# Understanding Visual Scences

## Dependency Graphs, Word Senses, and Multimodal Embeddings

Mirella Lapata
School of Informatics
University of Edinburgh

# Joint Work with

Carina Silberer

Spandana Gella

Frank Keller

Jasper Uijilings

# Structure in Multimodal Processing

Lots of recent in work on multimodal processing:

- image description generation;
- visual question answering;
- multimodal machine translation;
- video summarization.

# Structure in Multimodal Processing

Lots of recent in work on multimodal processing:

- image description generation;
- visual question answering;
- multimodal machine translation;
- video summarization.

We need to understand the meaning of images and text:
Who does what to whom?

# Structure in Multimodal Processing

Lots of recent in work on multimodal processing:

- image description generation;
- visual question answering;
- multimodal machine translation;
- video summarization.

We need to understand the meaning of images and text:
Who does what to whom?

Understanding requires structure, not just an unordered set of labels:

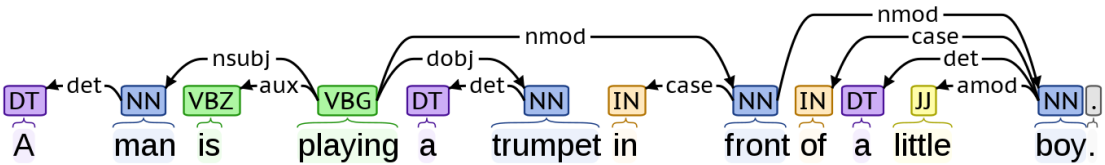
- linguistic structure;
- image structure.

# Structure in Multimodal Processing

A man is playing a trumpet in front of a little boy.

# Linguistic Structure

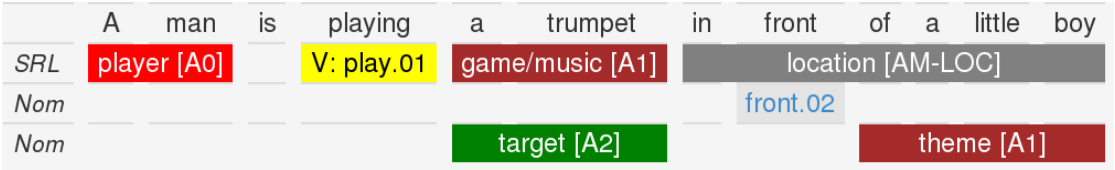
Output of dependency parser (with PoS labels):




http://nlp.stanford.edu:8080/corenlp/process

# Linguistic Structure

Output of a semantic role labeler (with word senses):




http://cogcomp.cs.illinois.edu/page/demo_view/srl

# Structure in Multimodal Processing

Linguistic structure:

- discrete base units (words), ordered in 1D;
- span-based labels (e.g., PoS, phrases);
- tree-based hierarchies;
- clear distinction between syntax and semantics;
- canonical representations defined by linguistic theory.

# Structure in Multimodal Processing

Linguistic structure:

- discrete base units (words), ordered in 1D;
- span-based labels (e.g., PoS, phrases);
- tree-based hierarchies;
- clear distinction between syntax and semantics;
- canonical representations defined by linguistic theory.

Now let's compare this to image structure.

# Image Structure

Output of an image labeler:

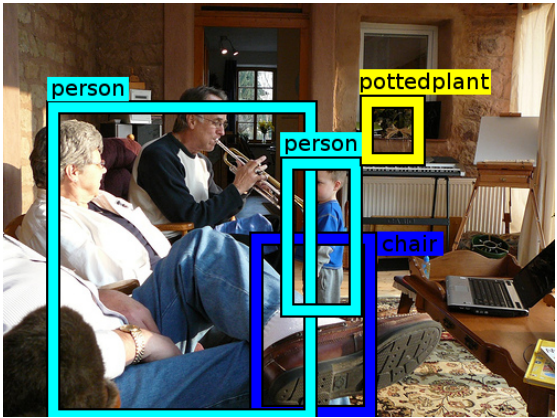
people
man
furniture
room
adult
group
child
indoors
woman
seat
family
music
education
sit

https://www.clarifai.com/demo

We could also label: attributes, scene type, colors, textures, etc.

# Image Structure

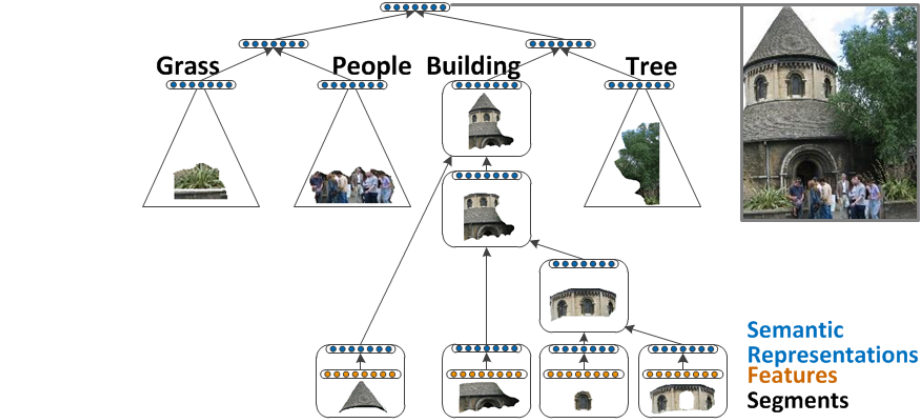
Output of an object recognizer:




Output of FastRCNN model with AlexNet architecture trained on PASCAL VOC 2007.

# Image Structure

Hierarchical segmentation (indicates part-whole relationships):




http://www.socher.org/index.php/Main/ParsingNaturalScenesAndNaturalLanguageWithRecursiveNeuralNetworks

# Structure in Multimodal Processing

Linguistic structure:

- discrete base units (words), ordered in 1D;
- span-based labels (e.g., PoS, phrases);
- tree-based hierarchies;
- clear distinction between syntax and semantics;
- canonical representations defined by linguistic theory.

# Structure in Multimodal Processing

Linguistic structure:

- discrete base units (words), ordered in 1D;
- span-based labels (e.g., PoS, phrases);
- tree-based hierarchies;
- clear distinction between syntax and semantics;
- canonical representations defined by linguistic theory.

Image structure:

- continuous base units (pixels), ordered in 2D;
- region-based labels (e.g., objects, attributes);
- part–whole structure;
- no clear distinction between syntax and semantics;
- no “correct” canonical representations.

# Representational Divergence

Representational divergence: for multimodal processing, we need to fuse linguistic and image structures, but they are very different.

# Representational Divergence

Representational divergence: for multimodal processing, we need to fuse linguistic and image structures, but they are very different.

Hypothesis: We need to align visual representations.

Two examples in this talk:

- visual dependency representations;
- visual sense disambiguation.

