DeViSE: A Deep Visual-Semantic Embedding Model

Presenters: Ji Gao, Fandi Lin

Motivation

Visual recognition systems experience problems with large amount of categories.

- Insufficient labeled training data
- Blurred distinction between classes

How do we improve predictions of unknown categories?

Background

N-way discrete classifiers

- Labels treated as unrelated
- Semantic information not captured

Result: These systems cannot make zero-shot predictions without additional information, i.e. text data.

Related Work

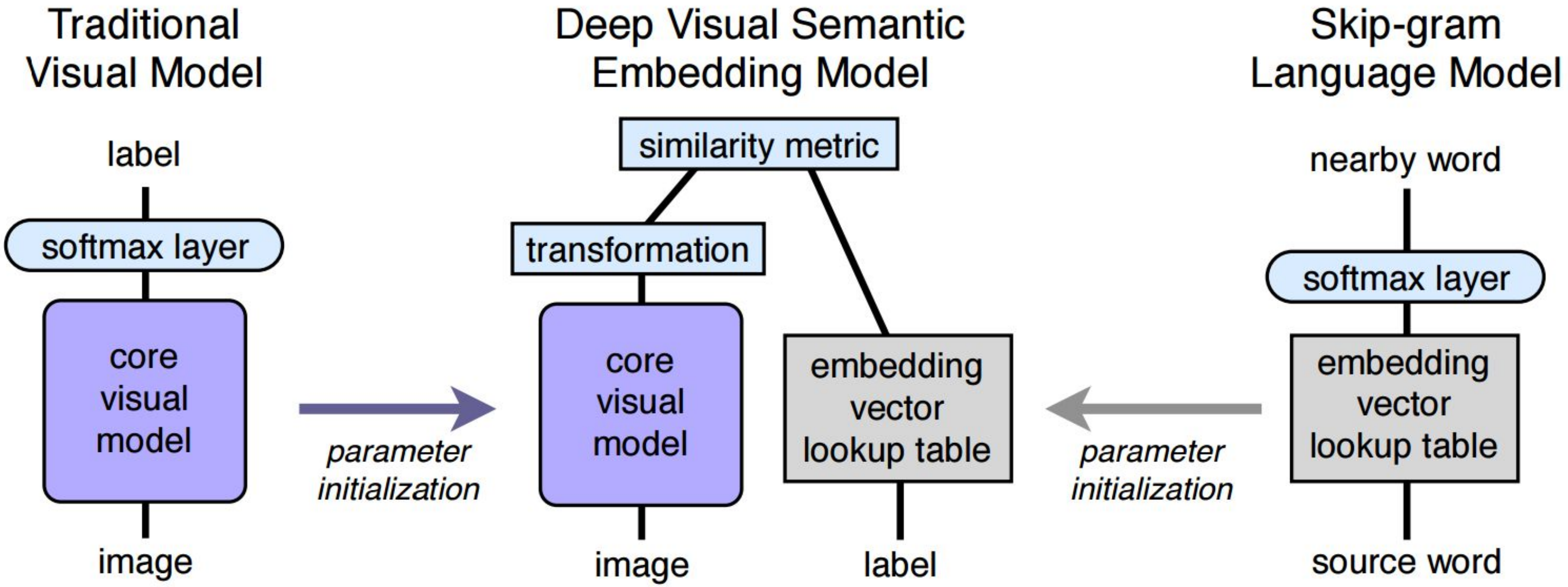
WSABIE: Linear map from image features to embedding space. Only used training labels.

Socher et al: Linear map from image features to embedding space. Outlier detection. Only 8 known and 2 unknown classes.

Other work that has shown zero-shot classification relies on curated information.

