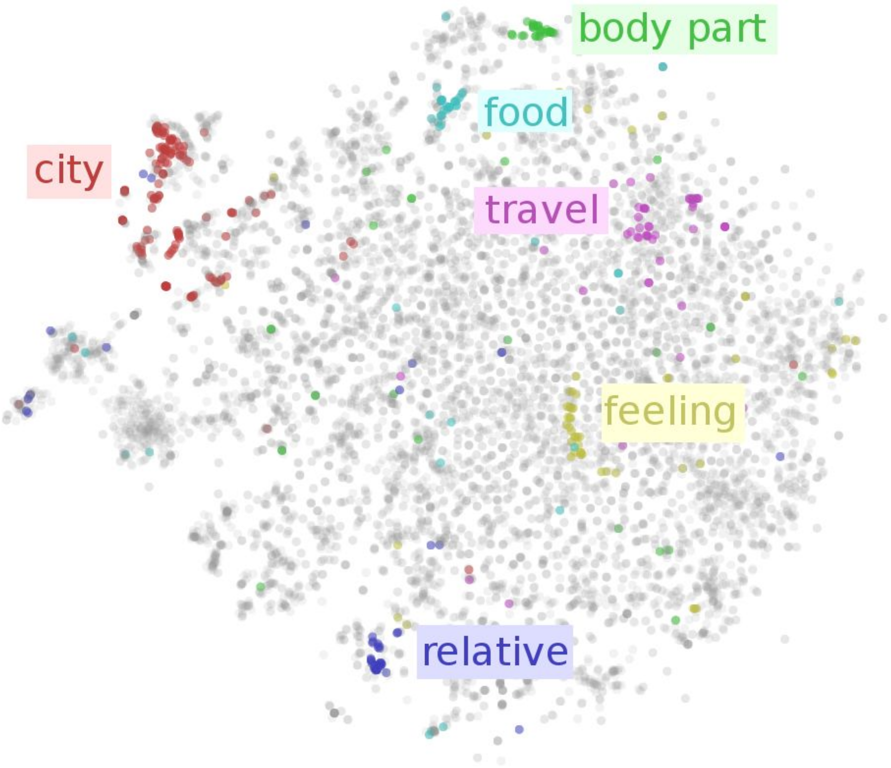
Gender encoded into subspace



comparative - superlative info

Mikolov et al., 2013

Word Embeddings - Skip-gram

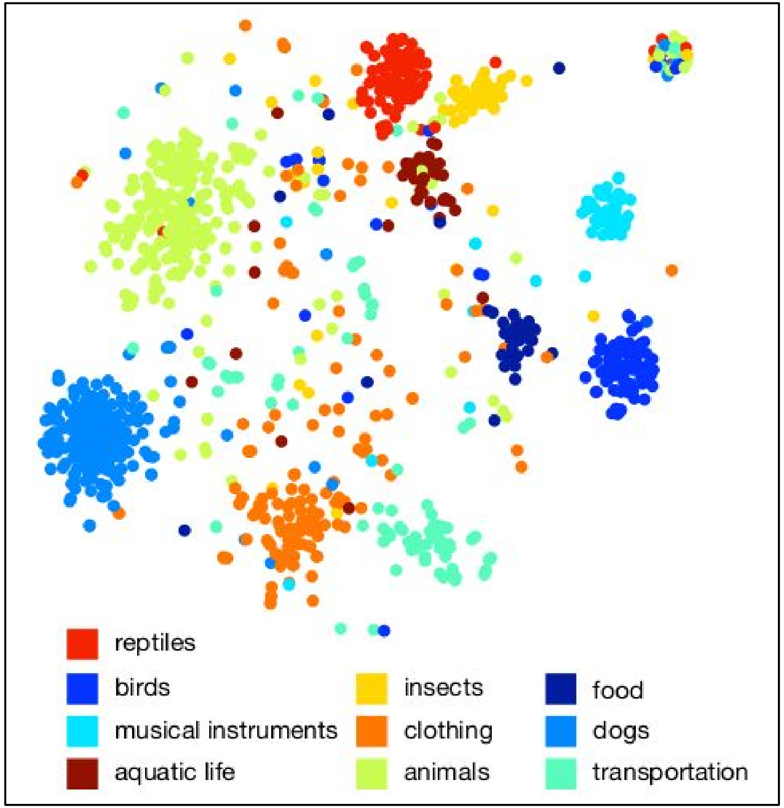


Sebastian Ruder

Word Embeddings - Skip-gram

Step 3: Train a LM on **5.7M documents** from wikipedia