# Spatial Relations

We need a grammar that defines the relations between the objects in an image: Visual Dependency Grammar (Elliott & Keller 2013).

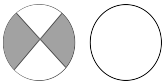
It assumes eight relations that can hold between pairs of objects, based on three geometric properties:

- pixel overlap;
- angle between objects;
- distance between objects.

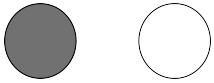
# Spatial Relations

X *on* Y

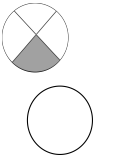
X *surrounds* Y



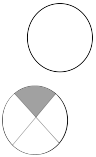
X *beside* Y



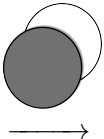
X *opposite* Y



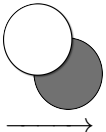
X *above* Y



X *below* Y



X *infront* Y



X *behind* Y

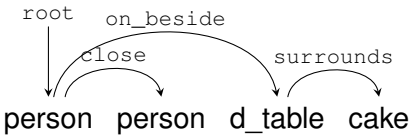
# Visual Tuples

An image represented a bag of VDR tuples (Ortiz et al., 2015).

person `close` person
person `on_beside` d_table
d_table `surrounds` cake
person `near` cake
person `close` d_table
person `above_close` cake

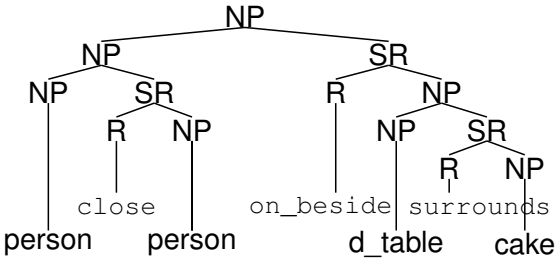
# Visual Dependency Representations

An image is represented as a dependency tree (Silberer et al., 2017).

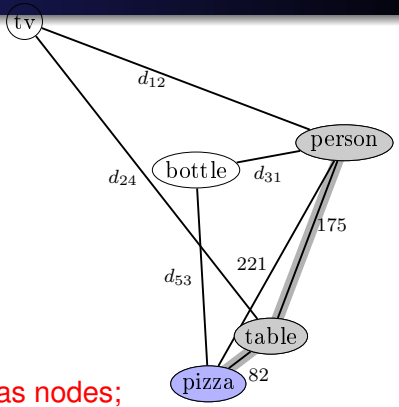


# Visual Constituency Representations

An image is represented as a constituency tree (Silberer et al., 2017).

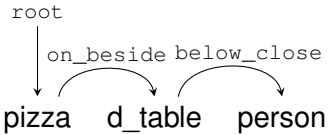


# Tree Construction



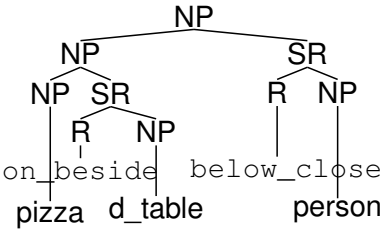

- Build a fully connected graph with all objects as nodes;
- edge weights correspond to spatial distance;
- minimum spanning tree (MST): visual dependency representation;
- use grammar to generate visual constituency representation.

# Tree Construction




- Build a fully connected graph with all objects as nodes;
- edge weights correspond to spatial distance;
- minimum spanning tree (MST): visual dependency representation;
- use grammar to generate visual constituency representation.

# Tree Construction




- Build a fully connected graph with all objects as nodes;
- edge weights correspond to spatial distance;
- minimum spanning tree (MST): visual dependency representation;
- use grammar to generate visual constituency representation.

# Image Description Generation via Machine Translation

- Repurpose existing NLP technology to construct visual representations;
- use machine translation models: focus on tree-to-string translation;

# Image Description Generation via Machine Translation

- Repurpose existing NLP technology to construct visual representations;
- use machine translation models: focus on tree-to-string translation;
- trees are task-independent, do not take descriptions into account:

# Image Description Generation via Machine Translation

- Repurpose existing NLP technology to construct visual representations;
- use machine translation models: focus on tree-to-string translation;
- trees are task-independent, do not take descriptions into account: **create parallel corpus of trees with multiple descriptions;**

# Image Description Generation via Machine Translation

- Repurpose existing NLP technology to construct visual representations;
- use machine translation models: focus on tree-to-string translation;
- trees are task-independent, do not take descriptions into account: **create parallel corpus of trees with multiple descriptions;**
- translation is loose: not all visual objects are verbalized; multiple descriptions can focus different aspects of a scene:

# Image Description Generation via Machine Translation

- Repurpose existing NLP technology to construct visual representations;
- use machine translation models: focus on tree-to-string translation;
- trees are task-independent, do not take descriptions into account:
  **create parallel corpus of trees with multiple descriptions;**
- translation is loose: not all visual objects are verbalized; multiple descriptions can focus different aspects of a scene:
  **generation model performs content selection.**

# Parallel Corpus Creation

**Step 1: Grounding objects to linguistic expressions.**

person
d_table
person
cake
plate
cup

*Little kids sitting around a table that has a birthday cake on it.*
*A group of young children standing around a cake.*

# Parallel Corpus Creation

**Step 1: Grounding objects to linguistic expressions.**

person
d_table
person
cake
plate
cup

*[Little kids]*$_{A1}$ *sitting*$_{sit.01}$ *[around a table]*$_{A2}$ *that has*$_{has.01}$ *[a birthday cake]*$_{A2}$ *on it.*
*[A group of young children]*$_{A1}$ *standing*$_{stand.01}$ *[around a cake]*$_{A2}$*.*

# Parallel Corpus Creation

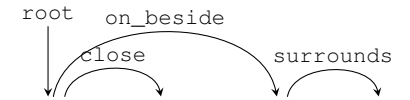
**Step 1: Grounding objects to linguistic expressions.**

person
d_table
person
cake
plate
cup

*[Little kids]*$_{A1}$ *sitting*$_{sit.01}$ *[around a table]*$_{A2}$ *that* *has*$_{has.01}$ *[a birthday cake]*$_{A2}$ *on it.*
*[A group of young children]*$_{A1}$ *standing*$_{stand.01}$ *[around a cake]*$_{A2}$*.*

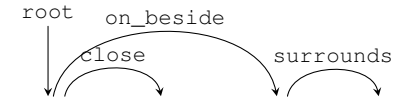
# Parallel Corpus Creation

**Step 2: Render scenes as trees and generate corpus.**



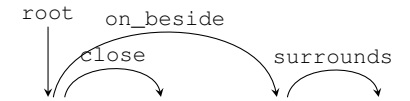

Kids sitting around a table.

---




A table that has a birthday cake.

---




Children standing around a cake.

