An old Artificial Intelligence dream that comes true: Merging language and vision modalities

Raffaella Bernardi

University of Trento

May 5th, 2017
Credits

L. Fei Fei, Tamara Berg, Andrej Karpathy, Angeliki Lazaridou, Elia Bruni, Marco Baroni, Chris McCormick
An old AI dream
Words can be represented by vectors harvested from a corpus of texts counting word co-occurrences.
From words to Meaning Representation

Predict the context: Word2Vec (Skip-Gram)

Instead counting words co-occurrences, the vector representing a word can be learned by *predicting* its nearby word. (Mikolov et al, 2013)
Vector Representations

Successful

- Lexical meaning
  - Synonyms
  - Concept categorization (e.g. car ISA vehicle)
  - Selectional preferences (e.g. eat chocolate vs. *eat sympathy)
  - relation classification (exam-anxiety CAUSE-EFFECT relation)
  - salient properties (car-wheels)

- Compositionality: Phrase and Sentence
  - similarity
  - entailment
Grounding language representations into the world. Point to the reference of our mental representation.
From Pixels to Meaning Representation

Gap

- To bridge the gap between pixels and "meaning"

<table>
<thead>
<tr>
<th>What we see</th>
<th>What a computer sees</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="La_Gare_Montparnasse_1895" alt="Image" /></td>
<td>0 3 2 5 4 7 6 9 8</td>
</tr>
<tr>
<td></td>
<td>3 0 1 2 3 4 5 6 7</td>
</tr>
<tr>
<td></td>
<td>2 1 0 3 2 5 4 7 6</td>
</tr>
<tr>
<td></td>
<td>5 2 3 0 1 2 3 4 5</td>
</tr>
<tr>
<td></td>
<td>4 3 2 1 0 3 2 5 4</td>
</tr>
<tr>
<td></td>
<td>7 4 5 2 3 0 1 2 3</td>
</tr>
<tr>
<td></td>
<td>6 5 4 3 2 1 0 3 2</td>
</tr>
<tr>
<td></td>
<td>9 6 7 4 5 2 3 0 1</td>
</tr>
<tr>
<td></td>
<td>8 7 6 5 4 3 2 1 0</td>
</tr>
</tbody>
</table>

Source: S. Narasimhan
From Pixels to Meaning Representation

Challanges
From Pixels to Meaning

Abstract Features

Input

Input space

Feature space

Motorbikes

“Non”-Motorbikes

Feature representation

Learning algorithm
Applications: Traditional CV tasks

Objects

**Image classification**: assigning a label to the image.

![Image of a cow and a car with labels]

**Object localization**: define the location and the category.

![Image of a cow and a car with location and category]

Similarly, scene recognition.
First Revolution: Big dataset

ImageNet

Image database organized according to the WordNet hierarchy. Stanford Vision Lab, Stanford University & Princeton University.

- Challenges: 2007-present
- AMT: 48,940 annotators from 167 countries
- 15M images
- 22K categories of objects
From Pixels to Features

Two methods

- Bag of Visual words (BoVW) (Sivic and Zisserman, 2003)
Keypoints detectors To locate interesting points/content, various kinds of low-level features detectors exists:

- edge detection: the lines we would draw – encode shape info
- corner detection

Local description The identified interesting points are then described: clustered into regions and transformed into vectors representing the region. Several local descriptors exist, e.g:

- Textons (Leung and Malik ’01)
- HoG (Histograms of Oriented Gradients) (Dalal and Triggs ’05)

The low-level features can capture eg. Color, Texture, Shape,

Bag of Visual Words The local descriptions are clustered to obtain the Visual Words that are used to obtain the vector representation of the image.
From Pixels to Features: BoVW

BoVW’s pipeline
Second Revolution: End-to-end systems
Convolutional Neural Networks

ImageNet Classification with Deep Convolutional Neural Networks

- 2012: Krizhevsky outperformed the other systems using CNN
- 2013: half of the systems used CNN
- 2014: All of the systems used CNN.
Deep Learning

- Deep architectures can be representationally efficient.

- Natural progression from low level to high level structures.

