Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics

D. Kiela and L. Bottou, 2014

Research questions and motivations

Studies show that:

- human semantic knowledge relies heavily on perceptual information
- multi-modal semantic representation models outperform uni-modal linguistic models

There has been research into multi-modal representation models that apply deep learning

Key contributions of the paper

Use CNNs in multi-modal semantics

First approach to exclusively use deep learning to get input representations

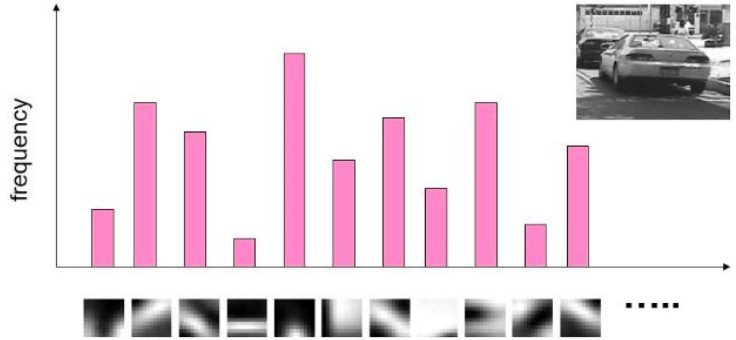
Methodology

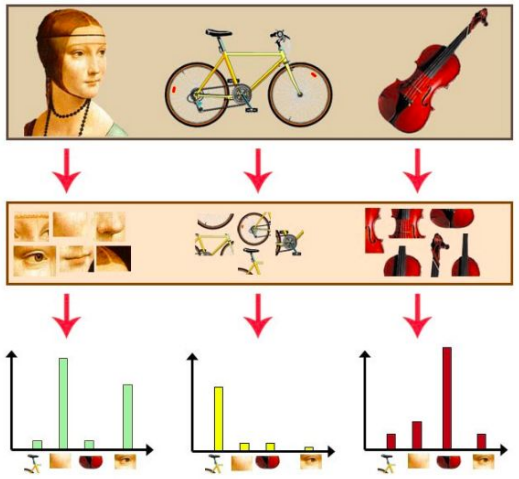
Representations:

- Perceptual (visual) representation
- Linguistic representation

Perceptual representation - baseline

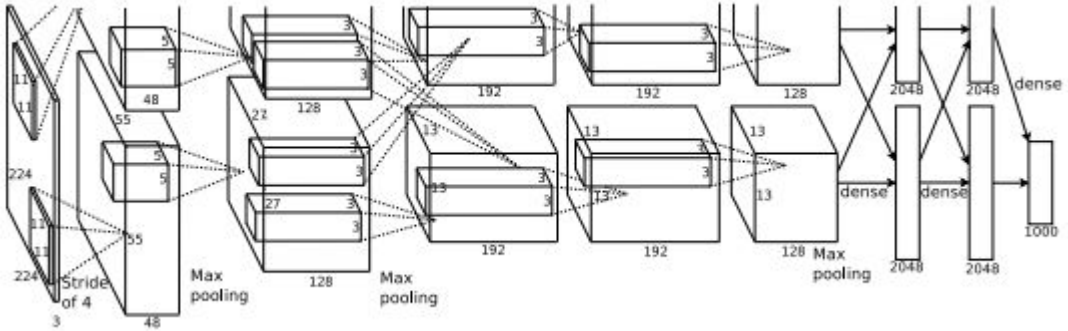
The perceptual component is often an instance of the bag-of-visual-words (BOVW) - akin to BOW but for visual features





Perceptual representation

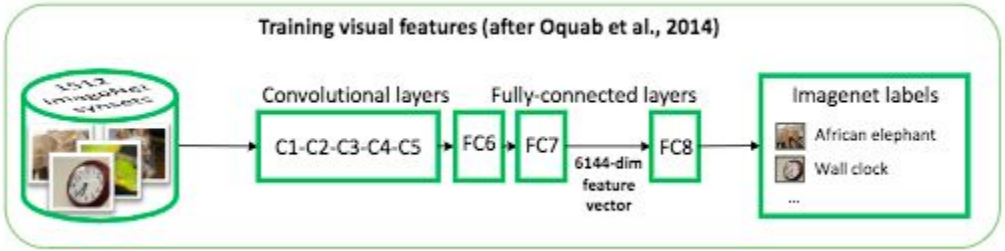
Follow approach described by Oquab et al. (2014)