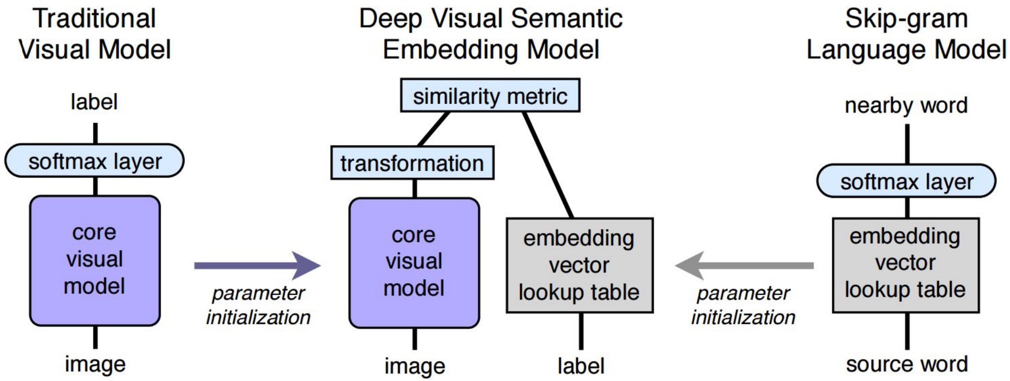
Proposed Method

Combine a traditional Visual model with a language model.



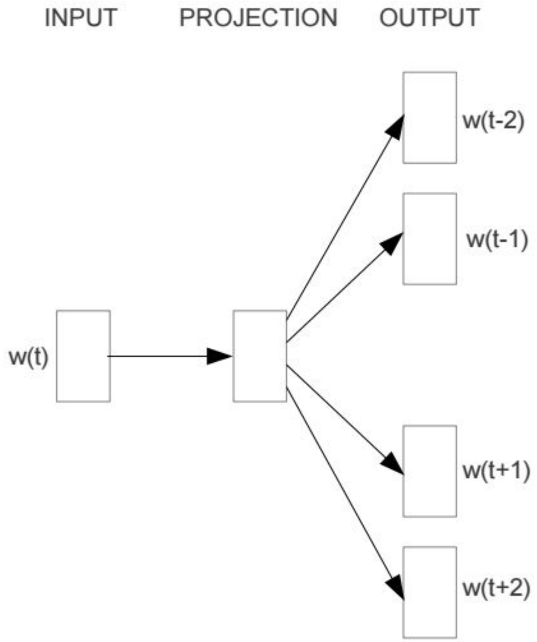
Proposed Method

1. Train a language model for semantic information
2. At the same time, train a CNN for images
3. Initialize the combined model using pre-trained parameters
4. Train the combined model



Skip-gram language model

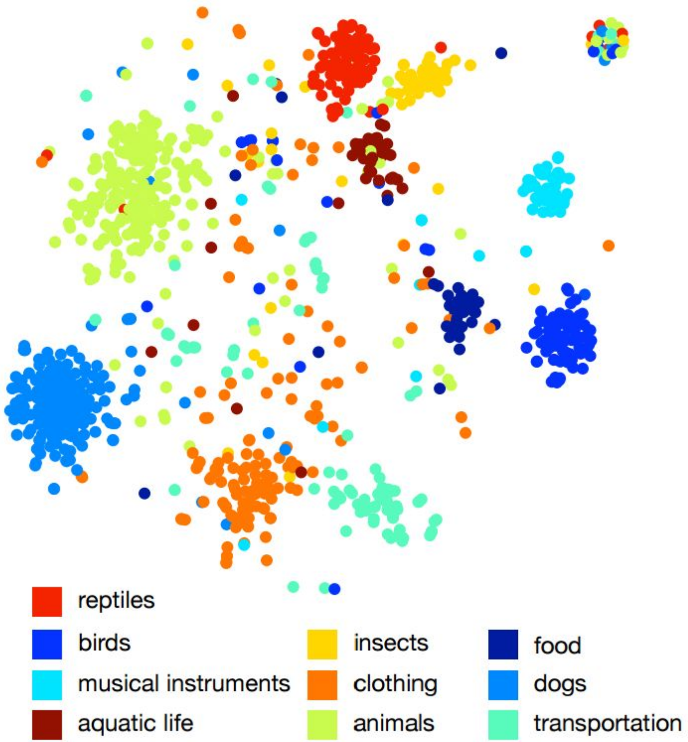
- *Efficient estimation of word representations in vector space, ICLR 2013*
- Skip-gram: **a generalization of *n*-grams which skips the words between**
- Skip-gram model: Learn a NN from a word to predict nearby words.



