Learning Multimodal Word Representations: Language Grounding in Visual Context and Visual Question Answering

12-12-2017
Patrick Gallinari
Université Pierre et Marie Curie – Paris 6, France
patrick.gallinari@lip6.fr
General context

• Multimodal machine learning with text and image sources
  • Complementary role of text and image
  • Improve language/image processing tasks
General context

• Distributional hypothesis
  • Text
    • Linguistic items with similar context distributions should have similar meanings
    • Meaning is represented by a vector summarizing the context distribution → semantic space
      • Similarity measures in the semantic space represents word meaning similarity
    • Meaning is entirely defined by co-occurrence patterns and not by real world situations – similar to AI symbol grounding problem
    • Bruni et al 2012 -> according to semantic similarity sky is green, flour is black, etc
  • Image
    • Object recognition methods extract vector based representations from natural images
      • Bag of visual words (image patches)
      • Convolutional Neural Networks introduce hierarchical visual vectors representations of images
    • Distributional hypothesis for images (e.g. Bruni et al 2012)
      • Semantically similar objects tend to occur in similar environments in images
      • Importance of the visual context
General context: examples

• Grounded Language Model
  • Caption generation: text generation learned from image representations

• Visual QA: classification task, classifiers learned from joint image + query text representation

Fig. Vinyals et al. 05

VQA Real

Visual Genome
General context: examples

- **Visual Common Sense**
  - Modeling text-image bias
    - Frequency with which one refers to things or actions does not correspond to real world frequencies.
    - People tend to eliminate common things
    - Introduces a bias of language wrt the « real world frequencies »
    - Both textual and image data have biases about the information they encode about context
      - Motivates multi-modal approaches for learning word representations

Fig. from Misra et al. 2016
Illustrating human reporting bias for image annotation
General context: example

- Cross modal mapping (here image to text)
  - Objective
    - Learn a textual description of an image
      - i.e. using an image as input, generate a sentence that describes the objects and their relation!
  - Neural image caption generator (Vinyals et al. 2015)
    - Inspired by a translation approach but the input is an image
      - Use a RNN to generate the textual description, word by word, provided a learned description of an image via a deep CNN
    - Makes use of attention mechanism

Figure from Vinyals et al. 2015
Background - Models – CNNs for classification and segmentation

- CNNs state of the art models for image labeling and segmentation

- Layers build increasingly abstract vector representations of input image

- Learned representations on large datasets (Imagenet) may be used for other tasks or other datasets
Background- - Models – CNNs for classification and segmentation (Durand 2016)
Background - - Recurrent neural networks

Language models

• Objective:
  • Probability models of sequences \((x^1, x^2, ..., x^t)\)
  • Items may be words or characters
  • Estimate:
    • \(p(x^t|x^{t-1}, ..., x^1)\)

• Example
  • « S’il te plaît... dessine-moi ...» what next ?
  • « \(x^1x^2x^3 \ldots \ldots \ldots \ldots \ldots \ldots x^{t-1} \ldots \) » what is \(x^t\)?

\[
\hat{y}^t = g(Vs^t) \\
\]

\[
s^t = f(Ws^{t-1} + Ux^t)
\]

\[
\]
Background- - Language models – example
(Karpathy 2015- https://karpathy.github.io/2015/05/21/rnneffectiveness/)

• Training on Tolstoy’s War and Peace a character language model
  • Stacked recurrent networks (LSTM)

```
tyntd-iafhatawiaoihrdemot lytdw  e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e plia tklrgd t o idoe ns,smtt h ne etie h,hregtrs nigtike,aoaenns lng
```

```
"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwy fil on aseterlome
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort
how, and Gogition is so overetical and ofter.
```

```
"Why do what that day," replied Natasha, and wishing to himself the fact the
princess, Princess Mary was easier, fed in had oftened him.
Pierre aking his soul came to the packs and drove up his father-in-law women.
```
Background: Learning Grounded word representation