Network structure presented in Krizhevsky et al. (2012)

Perceptual representation

Trained on 1.6m ImageNet images associated with 1512 output categories from output layer



In: 224x224 RGB images

Out: 6144-dimensional feature vector

Perceptual representation

Two ways to aggregate feature vectors for every concept:

- CNN-mean

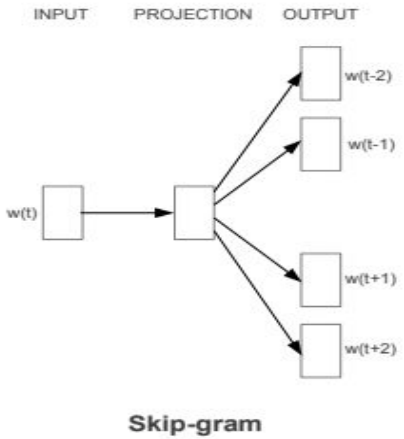
  [3, 1, 1] + [0, 1, 2] = [1.5, 1, 1.5]

- CNN-Max

  [3, 1, 1] + [0, 1, 2] = [3, 1, 2]

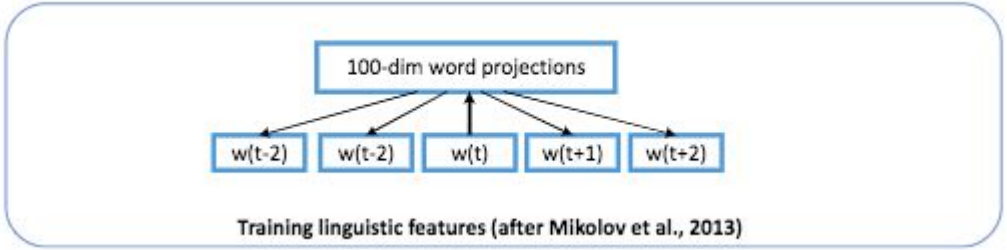
  Makes sense here as feature vectors are sparse (22% non-zero coefficients)

Linguistic representation



Model architecture as shown in Mikolov et al. (2013)