- 20 word window
- Hierarchical Softmax
- 500D vectors



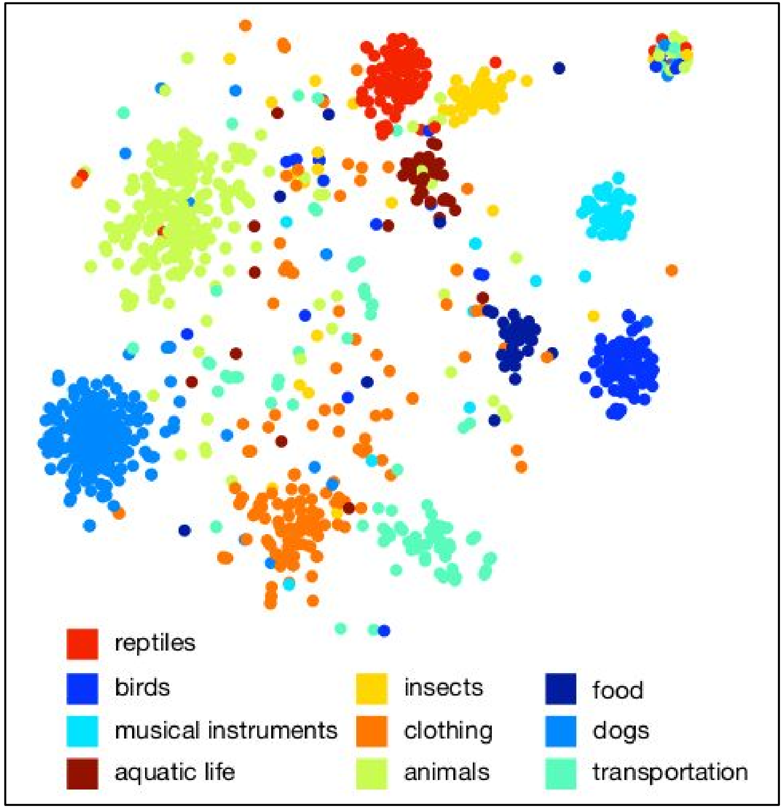
Q: What about multi-word classes like “snow leopard”?

Frome et al., 2013

Step 1: Train a CNN for classification
Step 2: Abandon Softmax
Step 3: Train a skip-gram LM