# MT Model: Surface Realization

We train a translation model on our parallel corpus using the MT framework implemented in Moses (Koehn et al., 2007):

$$t^* = \arg\max_t P(t|s)$$

$$P(t|s) = \arg\max_d \left( \sum_{k=1}^{K} \lambda_k h_k(d) \right)$$

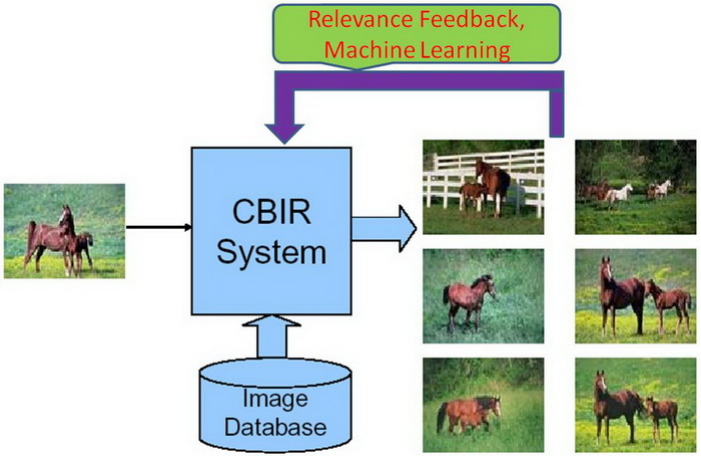
- $d \in D(s,t)$ are derivations in a synchronous grammar;
- $h_k$ feature functions (language model, translation table, word penalty model);
- constants $\lambda_k$ scale different models, tuned during training.

# MT Model: Content Selection

At test time we must decide which objects to talk about:

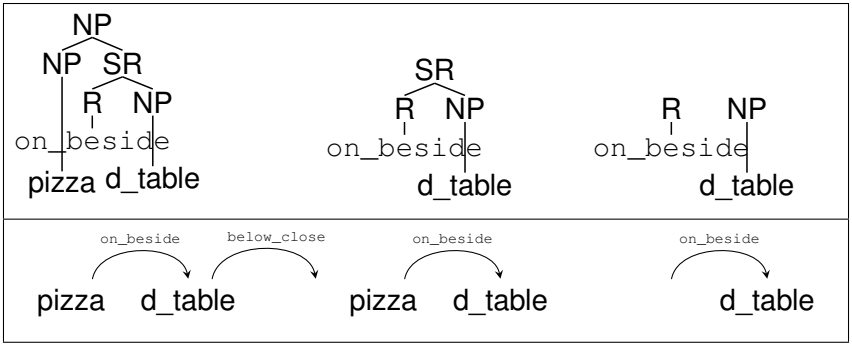
- predict whether a detected object is relevant for scene;
- we use logistic regression with $l_2$ regularization;
- trained on positive and negative instances;
- positives: objects aligned to SRL arguments;
- negatives: unaligned objects;
- features: object detection score, relative size, relative distance between two objects, object occurrences, spatial features.

# Query-by-Example Image Retrieval

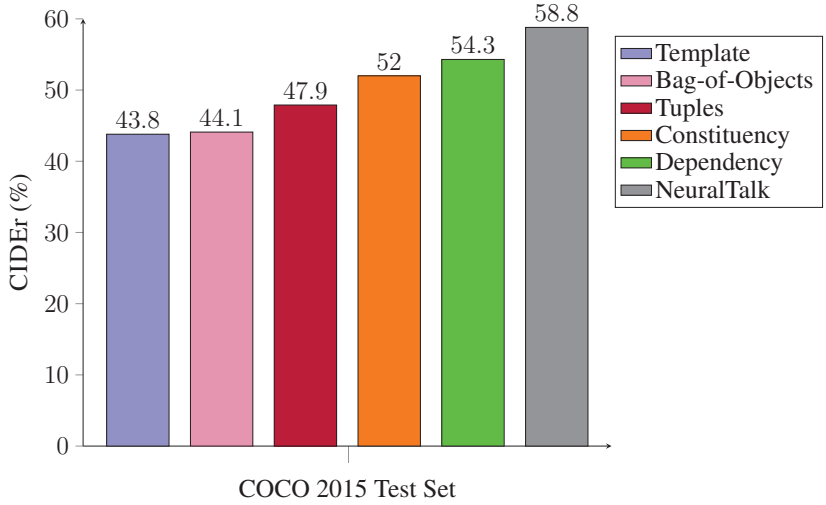


# Query-by-Example Image Retrieval

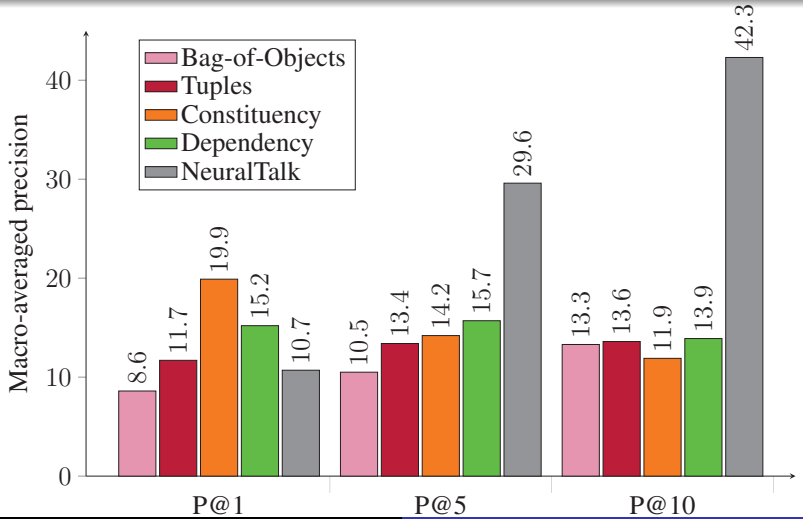
- Let $\mathcal{I}$ denote an image collection;
- for every image $q$ produce a ranking in order of similarity to $q$;
- subtree kernels measure similarity of constituent trees;
- partial tree kernels measure similarly of dependency trees.



# Results: Image Description Generation



# Results: Image Retrieval



# Example Output

| | | |
|---|---|---|
| Template | 5) a couch has a couch | 2) an airplane is near a car |
| Tuples | 4) the room has a couch | 5) a airplane sitting on a street |
| Dependency | 1) a dog sitting on a couch | 3) a airplane parked next to a car |
| Constituency | 2) dog laying on a couch | 4) a airplane parked next to a car |
| Human | 3) a dog is looking at something | 1) a large plane with a red tail |

# Aligning Actions and Verbs

So far, we have looked at syntactic structure only: how do the objects in an image relate to each other.

To really understand the content of an image, we need semantics: represent the event depicted, its participants, and the roles they play.

We can achieve this using verb senses:

- well established in linguistics (e.g., WordNet);
- more general that the action labels used in computer vision;
- can be aligned with both sentences and images.

