A picture is worth five captions

Learning visually grounded word and sentence representations

Grzegorz Chrupała

(with Ákos Kádár and Afra Alishahi)

Learning word (and phrase) meanings

- Cross-situational

- Distributional

```
the cat sat on
the mat

the dog chased
the cat

funniest cat
video ever lol
```

Distributional

- Very popular in Cogsci and NLP
  - LSA, LDA, word2vec, …
- Massive amounts of data
- Recent focus on compositionality

Cross-situational

- Synthetic data (Fazly et al. 2010)

Utterance: *Joe is happily eating an apple*
Scene: {`joe`, `quickly`, `eat`, `a`, `big`, `red`, `apple`, `hand`}

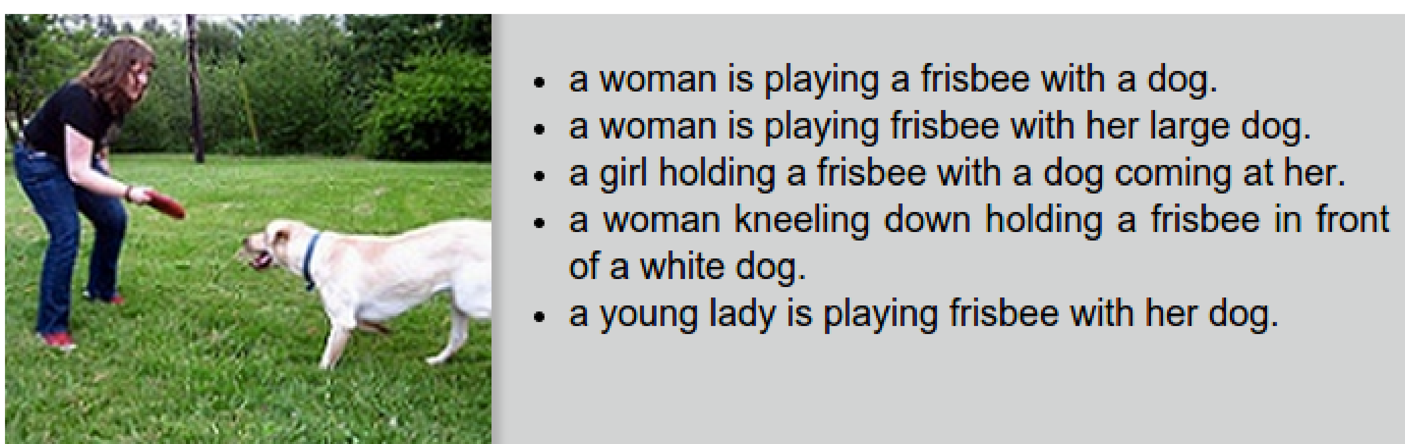
- "Coded" scene representations (Frank et al. 2009)
- But natural scenes are not sets of symbols

Real scenes

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

Captioned images

Young et al. 2014
Denotational semantics –
only use images as opaque ids



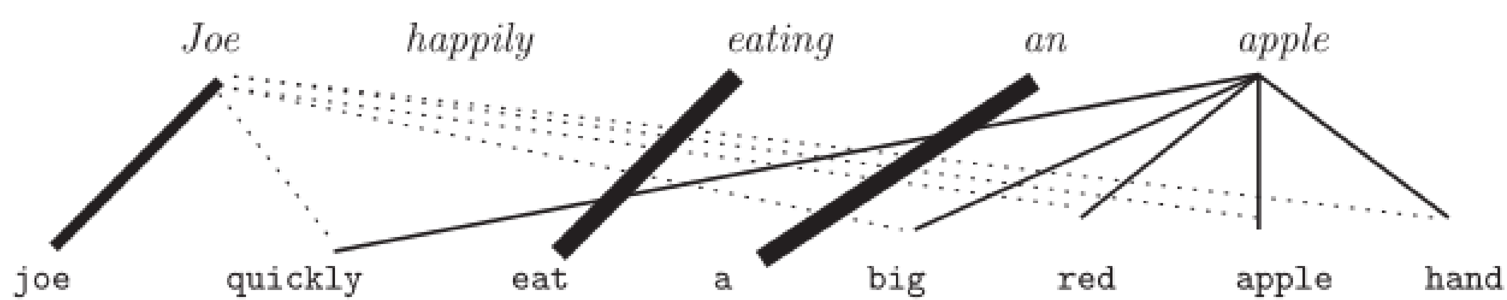
Several works on generating captions – use actual **image features**