Word Embeddings - Skip-gram

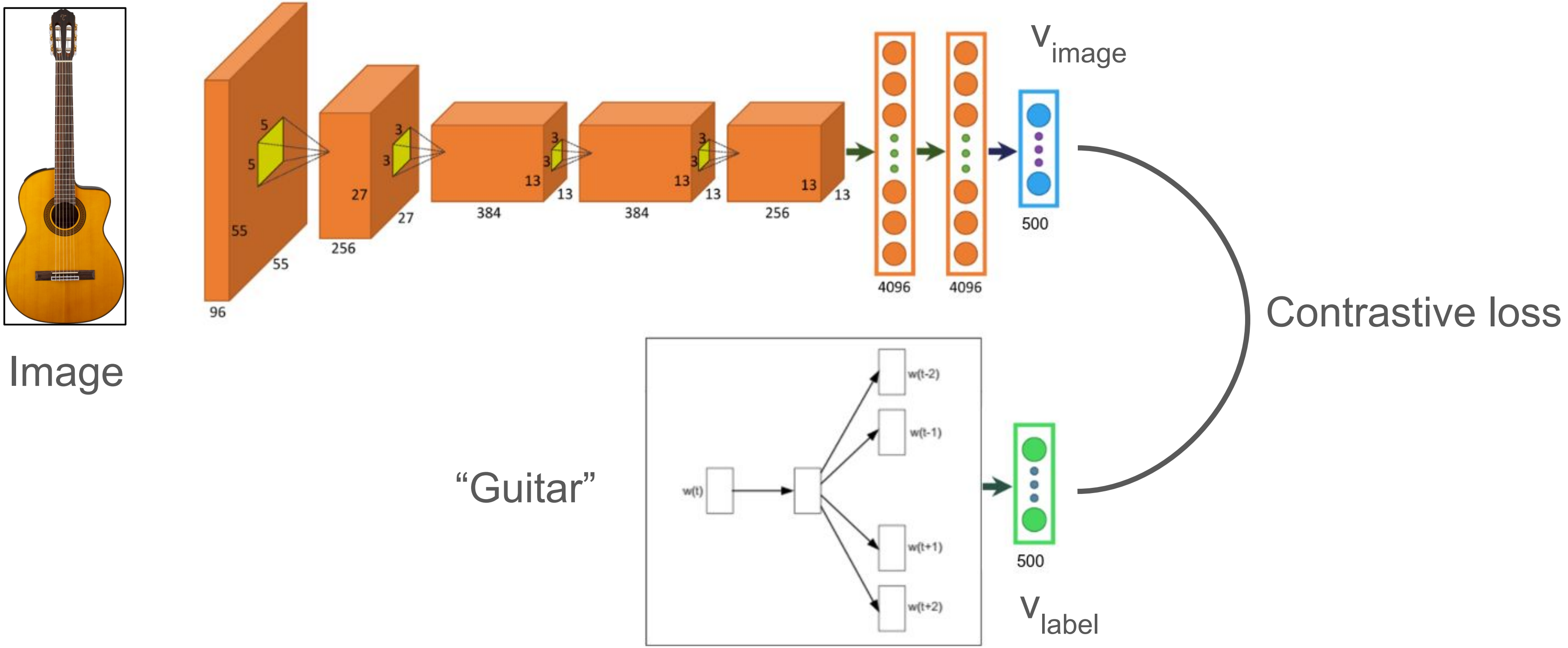
Tiger Shark	Car
Bull shark Blacktip shark Shark Blue shark ...	Cars Muscle car Sports car Automobile ...



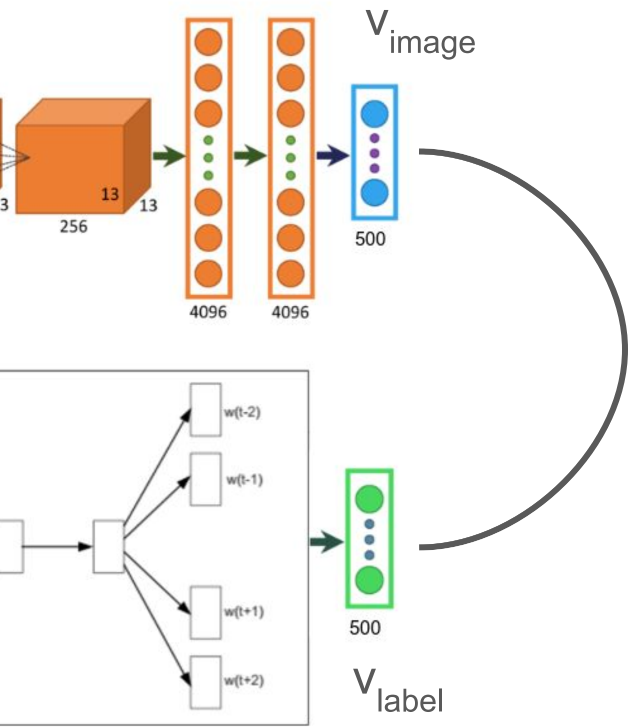
Frome et al., 2013

Step 1: Train a CNN for classification
Step 2: Abandon Softmax
Step 3: Train a skip-gram LM

