Learning language through pictures

Grzegorz Chrupała, Ákos Kádár and Afra Alishahi

Tilburg University

Word and phrase meanings

- Perceptual clues

- Distributional clues

the **cat** sat on the mat

the dog chased the **cat**

funniest **cat** video ever lol

Real scenes

- Harder
  - objects need to be identified
    - invariances detected
- But also easier
  - better opportunities for generalization

Cross-situational learning

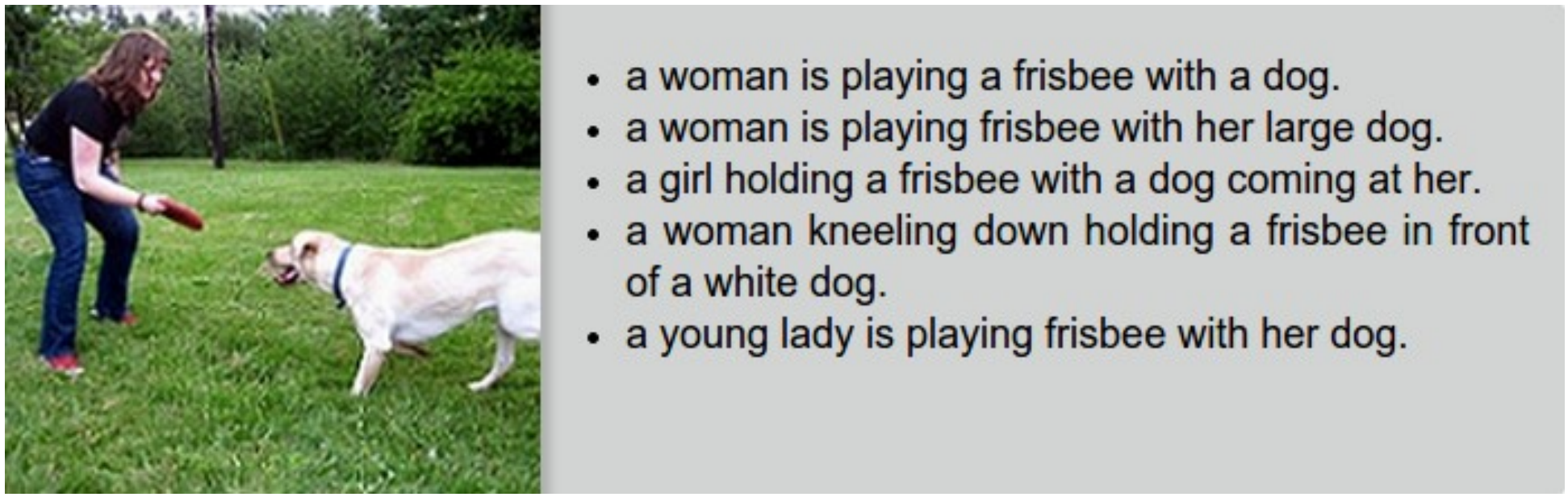
- Synthetic data (Fazly et al. 2010)
  - Utterance: *a bird walks on a beam*
  - Scene: `{bird, big, legs, walk, wooden, beam}`

- "Coded" scene representations
  (Frank et al. 2009)

Cross-situational learning

- Synthetic data (Fazly et al. 2010)
  - Utterance: *a bird walks on a beam*
  - Scene: `{bird, big, legs, walk, wooden, beam}`
- "Coded" scene representations
  (Frank et al. 2009)
- Natural scenes not set of symbols