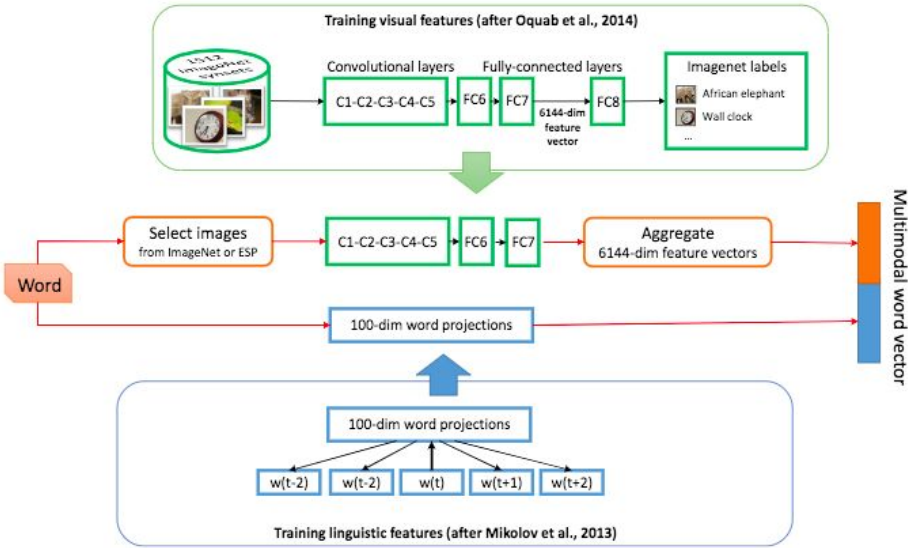
Linguistic representation



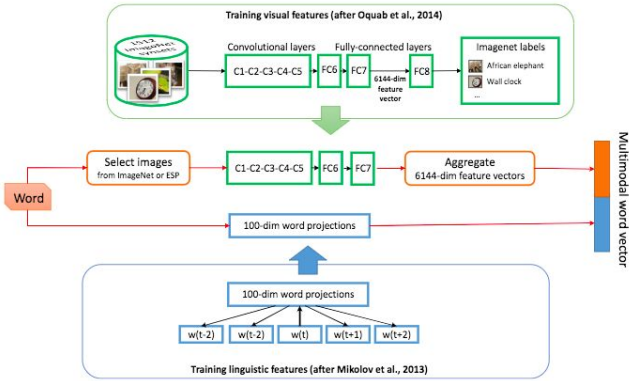
Trained on corpus consisting of:

- Text8 corpus of Wikipedia text (400m words)
- British National Corpus (100m words)

Multi-modal representation



Multi-modal representation



Follow approach used in previous research:

1) Center output vectors

$$C_n = I_n - \frac{1}{n}\mathbb{O}$$

2) L2-normalize vectors

$$|\mathbf{x}| = \sqrt{\sum_{k=1}^{n} |x_k|^2}$$

3) Concatenate linguistic and visual vector

$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \,||\, (1-\alpha) \times \vec{v}_{vis}$$

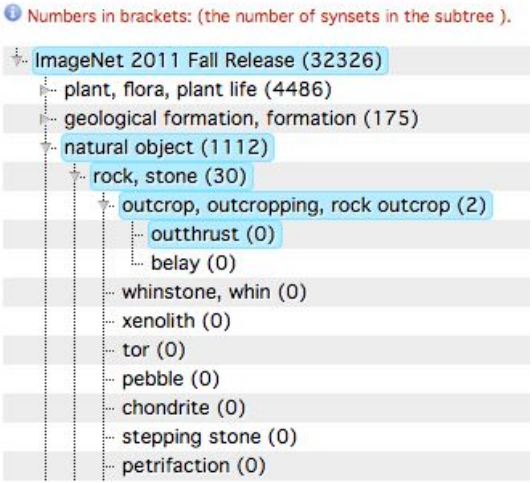
Where α is an optional tuning parameter

Experimental setup

Visual data - dataset

ImageNet:

- 12.5m photos in ~ 22k synsets
- Organized according to WordNet hierarchy
- All photos are manually labeled
- High quality photos w/object usually centered



Experimental setup

Visual data - dataset

ESP Game:

- 100k images covering 20 515 unique words
- Collected by means of a game where two players independently label the photos and have to agree on tags
- Images can contain more than one object and on average contains 14 tags
- No weighting on tags -> can't discern most important image features

Experimental setup

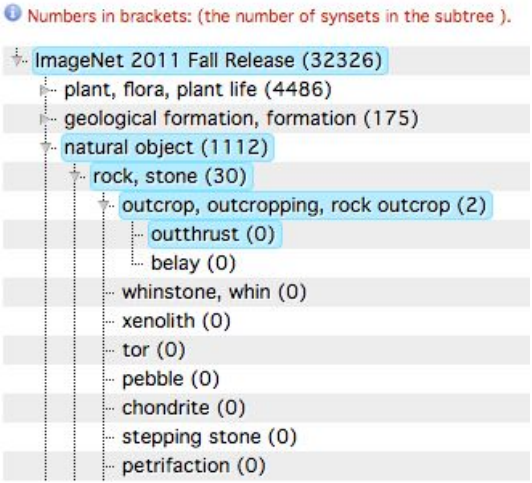
Visual data - image selection

ImageNet - for higher level concepts:

- Sample 1000 images from subtree of concept
- Fallback: sample from subtree of hypernym

ESP game:

- Apply ImageNet logic



Experimental setup

Visual data - image processing

ImageNet:

1) Largest centered square is resampled to form a 256x256 image
2) Crop 16px off all borders to obtain 224x224, then subtract 128 from all image components

$I_{original}$

I_1

I_2

Experimental setup

Visual data - image processing

ESP Game:

1) Scaled to fit inside 224x224 rectangle
2) Centered, added zero padding, and subtract 128 from all image components

$I_{original}$

I_1

I_2

Experimental setup

Visual data - image processing

BOVW features (baseline):

1) Compute Dense Scale Invariant Feature Transform (DSIFT) descriptors
2) Descriptors are then clustered using mini-batch k-means w/100 clusters
3) Each image is then represented by a bag of clusters quantized as 100-dimensional feature vectors
4) ....
5) ....
6) These vectors are then combined into visual concept representations by taking their mean

Experimental setup

Evaluation

Evaluated on two semantic relatedness datasets:

- WordSim353
- MEN

Experimental setup

Evaluation - datasets

WordSim353:

- Most widely used evaluation dataset for distributional semantics
- 353 concept pairs
- Similarity rating provided by human annotators
- Has some idiosyncrasies:
    - Includes named entities such as “OPEC”, “Arafat”
    - Includes abstract words such as “*antecedent*”

Experimental setup

Evaluation - datasets

MEN:

- 3000 word pairs w/751 unique words
- In part designed to:
  - Alleviate issue of uncommon words in WordSim353
  - Be used with ESP Game -> only words w/at least 50 images in ESP Game used

Experimental setup

Evaluation - datasets

In total four evaluation datasets:

- **W353**
- **MEN**
- **W353-relevant** and **MEN-relevant**: Subsets of the full datasets where both words in the concept pair have images in both ImageNet and ESP Game

Evaluated in terms of Spearman p correlation with human-annotated relatedness ratings

Similarity between representations associated with a pair of words is calculated using cosine similarity

$$\cos(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \, \|v_2\|}$$

Results

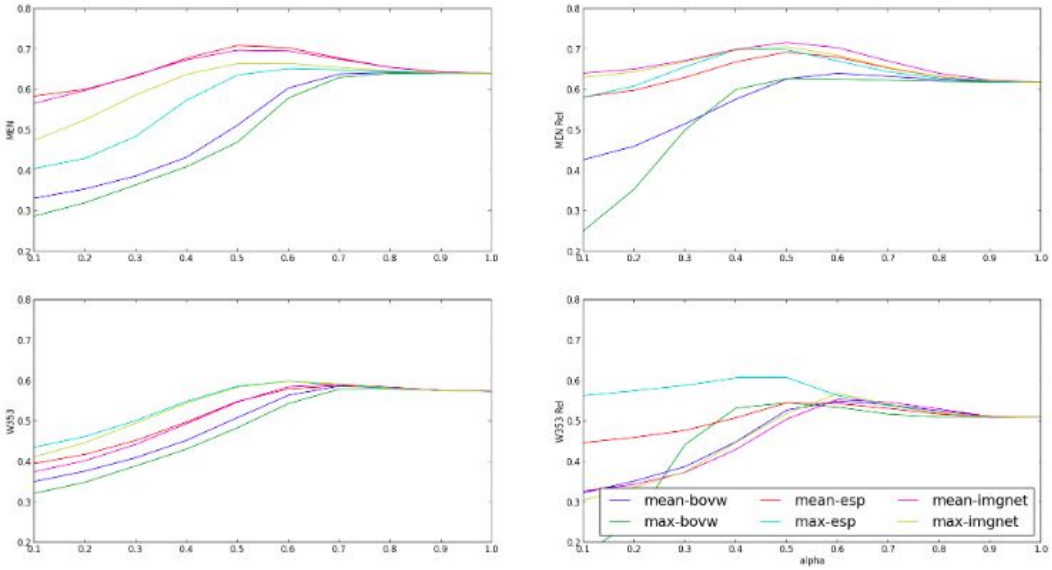
Dataset	Linguistic	Visual			Multi-modal		
		BOVW	CNN-Mean	CNN-Max	BOVW	CNN-Mean	CNN-Max
ImageNet visual features							
MEN	0.64	-	-	-	0.64	**0.70**	0.67
MEN-Relevant	0.62	0.40	0.64	0.63	0.64	**0.72**	0.71
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.30	0.32	0.30	0.55	0.56	**0.57**
ESP game visual features							
MEN	0.64	0.17	0.51	0.20	0.64	**0.71**	0.65
MEN-Relevant	0.62	0.35	0.58	0.57	0.63	0.69	**0.70**
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.38	0.44	0.56	0.52	0.55	**0.61**