# Word Sense Disambiguation

Word sense disambiguation is a standard NLP task:

(1) A man is **playing** a guitar.

(2) The children are **playing** across the street.

(3) Two men **playing** doubles tennis on a grass court.

# Word Sense Disambiguation

Word sense disambiguation is a standard NLP task:

(1) A man is **playing** a guitar.

play:1 perform music on musical instrument

(2) The children are **playing** across the street.

(3) Two men **playing** doubles tennis on a grass court.

# Word Sense Disambiguation

Word sense disambiguation is a standard NLP task:

(1) A man is **playing** a guitar.
play:1 perform music on musical instrument

(2) The children are **playing** across the street.
play:2 engage in a fun or recreational (childlike) activity

(3) Two men **playing** doubles tennis on a grass court.

# Word Sense Disambiguation

Word sense disambiguation is a standard NLP task:

(1) A man is **playing** a guitar.
play:1 perform music on musical instrument

(2) The children are **playing** across the street.
play:2 engage in a fun or recreational (childlike) activity

(3) Two men **playing** doubles tennis on a grass court.
play:3 engage in or make moves related to competition or sport

# Visual Sense Disambiguation

We can apply this task to an image/verb pair:

**play**

# Visual Sense Disambiguation

We can apply this task to an image/verb pair:

**play**

play:1 perform music on musical instrument

New task: visual sense disambiguation (VSD, Gella et al. 2016).

# Existing Action Recognition Datasets

| Dataset | Actions |
|---|---|
| PPMI (Yao & Fei-Fei 2010) | 24 |
| Stanford 40 (Yao et al. 2011) | 40 |
| PASCAL 2012 (Everingham et al. 2015) | 11 |
| TUHOI (Le et al. 2014) | 2974 |

# Existing Action Recognition Datasets

| Dataset | **Verbs** | Actions | Sense |
|---|---|---|---|
| PPMI (Yao & Fei-Fei 2010) | **2** | 24 | N |
| Stanford 40 (Yao et al. 2011) | **33** | 40 | N |
| PASCAL 2012 (Everingham et al. 2015) | **9** | 11 | N |
| TUHOI (Le et al. 2014) | **–** | 2974 | N |

- Actions: verb phrases or verb-object pairs;
- verb senses are more general than actions;
- no existing datasets with verb sense annotation.

# Dataset for Visual Verb Sense Disambiguation

Design a new dataset using images from:

- MSCOCO: 123k images with object labels, image descriptions:
  - not designed for action recognition;
  - use verbs in descriptions as labels.
- TUHOI: 10,805 images with object labels:
  - labeled with actions (verb-object pairs);
  - use verbs as labels.

# Dataset for Visual Verb Sense Disambiguation

We use the OntoNotes inventory of verb senses (less fine-grained than WordNet).
But: not all verb senses are visual.

Visual:

- S: (v) **play** (perform music on (a musical instrument)) *"He plays the flute"; "Can you play on this old recorder?"*

# Dataset for Visual Verb Sense Disambiguation

We use the OntoNotes inventory of verb senses (less fine-grained than WordNet).
But: not all verb senses are visual.

Visual:

- S: (v) **play** (perform music on (a musical instrument)) *"He plays the flute"; "Can you play on this old recorder?"*

Non-Visual:

- S: (v) **play** (use to one's advantage) *"She plays on her clients' emotions"*
- S: (v) dally, trifle, **play** (consider not very seriously) *"He is trifling with her"; "She plays with the thought of moving to Tasmania"*

# Dataset for Visual Verb Sense Disambiguation

We use the OntoNotes inventory of verb senses (less fine-grained than WordNet).
But: not all verb senses are visual.

Visual:

- S: (v) **play** (perform music on (a musical instrument)) *"He plays the flute"; "Can you play on this old recorder?"*

Non-Visual:

- S: (v) **play** (use to one's advantage) *"She plays on her clients' emotions"*
- S: (v) dally, trifle, **play** (consider not very seriously) *"He is trifling with her"; "She plays with the thought of moving to Tasmania"*

Solution: annotate only the visual senses:

- annotators decide which senses are visual (about 50% in MSCOCO);
- new annotators select correct visual sense for each image.

# Annotating Image and Verb with Visual Sense

- **Verb: play**
  - engage in or make moves related to competition or sport They played cards far into the night. more examples
  - engage in a fun or recreational (childlike) activity The children are playing across the street. more examples
  - perform or transmit music The band played all night long. more examples
  - perform/act a role, pretend He usually plays a villain in films. more examples
  - FISHING-exhaust by allowing to pull on the line John knows how to play a hooked fish.
  - None of the above

# Annotating Image and Verb with Visual Sense

- **Verb: play**
  - engage in or make moves related to competition or sport They played cards far into the night. more examples
  - engage in a fun or recreational (childlike) activity The children are playing across the street. more examples
  - perform or transmit music The band played all night long. more examples
  - perform/act a role, pretend He usually plays a villain in films. more examples
  - FISHING-exhaust by allowing to pull on the line John knows how to play a hooked fish.
  - None of the above

# VerSe Dataset

Comparison of VerSe with existing action recognition datasets:

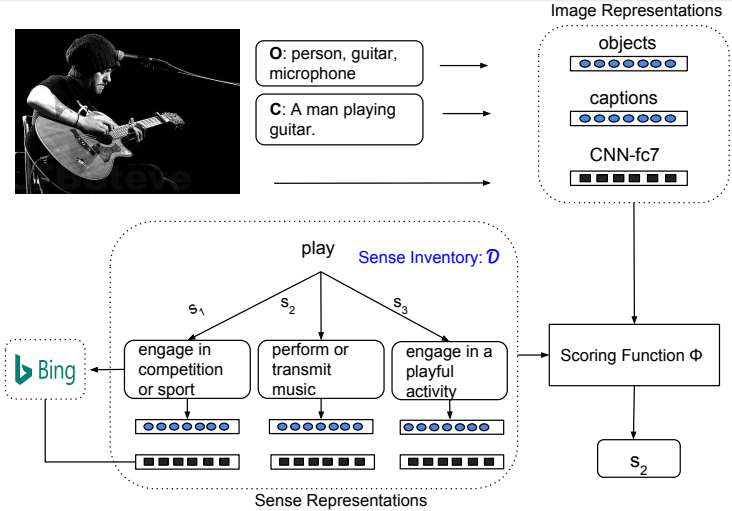
| Dataset | Verbs | Actions | Sense |
|---|---|---|---|
| PPMI (Yao & Fei-Fei 2010) | 2 | 24 | N |
| Stanford 40 (Yao et al. 2011) | 33 | 40 | N |
| PASCAL 2012 (Everingham et al. 2015) | 9 | 11 | N |
| TUHOI (Le et al. 2014) | – | 2974 | N |
| **VerSe** (our dataset) | **90** | – | **Y** (163) |