Step 4: Surgery



Step 4: Surgery



Contrastive loss

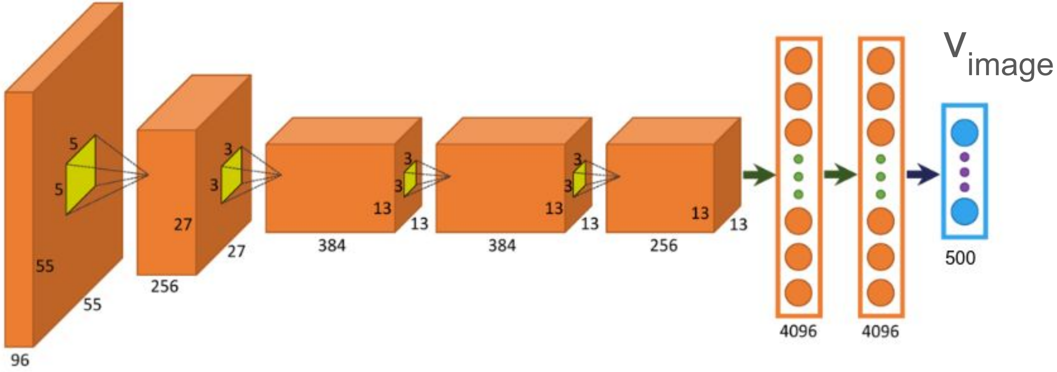
$$loss(image, label) = max \begin{cases} 0 \\ \Delta - v_{label} \cdot v_{image} + v_{neg} \cdot v_{image} \end{cases}$$

margin

random incorrect class

Inference - ZSL

When a new image comes in:

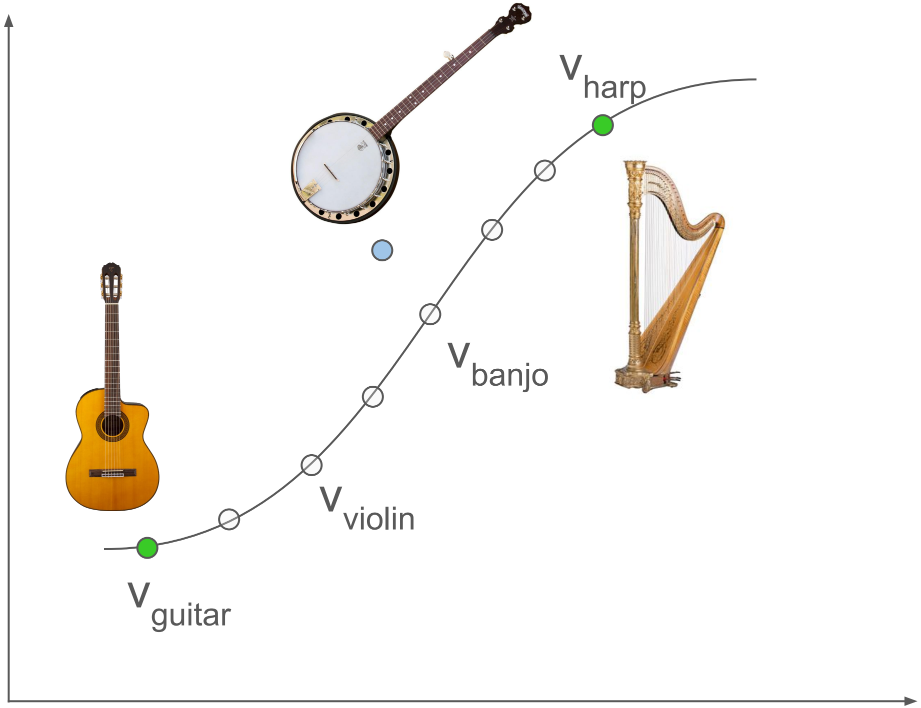


1. Push it through the CNN, get v_{image}

Inference - ZSL

When a new image comes in:

1. Push it through the CNN, get v_{image}

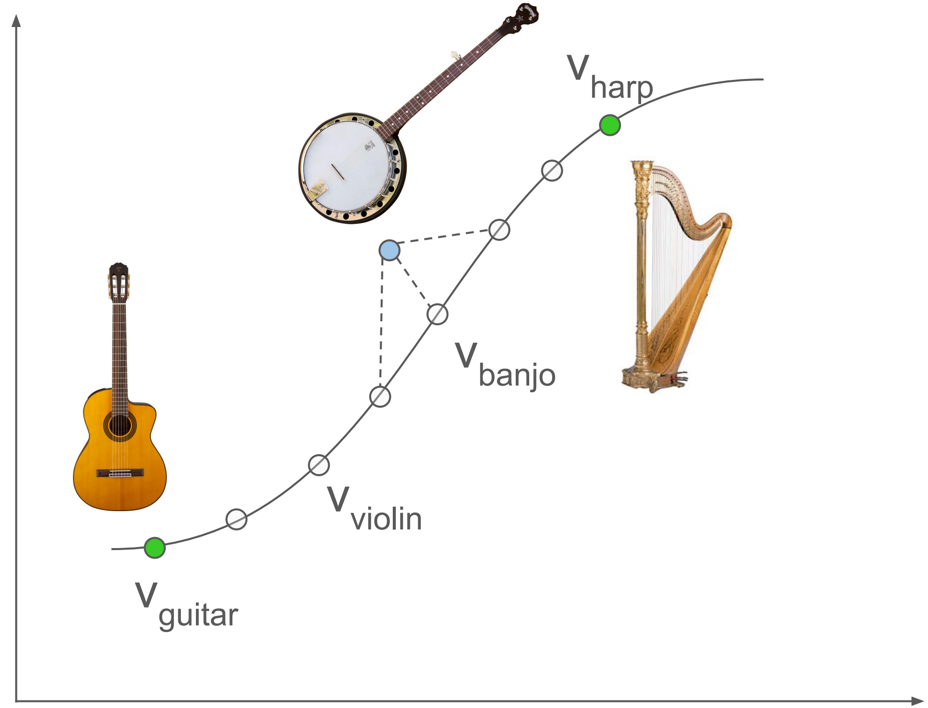


Inference - ZSL

When a new image comes in:

1. Push it through the CNN, get v_{image}
2. Find the **nearest** v_{label} to v_{image}

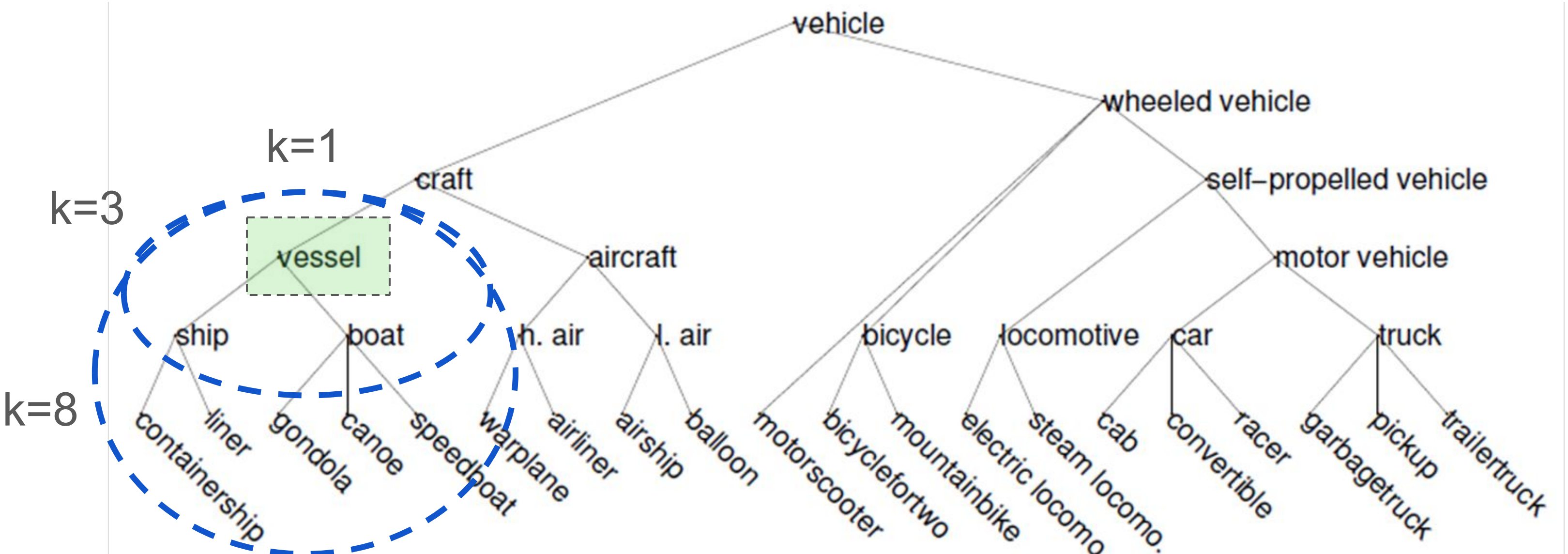
Potentially unseen labels!