Captioned images

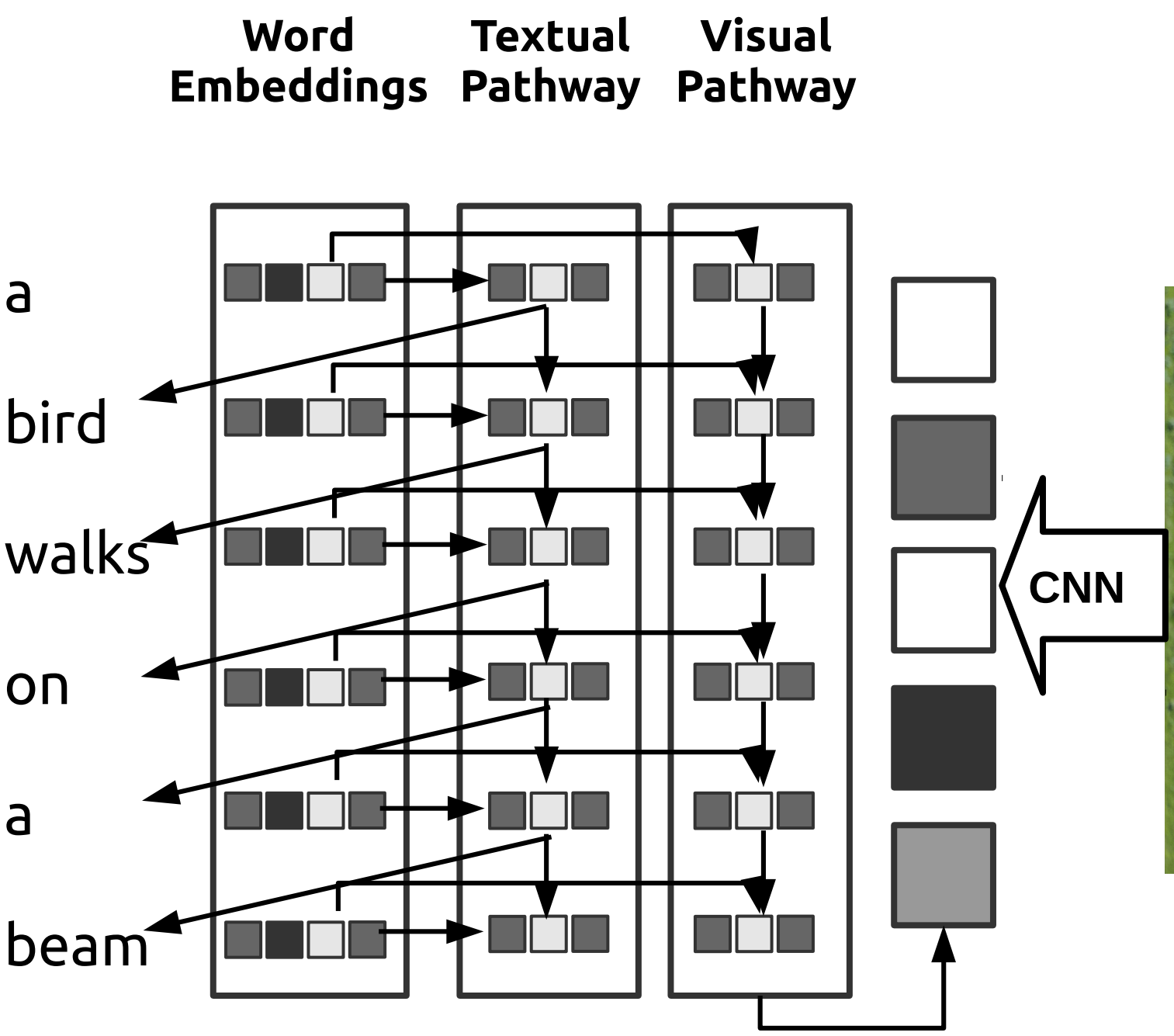


Recent works on generating image descriptions use actual image features.

IMAGINET
Multi-task language/image model

- Integrate linguistic and visual context

- Representations of phrases and complete sentences


Word Embeddings
Textual Pathway
Visual Pathway
a
bird
walks
on
a
beam
CNN

Some details

- Shared word embeddings – 1024 units
- Pathways – Gated Recurrent Unit nets
  - 1024 clipped rectifier units
- Image representations: 4096 dimensions
- Multi-task objective

Multi-task objective

$$L(\theta) = \alpha L^T(\theta) + (1 - \alpha) L^V(\theta)$$

- L^T – cross-entropy loss
- L^V – mean squared error

- Three versions
  - $a = 0$ – purely visual model
  - $a = 1$ – purely textual model
  - $0 < a < 1$ – multi-task model

Bag-of-words linear regression as a baseline

- Baseline
  - Input: word-count vector
  - Output: image vector
  - L2-penalized sum-of-squared errors regression