# VerSe Dataset

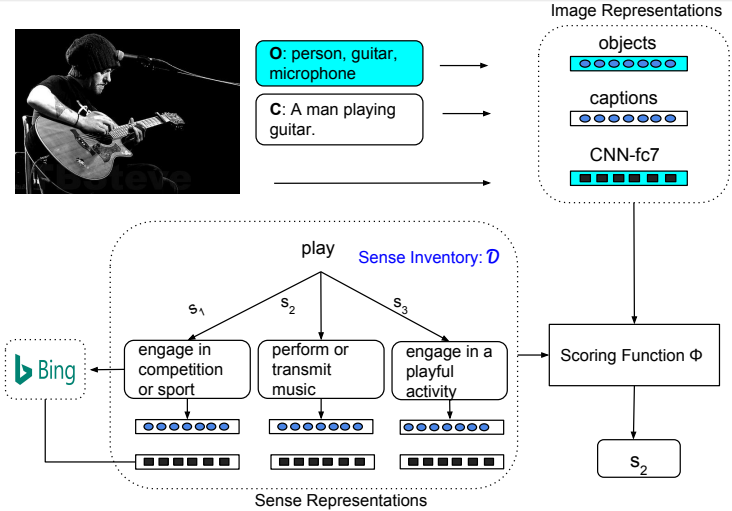
VerSe dataset divided into motion and non-motion verbs:

| Verb type | Verbs | Images | Senses | Examples |
|---|---|---|---|---|
| Motion | 39 | 1812 | 5.79 | run, walk, jump, swing, hit, kick |
| Non-motion | 51 | 1698 | 4.86 | sleep, sit, lean, read, write, look |

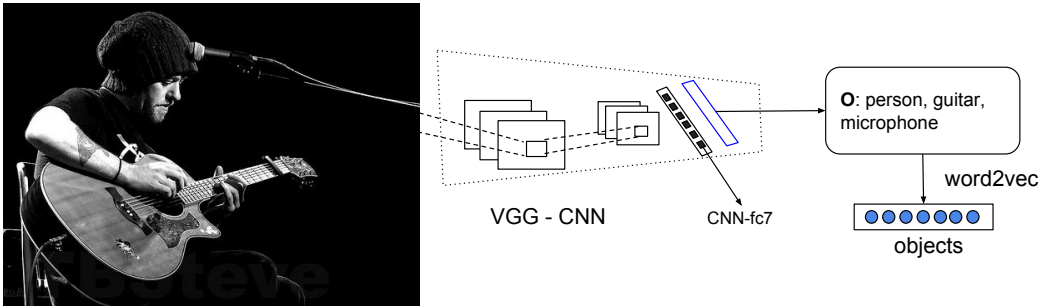
# Unsupervised Model for VSD



# Unsupervised Model for VSD

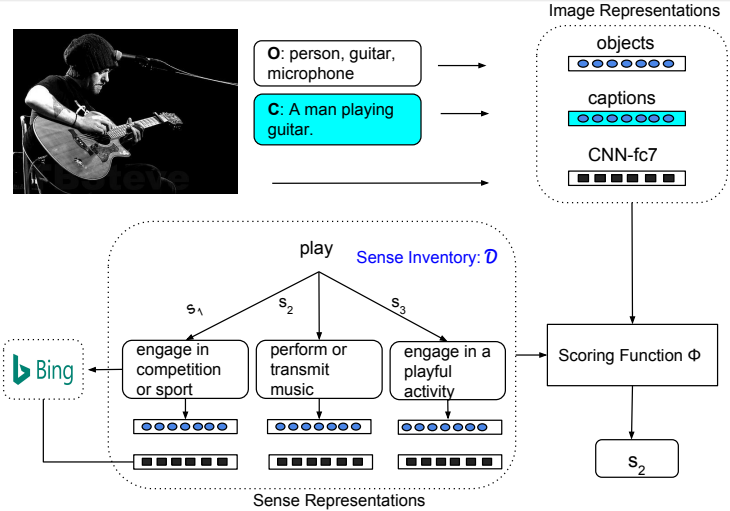


# Unsupervised Model for VSD




Object labels obtained using VGG (Simonyan & Zisserman 2014).

# Unsupervised Model for VSD



# Unsupervised Model for VSD

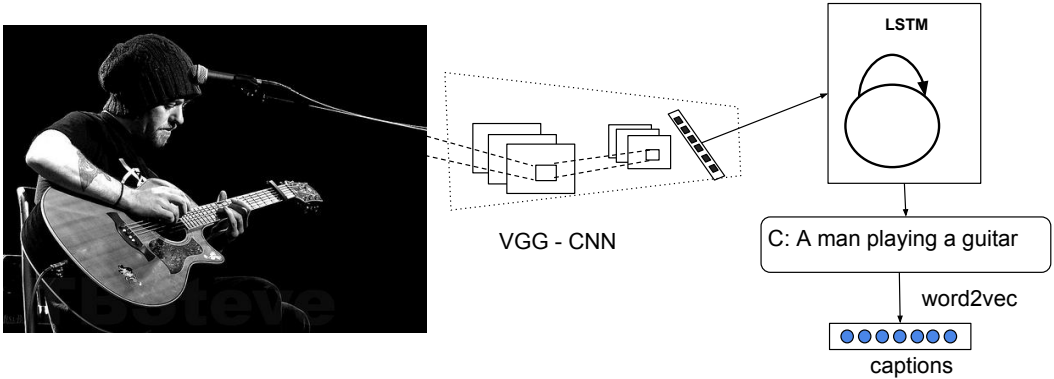
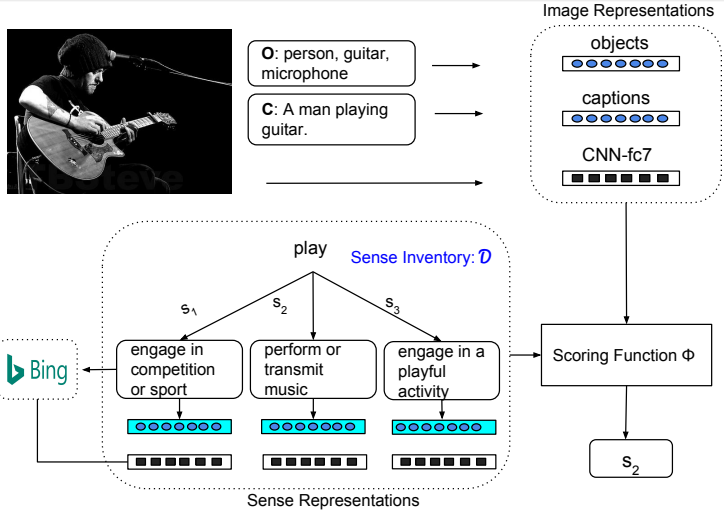
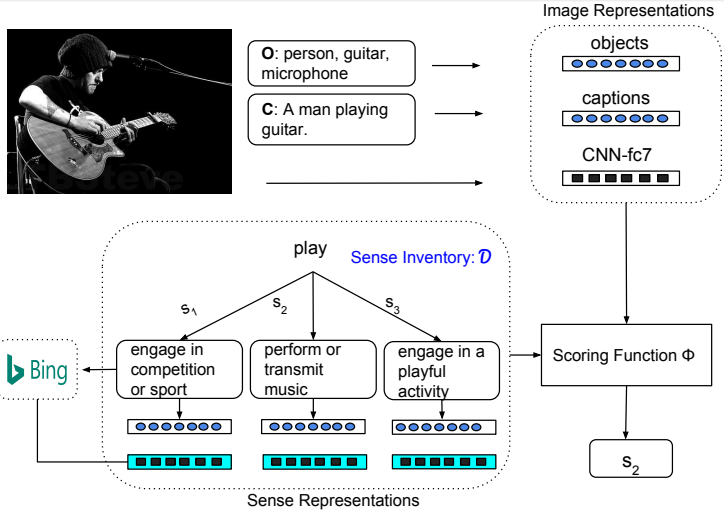