Results

Evaluation Metrics

- Flat hit @ k : Regular precision
- Hierarchical precision @ k:



Results on Imagenet

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

Softmax is hard to beat on raw classification on 1k classes

DeViSE gets pretty close with a **regression model!**

Frome et al., 2013

Results - Imagenet Classification

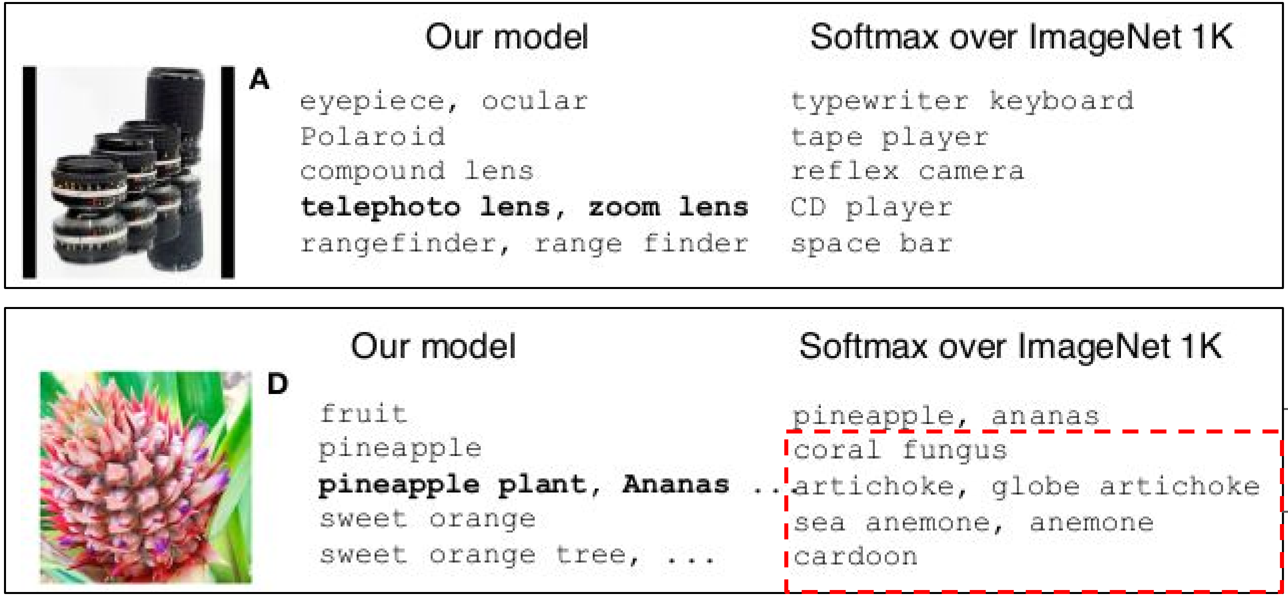
Model type	dim	Flat hit@k (%)				Hierarchical precision@k			
		1	2	5	10	2	5	10	20
Softmax baseline	N/A	**55.6**	**67.4**	**78.5**	**85.0**	0.452	0.342	0.313	0.319
DeViSE	500	53.2	65.2	76.7	83.3	0.447	**0.352**	**0.331**	**0.341**
	1000	54.9	66.9	78.4	**85.0**	**0.454**	0.351	0.325	0.331
Random embeddings	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

Hierarchical precision tells a different story

DeViSE finds labels that are **semantically relevant**

Frome et al., 2013

Results - Imagenet ZSL



Correct label @1

garbage?

Frome et al., 2013

Results - Imagenet ZSL

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

Our model	Softmax over ImageNet 1K
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, ...