- **Joint models**
  - Learn jointly the representations from visual and textual inputs
    - LDA (a latent variable generates the two modalities)
    - Auto-encoders: 1 for each modality + a joint representation used e.g. for classification
    - Extensions of Word2Vec: e.g. perceptual information about a concrete concept is introduced in the text model whenever it is encountered in text (e.g. Lazaridou et al. 2015)

- **Sequential models**
  - Learn separately text and image representations
  - Combine them
    - Middle fusion: e.g. concatenation of text and image representations, projections (e.g. SVD) or linear combinations of such concatenations
    - Late fusion: score combination of visual and textual representations, score are computed for a task (e.g. classification)
Objective

• Context
  • Language is ambiguous, biased and lacks common sense
  • Images are unequivocal depictions
  • Human meaning representation is grounded in physical reality and sensorimotor experience

• Traditional Distributional Semantic models lack common sense
  • Word meaning link to physical world
  • Common sense
    • e.g. How do we know
      • the different functions of objects – bowls can hold soup, knifes are used to cut meat
      • the relations between objects (spatial, temporal), …

• Question
  • Can language be grounded in the visual world?
    • Long term objective
  • Intermediate objective
    • Learn grounded representations for words
Importance of context

• Multimodal models show the complementarity of text and language
• They rely on direct image features + text alignment
• They ignore visual context
• Context brings a lot of information
  • How it is used, where it can be found, etc

Approach
• Hypothesis: visual context enhances the learned representation of words
• Ground words in visual context
• Without the need for text-image alignment
  • Trained on independent text / image corpora
  • Requires only that the 2 corpora share common entities
  • But contexts are different

Figure 3: Image of complex scene with lots of visual context for all objects
Background: Learning word vector representations
Language model/ Skip Gram model
(Mikolov et al. 2013)

• Predict the context word probability for a word in a sequence: \( p(w_{\text{context}} | w_{\text{input}}) \)

\[
p(w_{\text{context}} | w_{\text{input}}) = \frac{\exp(v_{w_{\text{context}}} \cdot v_{w_{\text{input}}})}{\sum_{w=1}^{V} \exp(v_{w} \cdot v_{w_{\text{input}}})}
\]

• \( v_{w} \) is the learned representation of the \( w \) vector (the hidden layer),
  \( v_{w_{\text{context}}} \cdot v_{w_{\text{input}}} \) is a dot product and \( V \) is the vocabulary size
Background: Learning word vector representations

Language model/ Skip Gram model
(Mikolov et al. 2013)

• Loss function: average log probability
  • \[ L = \frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t) \]
  • \( T \) is the number of words in the whole sequence used for training (roughly number of words in the corpus) and \( c \) is the context size

• In practice: simplified version of negative sampling
  • \[ l(w_{input}, w_{context}) = \log \sigma \left( v_{w_{context}} \cdot v_{w_{input}} \right) + \sum_{i=1}^{k} \log \sigma \left( -v_{w_i} \cdot v_{w_{input}} \right) \]
  • With \( w_i \) a negative example sampled from the word distribution and \( \sigma(x) = \frac{1}{1+\exp(-x)} \)
Related work: Multi-modal Skip-Gram - Lazaridou et al. 2015

- Extension of skip-gram for multi-modal data
  - For a subset of the corpus words, textual corpus concepts are linked to visual representation of concepts
    - Representation of words corresponding to image entities are forced to be close to the visual pretrained representation of the corresponding object
    - Joint learning, No context

Figure from Lazaridou et al. 2015
- Left: text Skip-Gram
- Right: visual component
  - visual vectors obtained from the upper layer of a CNN for 5000 words occurring > 500 times in Imagenet (about 5% tokens in the text corpus)
  - Visual word: average vector learned from 100 samples of this item (e.g. cat)
  - M matrix: maps the textual representation to the visual one for this subset of words
Grounding words in visual context: research questions

- Research questions
  - RQ1: what is a visual context, what is a context model?
  - RQ2: how to learn joint representations from texts and images using the context
  - RQ3: how can we evaluate the contribution of the visual modality to the final embeddings?