Image descriptions from Show and Tell (Vinyals et al. 2015).

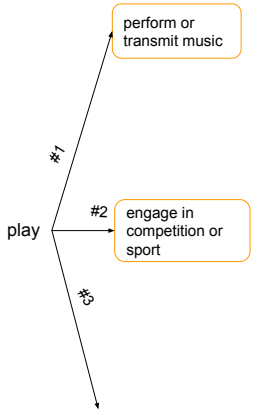
# Unsupervised Model for VSD



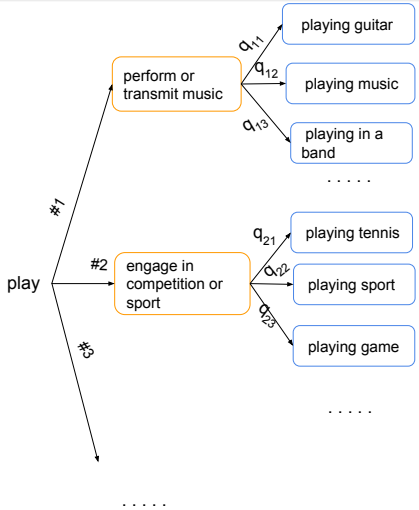
# Unsupervised Model for VSD



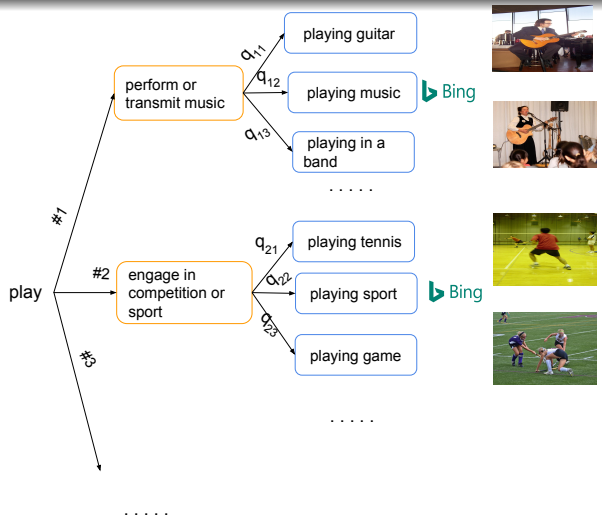
# Visual Representation for Senses



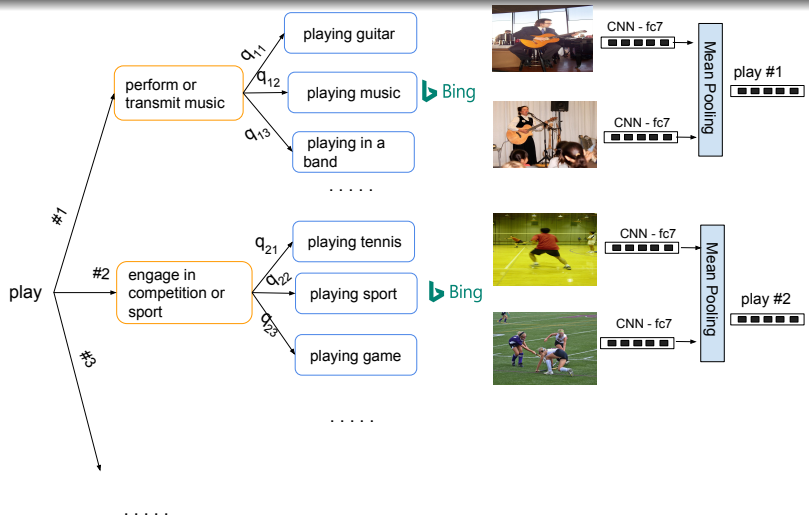
# Visual Representation for Senses



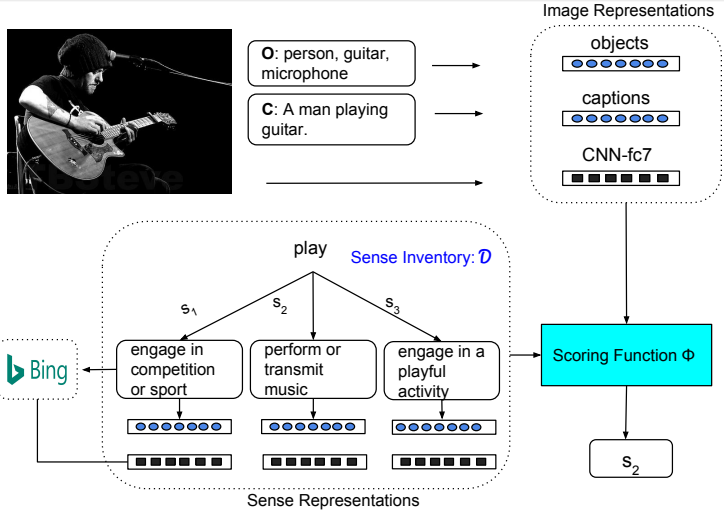
# Visual Representation for Senses



# Visual Representation for Senses



# Unsupervised Model for VSD



# Scoring Function

Use vector similarity (cosine) as scoring function:

$$\hat{s} = \arg\max_{s \in S(v)} \Phi(s, i, v, D)$$

# Scoring Function

Use vector similarity (cosine) as scoring function:

$$\hat{s} = \arg\max_{s \in S(v)} \Phi(s, i, v, D)$$

Representations:

- textual: O, C embeddings;

# Scoring Function

Use vector similarity (cosine) as scoring function:

$$\hat{s} = \arg\max_{s \in S(v)} \Phi(s, i, v, D)$$

Representations:

- textual: O, C embeddings;
- visual: CNN features;