Frome et al., 2013

Results - Imagenet ZSL

Data Set	Model	Flat hit @k			Hierarchical @k		
		1	5	20	1	5	20
3-hop	DeViSE	1.7	5.3	12.5	1.7	19.1	23.6
	Softmax	-	-	-	0	15.7	13
Imagenet 21k	DeViSE	0.8	2.5	6	0.8	7.2	9.6
	Softmax	-	-	-	0	7.1	6.5

3-hop: Unknown classes 3 hops away from imagenet labels

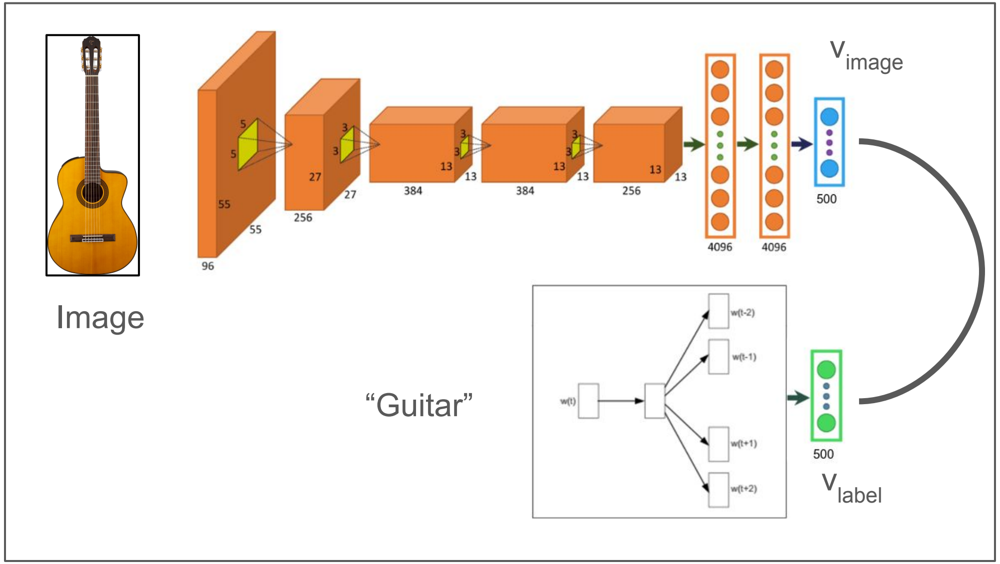
Imagenet 21k: **ALL unknown** classes

Chance: 0.00047
168x better!

Frome et al., 2013

Summary

Step 1: Train a CNN for classification
Step 2: Abandon Softmax
Step 3: Train a skip-gram LM
Step 4: Surgery
Step 5: Profit?



The Register®
Biting the hand that feeds IT

Googlers devise DeViSE: A thing-recognising FRANKENBRAIN

Machine-learning tech glues together image eyeballing and text grokking

The Register, 2013

Discussion

Embeddings are not fine-tuned during training

Semantic similarity is a happy coincidence

- sim(cat, kitten) = 0.746
- sim(cat, dog) = **0.761 (!!)**

Semantic similarity is a **depressing** coincidence

sim(happy, depressing) = ?

Discussion

D

Our model	Softmax over ImageNet 1K
fruit	pineapple, ananas
pineapple	coral fungus
pineapple plant, Ananas ...	artichoke, globe artichoke
sweet orange	sea anemone, anemone
sweet orange tree, ...	cardoon

Nearest neighbors of **pineapple**:

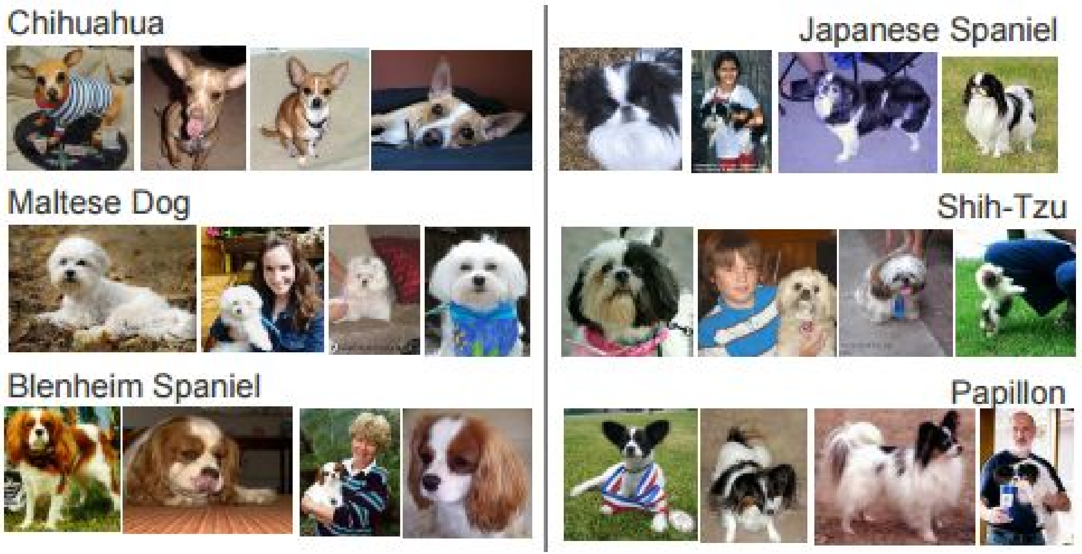
Pineapples, papaya, mango, avocado, banana ...