Correlations with human judgments

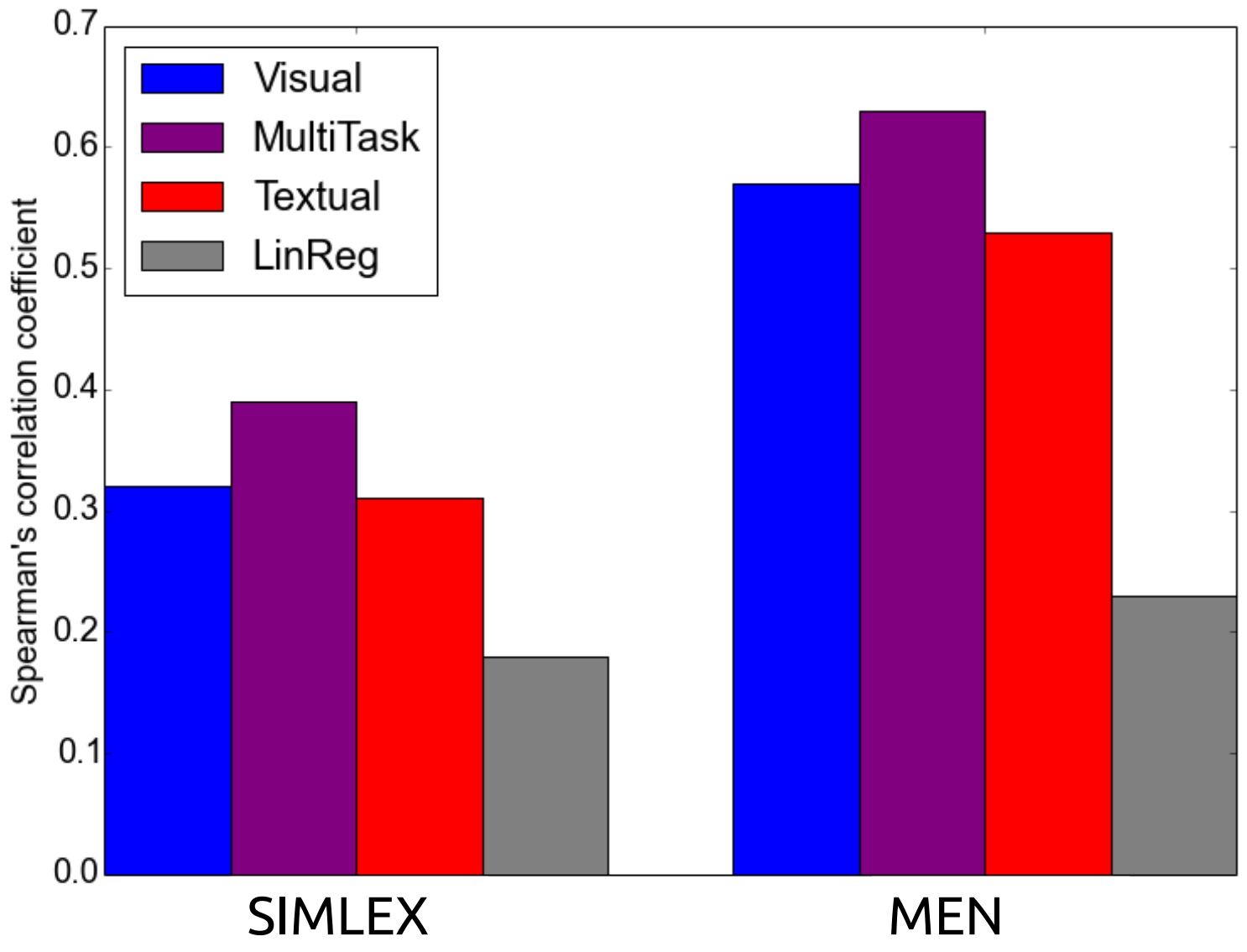


Image retrieval task

- Embed caption in visual space
- Rank images according to cosine similarity to caption