# Scoring Function

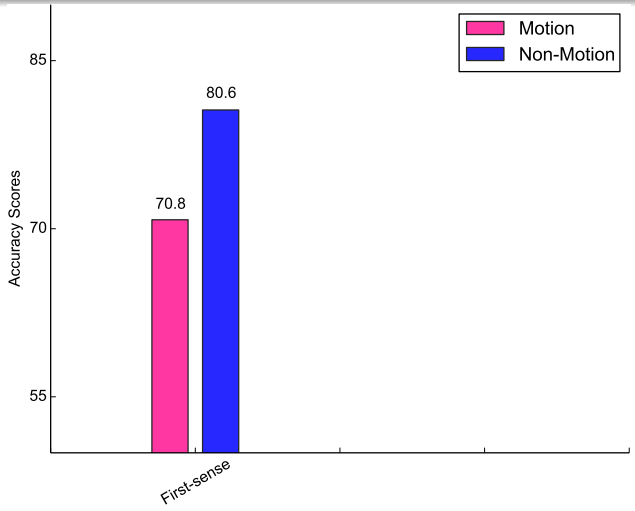
Use vector similarity (cosine) as scoring function:

$$\hat{s} = \arg\max_{s \in S(v)} \Phi(s, i, v, D)$$

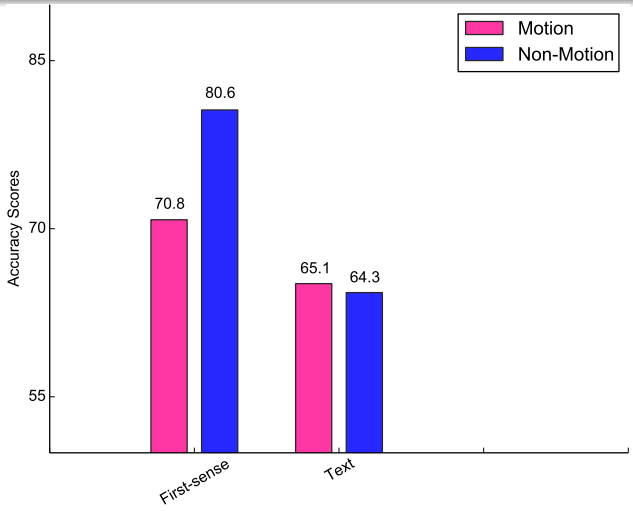
Representations:

- textual: O, C embeddings;
- visual: CNN features;
- multi-modal: fused textual and visual features using Canonical Correlation Analysis.

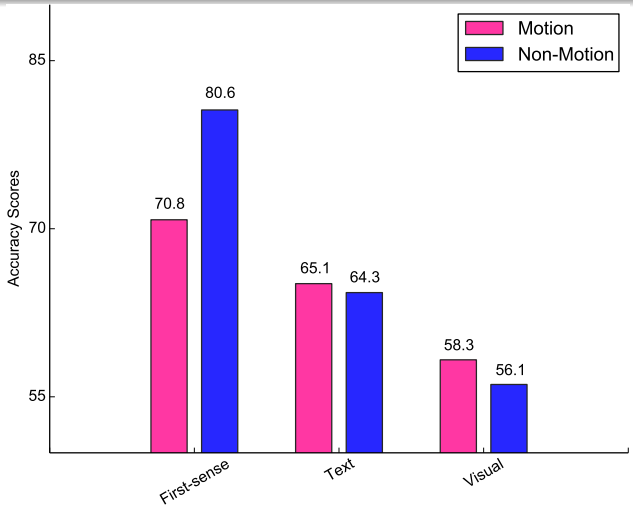
# Results


Motion
Non-Motion
85
70
55
Accuracy Scores
80.6
70.8
First-sense

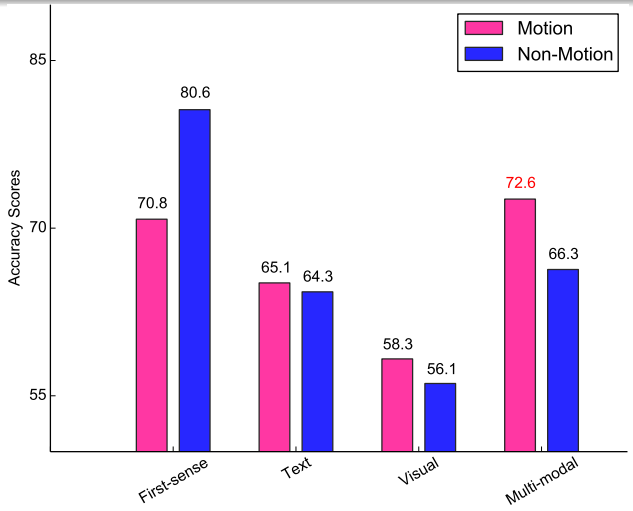
# Results


Motion
Non-Motion
Accuracy Scores
85
70
55
80.6
70.8
65.1
64.3
First-sense
Text

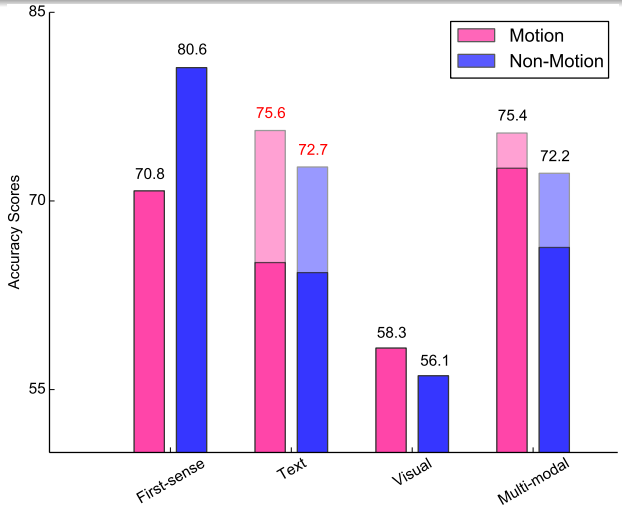
# Results


Motion
Non-Motion
Accuracy Scores
85
70
55
80.6
70.8
65.1
64.3
58.3
56.1
First-sense
Text
Visual

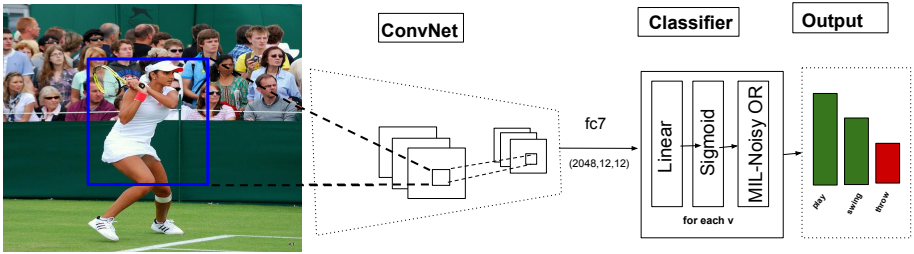
# Results


Motion
Non-Motion
Accuracy Scores
85
70
55
80.6
70.8
65.1
64.3
58.3
56.1
72.6
66.3
First-sense
Text
Visual
Multi-modal

# Results: Gold Standard Image Descriptions



# Verb Prediction




- detect the verbs that are present in an image (250 classes);
- use multiple instance learning (we do not know which bounding boxes correspond to which verbs).