Figure 1: Illustrative figure of the 3 research questions raised in our paper. RQ1: How to define a visual context? RQ2: How to integrate visual context use in a multimodal model? RQ3: How to evaluate the visual contribution to learned embeddings.
Context in visual data (RQ1)

- **Entity** $e$: object in the image
- **Context element** $c$: different possibilities corresponding to different « supervision » levels
  - High level context: surrounding objects
    - representation: learned embedding
  - Low level context: image patches
    - Full image with entity masked or random patch chosen around the entity
    - Representation: CNN encoding
  - Enhancing via spatial information
    - If entity localization in images is available, this information could be incorporated in the context representation
      - e.g. relative position of the bounding boxes of the object and the context
Model/ Loss functions (RQ2)

• Image context loss function
  
  \[ L_{Image} = \sum_{e \in D} \sum_{c \in C_e} \left( \log \sigma(f_\theta(c). t_e) + \sum_{c^-} \log \sigma(-f_\theta(c). t_e) \right) \]

  • \( f_\theta(c) \in R^d \): representation of context \( c \), \( D \) set of entities, \( t_e \) shared entity embedding
  • Similar to Skip-gram loss, except for functional \( f_\theta(c) \)

• Multimodal loss function
  
  \[ L(T, U, \theta) = L_{text}(T, U) + \alpha L_{image}(T, \theta) \]

  • \( T \): shared multi-modal text embeddings
  • \( U \): textual context parameters
  • \( \theta \): visual context parameters
Experiments/ Tasks (RQ3)

• No direct evaluation - like for all entity learning methods

• Word pairs similarity
  • Datasets: human judgements on word pairs similarity
  • Compare human similarity scores with cosine similarity of learned entity representation
  • Measure: Spearman coefficient between the two lists of similarities

• Feature norm prediction
  • Feature norms are lists of attributes for an object
  • Predict object attributes from their representation (e.g. is_red, can_fly) (McRae et al. 2005: properties concrete nouns)
  • 43 characteristics grouped in 9 categories, 417 entities
  • Linear SVM classifier for the 9 categories on entity representation

• Concreteness prediction
  • Predict word concreteness score (gold ratings provided on 3260 words)
  • SVM regression on entity representation
Experiments/ Datasets

• Text
  • Dump of Wikipedia
    • 4.2 M articles, 2.1 unique words

• Visual data
  • Visual Genome (Krishna et al 2017)
  • 108 k images, 4842 unique entities, complex scenes, 31 object instances per image on average

Fig: Krishna et al 2017
Experiments/ Models

• Baselines
  • Text alone
  • Text + direct visual features from objects
  • Text + visual context (sequential models)

• Model
  • Objects
  • Patches (Full image, random patches)
  • Spatial information
  • Ensemble model: context + direct visual features
Experiments/ results