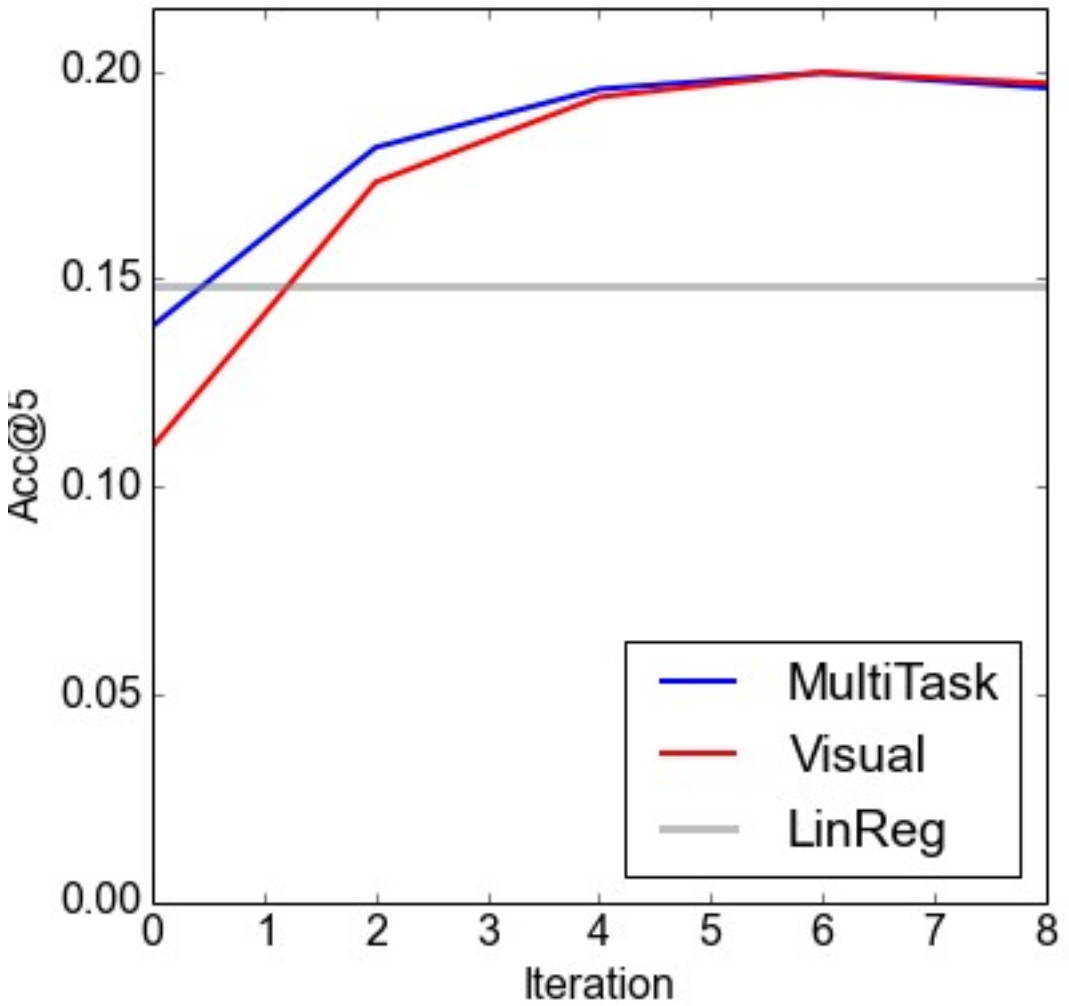
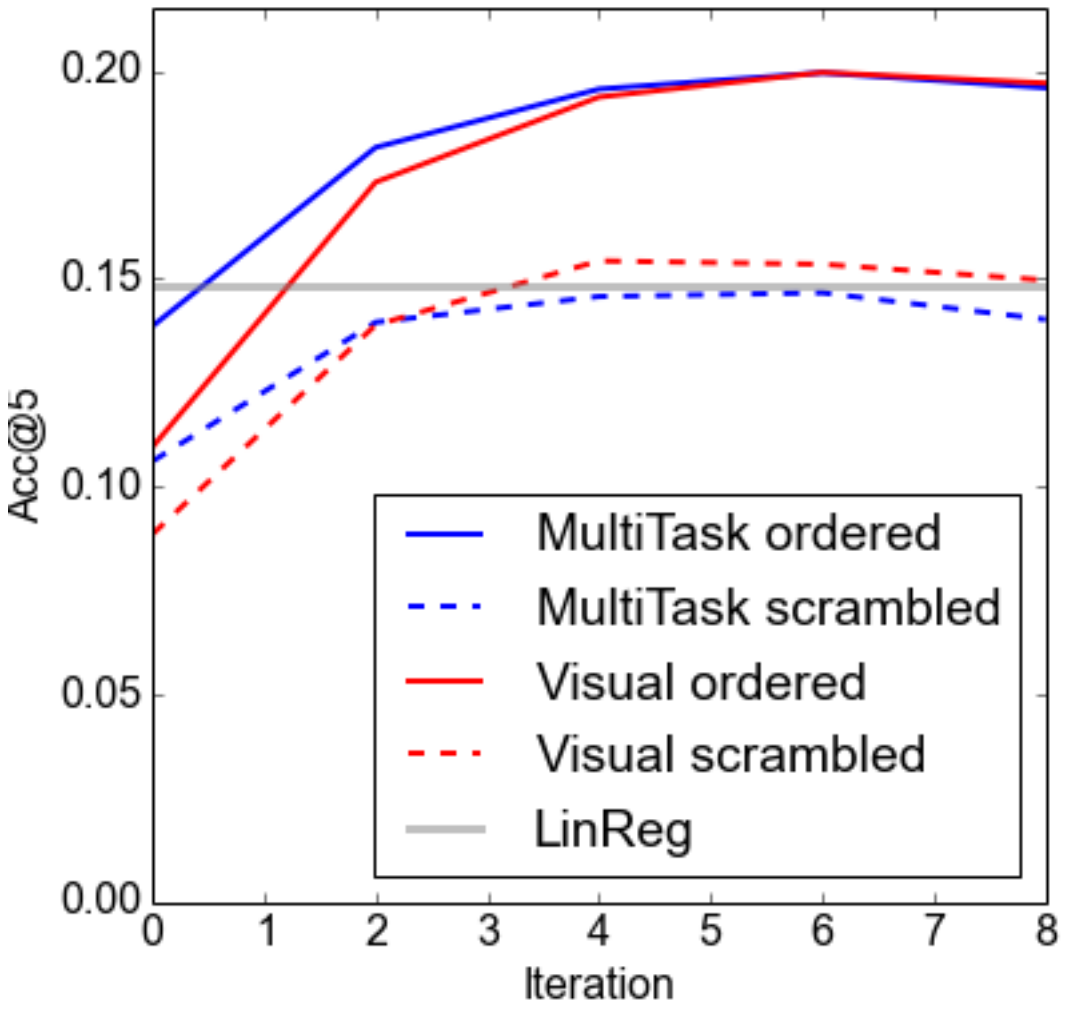
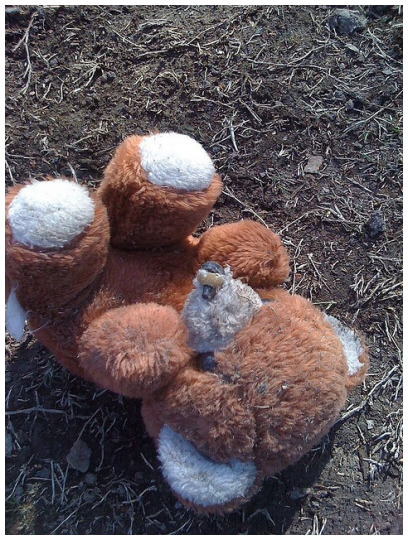