Discussion and conclusion

Dataset	Linguistic	Visual			Multi-modal		
		BOVW	CNN-Mean	CNN-Max	BOVW	CNN-Mean	CNN-Max
ImageNet visual features							
MEN	0.64	-	-	-	0.64	**0.70**	0.67
MEN-Relevant	0.62	0.40	0.64	0.63	0.64	**0.72**	0.71
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.30	0.32	0.30	0.55	0.56	**0.57**
ESP game visual features							
MEN	0.64	0.17	0.51	0.20	0.64	**0.71**	0.65
MEN-Relevant	0.62	0.35	0.58	0.57	0.63	0.69	**0.70**
W353	0.57	-	-	-	0.58	0.59	**0.60**
W353-Relevant	0.51	0.38	0.44	0.56	0.52	0.55	**0.61**

- Can performance gain be attributed to multitude of word labels in ESP Game?
- Impact of source dataset?
- Semantic similarity v semantic relatedness
  - WordSim353 captures both similarity and relatedness
  - MEN designed to capture relatedness only

Discussion and conclusion



$$\vec{v}_{concept} = \alpha \times \vec{v}_{ling} \,||\, (1-\alpha) \times \vec{v}_{vis}$$

Discussion and conclusion

W353-Relevant

ImageNet				ESP Game			
word1	word2	system score	gold standard	word1	word2	system score	gold standard
tiger	tiger	1.00	1.00	tiger	tiger	1.00	1.00
man	governor	0.53	0.53	man	governor	0.53	0.53
stock	phone	0.15	0.16	stock	phone	0.15	0.16
football	tennis	0.68	0.66	football	tennis	0.68	0.66
man	woman	0.85	0.83	man	woman	0.85	0.83
cell	phone	0.27	0.78	law	lawyer	0.33	0.84
discovery	space	0.10	0.63	monk	slave	0.58	0.09
closet	clothes	0.22	0.80	gem	jewel	0.41	0.90
king	queen	0.26	0.86	stock	market	0.33	0.81
wood	forest	0.13	0.77	planet	space	0.32	0.79

MEN-Relevant

ImageNet				ESP Game			
word1	word2	system score	gold standard	word1	word2	system score	gold standard
beef	potatoes	0.35	0.35	beef	potatoes	0.35	0.35
art	work	0.35	0.35	art	work	0.35	0.35
grass	stop	0.06	0.06	grass	stop	0.06	0.06
shade	tree	0.45	0.45	shade	tree	0.45	0.45
blonde	rock	0.07	0.07	blonde	rock	0.07	0.07
bread	potatoes	0.88	0.34	bread	dessert	0.78	0.24
fruit	potatoes	0.80	0.26	jacket	shirt	0.89	0.34
dessert	sandwich	0.76	0.23	fruit	nuts	0.88	0.33
pepper	tomato	0.79	0.27	dinner	lunch	0.93	0.37
dessert	tomato	0.66	0.14	dessert	soup	0.81	0.23

Discussion and conclusion

Use CNNs for visual representation, and first multi-modal model that uses deep learning for all input sources

Performance gain on both visual and multi-modal representations over linguistic and BOVW approaches

Approach is robust and works across several datasets with different semantic properties

Gain in multi-modal representation is due to intrinsic information captured in image and not result of accompanying labels

My opinion

Intriguing field of research

They use developed methodology wherever they can

They use several datasets that capture different properties

Discrepancy between dimensionality of input vectors (6144 v 100)

They do (rudimentary) error analysis, but they leave some important and impactful questions to further research

Future research

Include concreteness information or substitute metric such as image dispersion

Jointly learn multi-modal representations

Learn weighting parameters

Examine multi-modal distributional compositional semantics where the multi-modal representations are composed to obtain phrasal representations

Error analysis shows consistency in which words are rated the worst implies linguistic representation might be bad -> explore different ways to represent V_{ling} such as contextualized embeddings or even multilingual embeddings