Frome et al., 2013

Discussion

Categories are fine-grained

We TRUST softmax to distinguish them



Stanford Dogs Dataset - Khosla et al., 2011

Conclusion

Label spaces to embed semantic information

Shared embedding spaces

background knowledge for ZSL

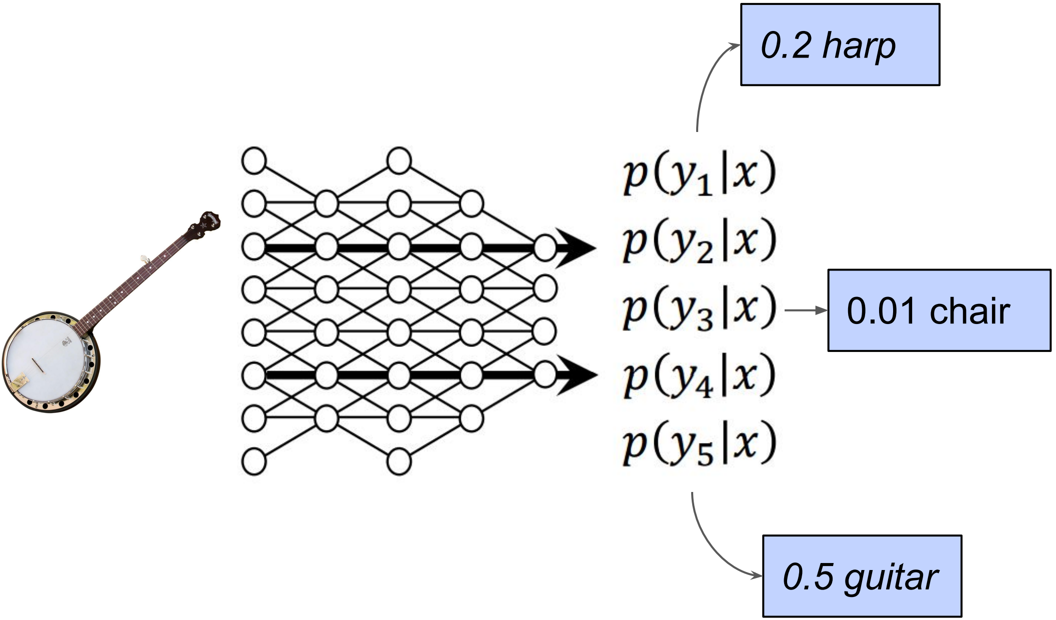
Zedonk

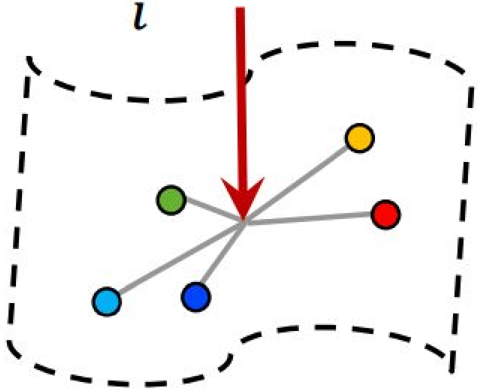
Thank you

Questions?

Bonus: ConSE

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





$$v_{label} = 0.2 \times v_{harp} + 0.5 \times v_{guitar} + 0.01 \times v_{chair} + \ldots$$

Norouzi et al., 2013