- Can share the lower-level representations for multiple tasks.
End-to-end systems

CNN: off-the-shelf vector representation

- Train a CNN on a vision task (e.g. AlexNet on ImageNet)
- Do a forward pass given an image input
- Transfer one or more layers (e.g. FC7 or C5)
Neural Networks

Example to compute a logical operator “Not And”

\[
\begin{array}{ccc}
  x_1 & x_2 & \text{“Not and”} \\
  0 & 0 & 1 \\
  0 & 1 & 1 \\
  1 & 0 & 1 \\
  1 & 1 & 0 \\
\end{array}
\]

Input 00 produces output 1 (since: \((-2) \times 0 + (-2) \times 0 + 3 = 3\) is positive) and similarly, 01 and 10; but the input 11 produces output 0 (since: \((-2) \times 1 + (-2) \times 1 + 3 = -1\) is negative.)
It’s a composition of functions (neurons) that goes from an n-dimensional vector to class scores.

Each neuron receives some inputs, performs a dot product and optionally follows it with a non-linearity. On the last (fully-connected) layer, they have a loss function (e.g., Softmax).
Traditional neural networks cannot use the information “about previous inputs” to inform later ones.

- **Recurrent neural networks (RNNs)** address this issue: They are networks with loops in them, allowing information to persist. They work well with short dependencies.

- **Long Short Term Memory (LSTM)** are a special kind of RNN, capable of learning long-term dependencies.
Language and Vision

Language and Visual Space can be combined!
Applications: Traditional CV tasks
Corpora as KB source: Object recognition


Not a Lemon, it’s more probable a Tennis Ball. Info come from a KB (word similarity list, extracted from internet – Google Sets).
Applications: Traditional CV tasks
Corpora as KB source: Action recognition

Thu Le Dieu, Jasper Uijlings and R. Bernardi (2010, 2011)
Applications: Traditional NLP tasks


<table>
<thead>
<tr>
<th></th>
<th>planet</th>
<th>night</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>moon</td>
<td>10</td>
<td>22</td>
<td>22</td>
<td>0</td>
</tr>
<tr>
<td>sun</td>
<td>14</td>
<td>10</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>dog</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>20</td>
</tr>
</tbody>
</table>
Applications: Traditional NLP tasks

Task 1  Predicting human **semantic relatedness** judgments
        Improved!

Task 2  **Concept categorization**, i.e. grouping words into
classes based on their semantic relatedness (*car* ISA
  *vehicle*; *banana* ISA *fruit*)
        Improved!

Task 3  Find **typical color** of concrete objects (*cardboard is
  brown*; *tomato is red*)
        Improved!

Task 4  Distinguish **literal vs. non-literal** usages of color
  adjectives (*blue uniform* vs. *blue note*)
        Improved!
New Language and Vision Tasks

Zero-shot (or Cross-modal) Mapping: training

“Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world” A. Lazaridou, E. Bruni and M. Baroni (ACL 2015)
New Language and Vision Tasks

Zero-shot (or Cross-modal) Mapping: testing

Step 1 Obtain “parallel data” of linguistic and visual vectors of concepts.
Step 2 Learn a cross-modal mapping between the two semantic spaces
Step 3 Map the unknown concept onto the linguistic/visual space
Step 4 Obtain a label through nearest neighbor search
New Language and Vision tasks

Fast Mapping
New LaVi Applications: Image Captioning

- Approaches: Retrieve vs. Generate
- Frameworks: Pipeline of predictions vs. End-to-end

A man is throwing a frisbee in a park
New LaVi Applications: IC

Approach: Pipeline

E.g., Kulkarni et al. (2011)

![Diagram of LaVi Applications]

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.
New LaVi Applications: IC

Approach: End-to-end

E.g., Karpathy and Fei Fei (2015)

Figure 4. Diagram of our multimodal Recurrent Neural Network generative model. The RNN takes a word, the context from previous time steps and defines a distribution over the next word in the sentence. The RNN is conditioned on the image information at the first time step. START and END are special tokens.
New LaVi Applications: IC

Further info

- **Datasets** Flickr, Pascal, MS-COCO (164K images, 5 captions each)
- **Survey** Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures, Bernardi et al. JAIR 2016
- **Very good talk** by Karpathy (2015):
  https://www.youtube.com/watch?v=ZkY7fAoaNcg
New LaVi Applications: IC

Limitations

- Evaluation Measures: Bleu, Rouge, etc. but not precise.
- No reasoning
New LaVi Applications: VQA


What colour is the moustache made of?