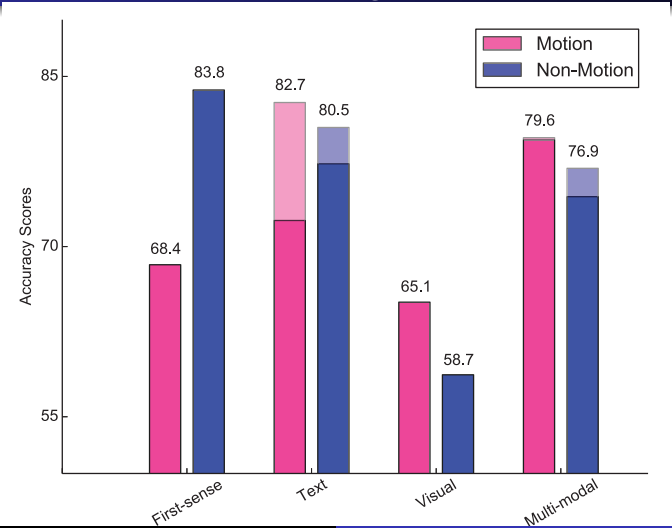
# Examples: Verb Prediction

play, perform

hit, swing, play

hold, sit, use

# Verb Prediction and Sense Disambiguation



# Conclusions

- Image understanding (like text understanding) requires structured representations;
- for multimodal tasks, we need to align linguistic and image structure;
- syntactic example: visual dependency representations align geometric structure of an image with syntactic structure of a sentence;
- application in image description and image retrieval;
- semantic example: visual word senses align event depicted in an image with event described in a sentence;
- unsupervised VSD model using multimodal embeddings.

# Other Approaches to Image Structure

Other approaches that align linguistic structure and image structure:

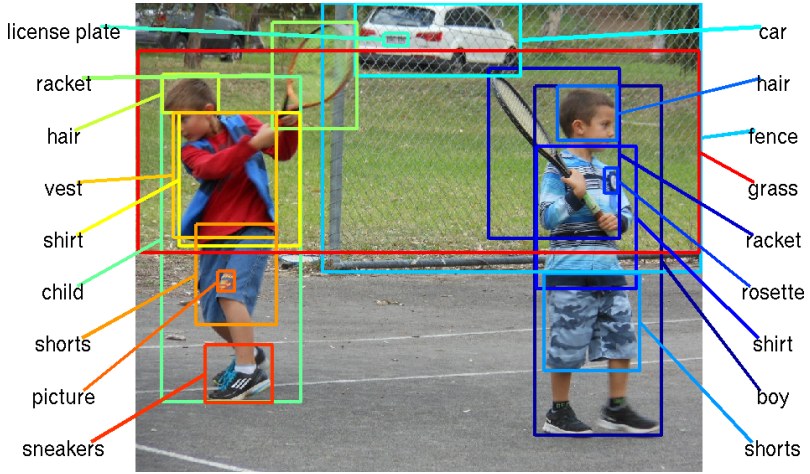
Scene (description) graphs (Johnson et al. 2015; Aditya et al. 2015):

- triples of object, attribute, relation;
- aligned with image regions and region descriptions;
- no explicit alignment with linguistic structure (but could be derived).

Visual semantic roles (Yatskar et al. 2016):

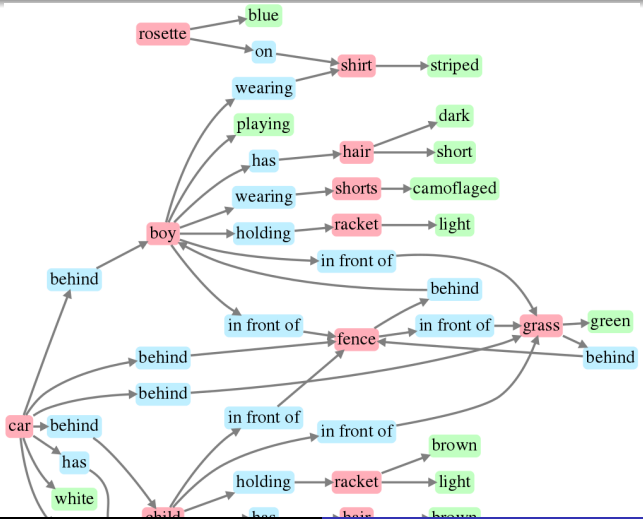
- uses semantic frames from FrameNet;
- annotates images with frames, participants, and roles;
- not aligned with regions or image descriptions; no verb senses.

# Scene Graphs




http://cs.stanford.edu/people/jcjohns/cvpr15_supp/

# Scene Graphs



# Visual Semantic Roles

| practicing | | | |
|---|---|---|---|
| agent | skill | tool | place |
| man | music | violin | room |

http://imsitu.org/demo/

# References I

Aditya, S., Yang, Y., Baral, C., Fermuller, C., & Aloimonos, Y. (2015). From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292*.

Elliott, D., & Keller, F. (2013). Image description using visual dependency representations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, (pp. 1292–1302), Seattle, WA.

Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J. M., & Zisserman, A. (2015). The Pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, *111*, 98–136.

Gella, S., Lapata, M., & Keller, F. (2016). Unsupervised visual sense disambiguation for verbs using multimodal embedding. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, (pp. 182–192), San Diego, CA.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D. A., Bernstein, M., & Fei-Fei, L. (2015). Image retrieval using scene graphs. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, (pp. 3668–3678), Boston, MA.

Le, D.-T., Uijlings, J., & Bernardi, R. (2014). *Proceedings of the Third Workshop on Vision and Language*, chap. TUHOI: Trento Universal Human Object Interaction Dataset, (pp. 17–24). Dublin City University and the Association for Computational Linguistics.

# References II

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, *abs/1409.1556*.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, (pp. 3156–3164).

Yao, B., & Fei-Fei, L. (2010). Grouplet: A structured image representation for recognizing human and object interactions. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, (pp. 9–16). IEEE.

Yao, B., Jiang, X., Khosla, A., Lin, A. L., Guibas, L., & Fei-Fei, L. (2011). Human action recognition by learning bases of action attributes and parts. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, (pp. 1331–1338). IEEE.

Yatskar, M., Zettlemoyer, L., & Farhadi, A. (2016). Situation recognition: Visual semantic role labeling for image understanding. In *Computer Vision and Pattern Recognition*.