Skip-gram language model

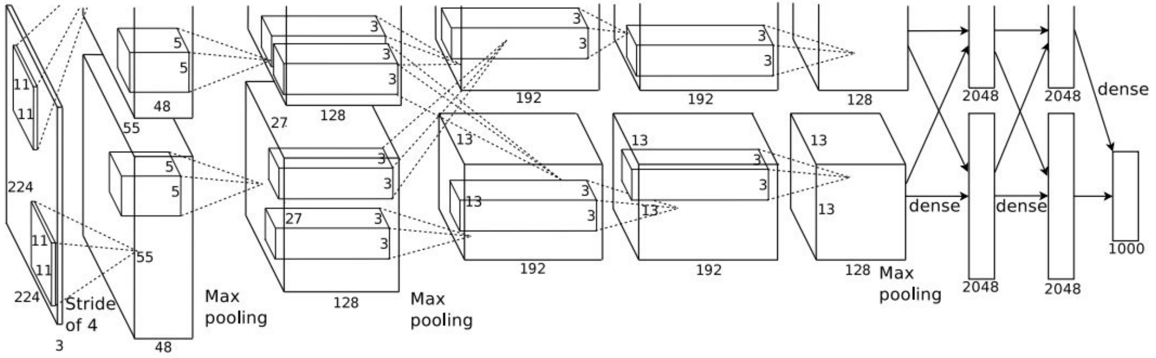
Learn the relationship between labels.

- **Data: 5.7 million documents (5.4 billion words) extracted from wikipedia.org**

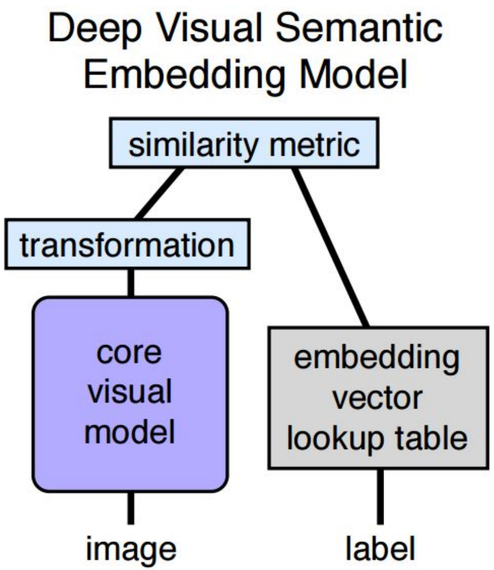


CNN model

- **AlexNet**
- **Winner of ILSVRC 2012**
- **5 conv layers**



Combined model



Use a linear embedding layer to map the features extracted before Softmax(4096d) to match the size of the language model(500 or 1000d).

Loss function:

$$loss(image, label) = \sum_{j \neq label} \max[0, margin - \vec{t}_{label} M \vec{v}(image) + \vec{t}_j M \vec{v}(image)]$$

Experiment

Task:

- **Image classification**
- **Zero-shot image classification**

Experiment - With same label set (not zero-shot)

Baselines:

- **Alexnet**
- **Random Embedding: Alexnet + a random vectors (instead of the language model)**

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

Table 1: Comparison of model performance on our test set, taken from the ImageNet ILSVRC 2012 1K validation set. Note that hierarchical precision@1 is equivalent to flat hit@1. See text for details.

Experiment: Zero-shot

Dataset:

- 2-hop: two clusters of labels
- 3-hop: three clusters of labels
- ImageNet2011: Use labels in ImageNet2011 that doesn’t appear in ImageNet2012

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

Experiment: Zero-shot

Comparing to pure CNN:

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

Experiment: Zero-shot

Compare to previous zero-shot result

Model	200 labels	1000 labels
DeViSE	31.8%	9.0%
Mensink et al. 2012 [12]	35.7%	1.9%
Rohrbach et al. 2011 [17]	34.8%	-

Conclusion

DeViSE achieves state-of-the-art performance in classification task, and also able to do zero-shot learning.

Suitable for large amount of data, and can handle labels with not enough number of data.

Show the power of combining image and semantic data.