Image retrieval and sentence structure

- **Original** versus **scrambled** captions



a brown teddy bear lying on top of a dry grass covered ground



a a of covered laying bear on brown grass top teddy ground . dry

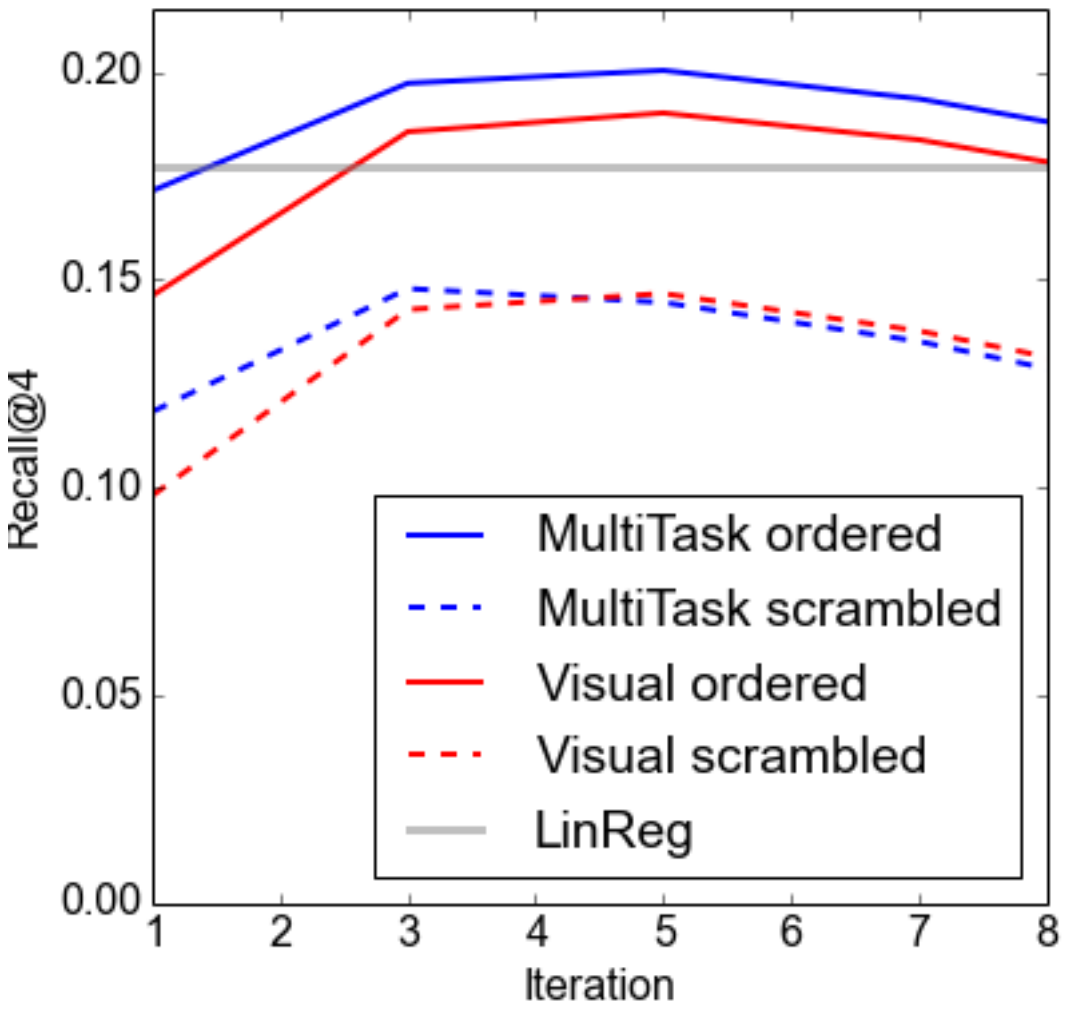
a variety of kitchen utensils hanging from a UNK board .

kitchen of from hanging UNK variety a board utensils a .

Paraphrase retrieval

- Record the final state along the visual pathway for a caption
- For each caption, rank others according to cosine similarity
- Are top-ranked captions about the same image?

Paraphrase retrieval



a cute baby playing with a cell phone

- small baby smiling at camera and talking on phone .
- a smiling baby holding a cell phone up to ear .
- a little baby with blue eyes talking on a phone .

phone playing cute cell a with baby a

- someone is using their phone to send a text or play a game .
- a camera is placed next to a cellular phone .
- a person that 's holding a mobile phone device

Imaginet:

- Learns visually-grounded word and sentence representations from multimodal data

- Encodes and uses aspects of linguistic structure

Current & future work

- Understand internal states
  - Poster at EMNLP **VL2015**
- Character level modeling

Thanks!

Compared to compositional distributional semantics

word embeddings	**distributional word vectors**
hidden states	**sentence vectors**
input-to-hidden weights	**projection to sentence space**
hidden-to-hidden weights	**composition operator**

All these are learned based on supervision signal from the two tasks

Compared to captioning

- Captioning (e.g. Vinyals et al. 2014)
  - Start with image vector
  - Output caption word-by-word
    - conditioning on image and seen words
- IMAGINET
  - Read caption word-by-word
  - Incrementally build sentence representation
    - while also predicting the coming word
  - Finally, map to image vector

Long term

- Character-level input
  - proof of concept working
- Direct audio input
- Need better story on
  - what should be learned from data
  - what should be hard-coded, or evolved