• Example results for RQ2

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Text</th>
<th>Similarity Evaluation</th>
<th>Feature-norm Prediction Task</th>
<th>Conc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>O + T</td>
<td>48 60 33 69 63</td>
<td>58 52 44 79 62 11 32 54 60</td>
<td>42.1</td>
</tr>
<tr>
<td>Joint</td>
<td>L + T</td>
<td>52 65 34 71 65</td>
<td>61 55 42 80 59 11 31 54 62</td>
<td>43.4</td>
</tr>
<tr>
<td>Objects</td>
<td>O + T</td>
<td>53 66 35 <strong>75 67</strong></td>
<td>62 55 46 <strong>82</strong> 61 13 33 55 61</td>
<td>42.9</td>
</tr>
<tr>
<td>Patches</td>
<td>P + T</td>
<td>53 65 35 72 <strong>67</strong></td>
<td>60 56 49 <strong>82</strong> 60 12 32 55 61</td>
<td>43.1</td>
</tr>
<tr>
<td>Pfull + T</td>
<td></td>
<td>53 65 34 73 65</td>
<td>60 55 44 <strong>82 63</strong> 14 32 55 59</td>
<td>43.2</td>
</tr>
<tr>
<td>Our models</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spatial</td>
<td>Sp(O, δ, ⊕) + T</td>
<td>52 66 36 73 64</td>
<td><strong>64 59</strong> 46 81 62 06 31 <strong>57 63</strong></td>
<td>42.5</td>
</tr>
<tr>
<td></td>
<td>Sp(O, c, ⊕) + T</td>
<td>54 66 35 72 64</td>
<td>62 56 <strong>52</strong> 80 61 13 <strong>34 57</strong> 58</td>
<td>43.7</td>
</tr>
<tr>
<td></td>
<td>Sp(O, δ, b) + T</td>
<td>54 <strong>68 38</strong> 73 66</td>
<td>63 56 48 81 60 13 32 <strong>56 63</strong></td>
<td>42.5</td>
</tr>
<tr>
<td></td>
<td>Sp(O, c, b) + T</td>
<td>55 67 34 <strong>75 64</strong></td>
<td>61 58 46 80 <strong>63 15 34 57</strong> 62</td>
<td><strong>44.4</strong></td>
</tr>
<tr>
<td>Ensemble</td>
<td>L + O + T</td>
<td>54 66 35 <strong>75 63</strong></td>
<td>63 55 50 <strong>82 60</strong> 10 33 55 59</td>
<td>43.9</td>
</tr>
</tbody>
</table>

Table 2: RQ2 experimental results on word similarity evaluation benchmarks, feature-norm prediction task, concreteness prediction task (Conc.). Concreteness measures are coefficients of determination ($R^2$) multiplied by 100.
Experiments/ results

• Main results on all three tasks/ RQ
  • Multi-modal embeddings > Text only embeddings
  • Surroundings > Direct features (word sim. task)
  • Ensemble model: surrounding and direct features complementary
  • Spatial information useful
  • High-level context > Low-level context
Multimodal Tucker Fusion for Visual Question Answering
H. Ben Younes, R. Cadene, M. Cord, N. Thome, ICCV 2017
Visual Question Answering

• Classification using multimodal inputs
• several datasets
Vanilla VQA Scheme

学习多模态词表征
VQA: Multimodal fusion strategies

• Linear models

Concatenation & projection: \( y = W \begin{bmatrix} q \\ v \end{bmatrix} \)

Element-wise sum: \( y = (Wq) + (Vv) \)

Element-wise product: \( y = (Wq) \odot (Vv) \)

Multi-layer perceptron: \( y = MLP \left( \begin{bmatrix} q \\ v \end{bmatrix} \right) \)

Works best among linear methods
VQA: multimodal fusion strategies

- Bilinear models: extract more relevant correlations between the two modes

\[
y^k = \sum_{i=1}^{d_q} \sum_{j=1}^{d_v} T^{ijk} q^i v^j
\]

- Full tensor: \( d_q = d_v = 2048 \) and 2000 classes, number of free parameters in \( T \approx 10^{10} \)
- Reduce the size of the tensor
VQA: multimodal fusion strategies

- Reducing tensor size
  - Tucker based decomposition + sparsity constraints on the core tensor $T_c$

- All the parameters ($W_v, W_q, W_o, T_c$) are learned end to end
VQA: attention process
VQA: attention process
Attention mechanism

What is sitting on the desk in front of the boys?

What are on the shelves in the background?

Tucker Decomposition with Structured Sparsity

Laptops

Books
VQA and reasoning

Many initiatives to improve datasets and evaluate reasoning as:

VQA v2.0 dataset and challenge 2017

[CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, Ross Girshick, CVPR 2017]
- Questions testing various aspects of visual reasoning including attribute identification, counting, comparison, spatial relationships, and logical operations.

Figure 1: Examples from our balanced VQA dataset.