Yellow
New LaVi Applications: VQA Model

\[ v^{i,q} = v^i \circ v^q \]

\[ s = W v^{i,q} + b \]

\[ p_a = \frac{e^{s_a}}{\sum_{a'} e^{s_{a'}}} \]
New LaVi Applications: VQA

Limitations

- Language prior problem: Blind models perform pretty well (50% accuracy on COCO-VQA!).
- Development of synthetic datasets: SHAPES, CLEVR, Yin and Yang.
- Development of new real image datasets: VQA2, FOIL, TDIUC
New LaVi Applications: VQA

Similar images different answers

Who is wearing glasses?

- man
- woman

**task 1:** classification
People riding bicycles down the road approaching a **dog.**

**task 2:** foil word detection
People riding bicycles down the road approaching a **dog.**

**task 3:** foil word correction
People riding bicycles down the road approaching a **bird.**
New LaVi Applications: VQA

Further info

- **Datasets** DAQUAR 2014, COCO-QA, VQA, Visual7W, Visual Genome.
- **Survey** Visual Question Answering: A Survey of Methods and Datasets Wu et ali, (2016)
Other applications

- Spoken VQA (posted on ArXiv on the 1st of May)
- Multimodal Machine Translation
- Image Generation
Can we (linguists) be happy?

Your answer!
de Vries et ali. “GuessWhat?! Visual object discovery through multi-modal dialogue” (2017 arXiv)

Goal of the game: locate an unknown object in a rich image scene by asking a sequence of questions.

Training data: human-playes games: 800K visual QA pairs on 66K images.
New LaVi Applications: IVQA
Interactive Visual Question Answering


Cooperative guessing game: Q-Bot has to select a particular unseen image among several ones that the A-Bot sees. The two agents communicate through NL. Trained on VisDial dataset.
Cutting-edge fancy models’ ingredients
Memory & Attention

- Generative adversarial networks (GAN): two neural networks competing against each other in a game framework.
- Memory & Attention: To focus on some parts of the visual vectors (stored in the memory) e.g. by using the linguistic query to “see” the image.
Attention

**A dog is standing on a hardwood floor.**
Pay attention to those sets! Learning quantification from images Sorodoc et. al. (Submitted)

Query: ___ fish are red.
Answers: (a) All, (b) Most, (c) Some, (d) Few, (e) No.
Datasets: Q-COCO

<table>
<thead>
<tr>
<th>ORIGINAL IMAGE</th>
<th>SCENARIO</th>
<th>ANNOTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>banana: healthy</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>orange: fresh, tasty/delicious</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>orange: healthy, tasty/delicious, appetizing, fresh, round</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>orange: tasty/delicious, appetizing, fresh, cooked</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>banana: laying, healthy, tasty/delicious, horizontal, fresh, whole</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>orange: laying, round, fresh, appetizing tasty/delicious, whole, healthy</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>orange: tasty/delicious, fresh</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>orange: fresh</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENERATED QUERIES</th>
<th>PROPORTION</th>
<th>GROUND-TRUE ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ___ oranges are fresh</td>
<td>100%</td>
<td>all</td>
</tr>
<tr>
<td>2. ___ oranges are whole</td>
<td>16.7%</td>
<td>few</td>
</tr>
<tr>
<td>3. ___ oranges are healthy</td>
<td>33.3%</td>
<td>some</td>
</tr>
<tr>
<td>4. ___ oranges are tasty/delicious</td>
<td>83.3%</td>
<td>most</td>
</tr>
<tr>
<td>5. ___ oranges are horizontal</td>
<td>0%</td>
<td>no</td>
</tr>
</tbody>
</table>
Datasets: Q-ImageNet