Gated recurrent units

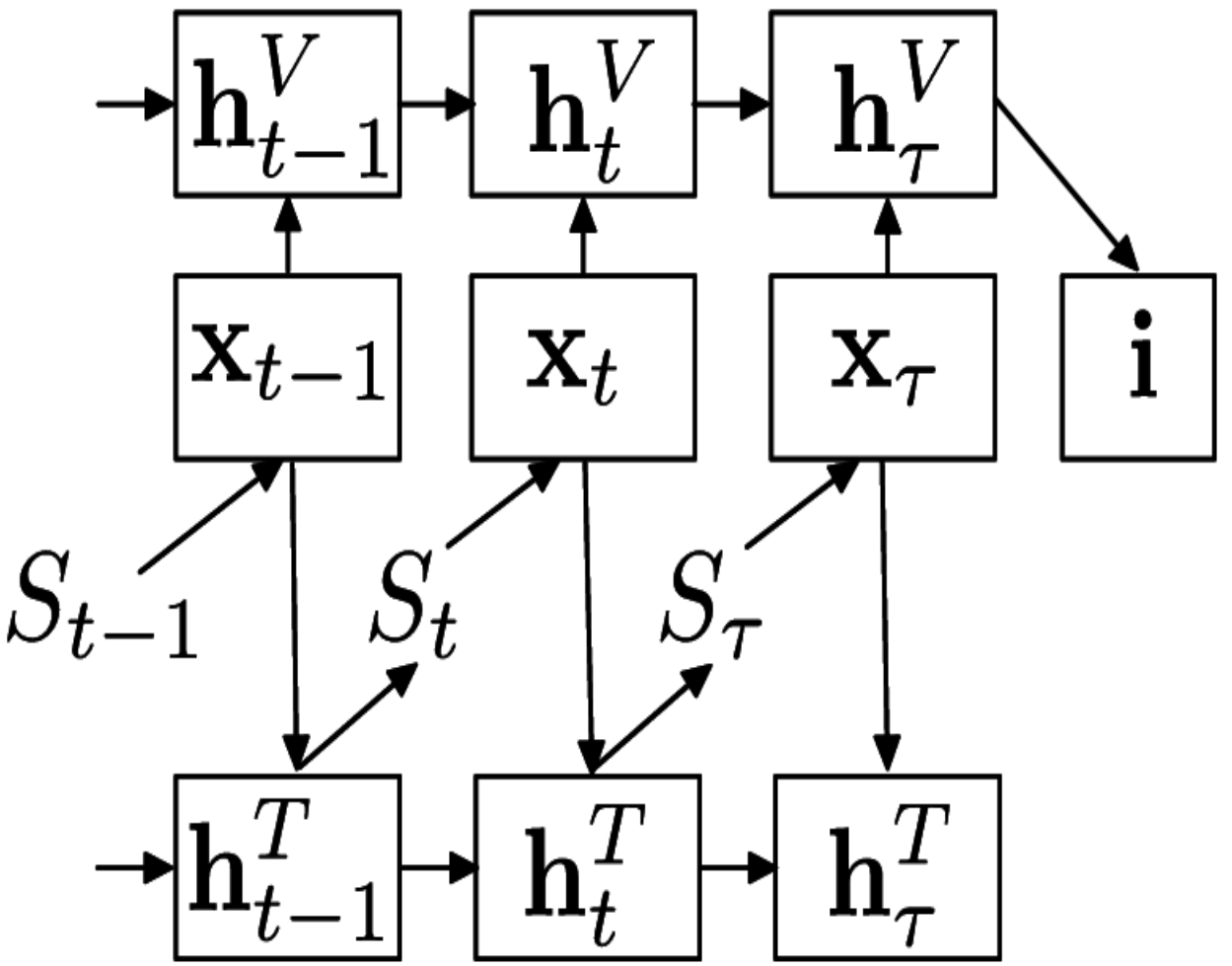
$$h_t^j = (1 - z_t^j) h_{t-1}^j + z_t^j \tilde{h}_t^j$$

$$z_t^j = \sigma_s(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1})^j$$

$$\tilde{h}_t^j = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1}))^j$$

$$r_t^j = \sigma_s(\mathbf{W}_r \mathbf{x}_r + \mathbf{U}_r \mathbf{h}_{t-1})^j$$

IMAGINET


$\mathbf{h}_{t-1}^{V}$
$\mathbf{h}_{t}^{V}$
$\mathbf{h}_{\tau}^{V}$
$\mathbf{x}_{t-1}$
$\mathbf{x}_{t}$
$\mathbf{x}_{\tau}$
$\mathbf{i}$
S_{t-1}
S_{t}
S_{τ}
$\mathbf{h}_{t-1}^{T}$
$\mathbf{h}_{t}^{T}$
$\mathbf{h}_{\tau}^{T}$