<table>
<thead>
<tr>
<th>ORIGINAL IMAGE</th>
<th>SCENARIO</th>
<th>ANNOTATION</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image 1" /></td>
<td><img src="image2.png" alt="Image 2" /></td>
<td><strong>dog</strong>: furry, black</td>
</tr>
<tr>
<td><img src="image3.png" alt="Image 3" /></td>
<td><img src="image4.png" alt="Image 4" /></td>
<td><strong>dog</strong>: furry, black, smooth</td>
</tr>
<tr>
<td><img src="image5.png" alt="Image 5" /></td>
<td><img src="image6.png" alt="Image 6" /></td>
<td><strong>rabbit</strong>: furry, white, brown</td>
</tr>
<tr>
<td><img src="image7.png" alt="Image 7" /></td>
<td><img src="image8.png" alt="Image 8" /></td>
<td><strong>dog</strong>: furry, black, brown, smooth</td>
</tr>
<tr>
<td></td>
<td><img src="image9.png" alt="Image 9" /></td>
<td><strong>dog</strong>: furry, black, gray</td>
</tr>
<tr>
<td></td>
<td><img src="image10.png" alt="Image 10" /></td>
<td><strong>hoop</strong>: white, red, round</td>
</tr>
<tr>
<td></td>
<td><img src="image11.png" alt="Image 11" /></td>
<td><strong>dog</strong>: black, white</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GENERATED QUERIES</th>
<th>PROPORTION</th>
<th>GROUND-TRUTH ANSWER</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ____ dogs are black</td>
<td>100%</td>
<td>all</td>
</tr>
<tr>
<td>2. ____ dogs are white</td>
<td>16.7%</td>
<td>few</td>
</tr>
<tr>
<td>3. ____ dogs are smooth</td>
<td>33.3%</td>
<td>some</td>
</tr>
<tr>
<td>4. ____ dogs are furry</td>
<td>83.3%</td>
<td>most</td>
</tr>
<tr>
<td>5. ____ dogs are red</td>
<td>0%</td>
<td>no</td>
</tr>
</tbody>
</table>
Sequential Processing

CNN+LSTM model

400 dimension visual vectors

LSTM Cell

LSTM Cell

LSTM Cell

LSTM Cell

Gist 1

Concatenation of Gists

Gist 2

'DBlind' LSTM

Dog embedding : 400

Black embedding : 400

Quantifier Prediction
Attention Mechanism: SAN’s attention layer

Yang, Z., et al. (CVPR 2016). Stacked attention networks (SAN) for image question answering.
Linguistically motivated NNs with stacked attention
Conclusion

- Impressive progress
- Hard but fun to learn
- A land of new ideas can be explored

My wish:

Combine language (pragmatics) with vision.
Other Useful Links

Neural Networks

- http://info.usherbrooke.ca/hlarochelle/neural_networks/content.html
- http://www.iro.umontreal.ca/~bengioy/dlbook/
- Blog posts: http://colah.github.io/
Other Useful Links

Language and Vision

- Describing Images in Sentences by Julia Hockenmaier
  http://nlp.cs.illinois.edu/HockenmaierGroup/EACLTutorial2014/index.html

- Vision and Language Summer Schools: 2nd edition 2016 (Malta). COST-ACTION.

- “Multimodal Learning and Reasoning”, Desmond Elliott, Douwe Kielay, and Angeliki Lazaridou (Tutorial at ACL 2016)


- “How we teach computers to understand pictures” TED Talk by Fei Fei Li.
Language and Vision Research Groups

- Stanford Vision Lab – Le Fei Fei http://vision.stanford.edu/
- University of North Carolina – Tamara Berg http://www.tamaraberg.com/
- Virginia University – Devi Parikh https://filebox.ece.vt.edu/~parikh/CVL.html
- Edinburgh University (M. Lapata, F. Keller)
- Facebook
- Google DeepMind
- More on the iV&L Net Cost Action http://www.cost.eu/COST_Actions/ict/Actions/IC1307
The team@UniTN

Ionut

Sandro

Ravi

Aurelie

me