Visually grounded word and sentence representations

- Learn from
  - linguistic context
  - (non-symbolic) visual context
- Compositionality
  - Word, phrase and sentence representations

Aligning words and image features

Based on word learning model for synthetic data

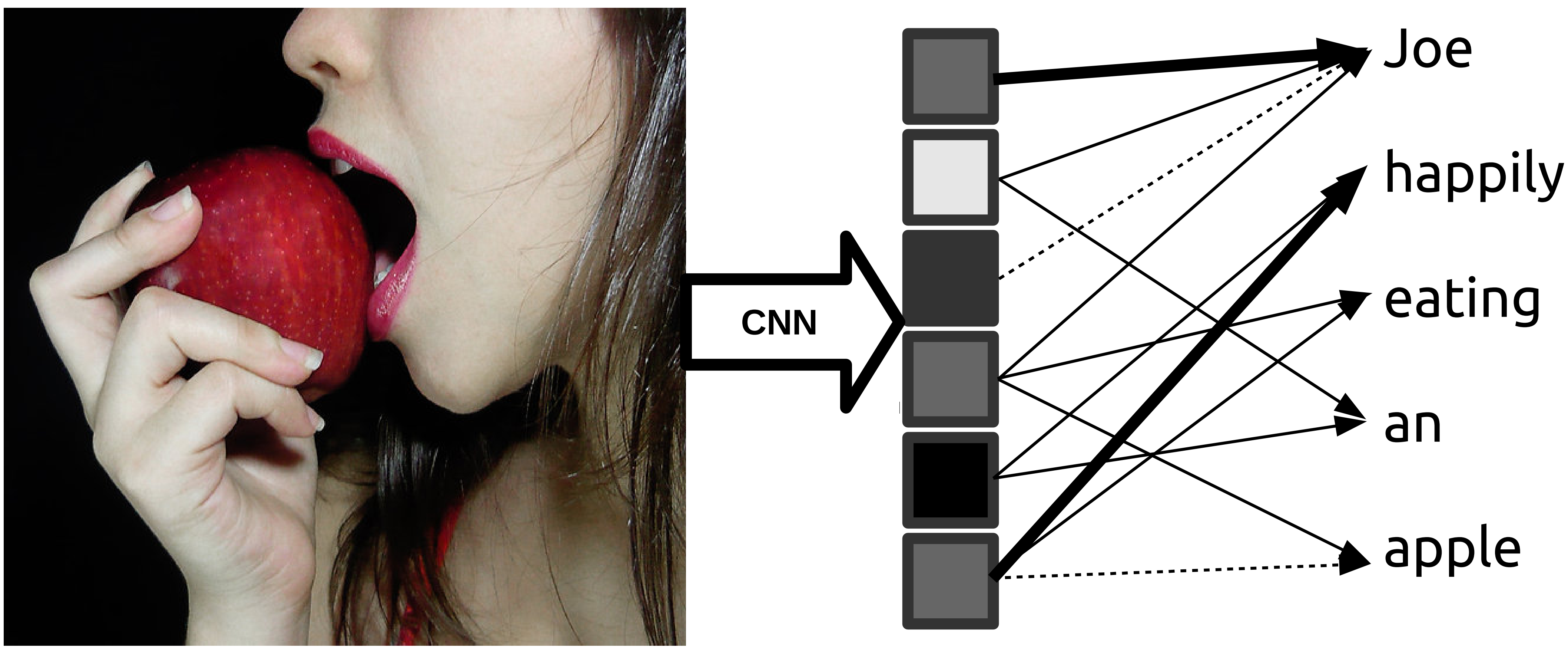


A Probabilistic Computational Model of Cross-Situational Word Learning

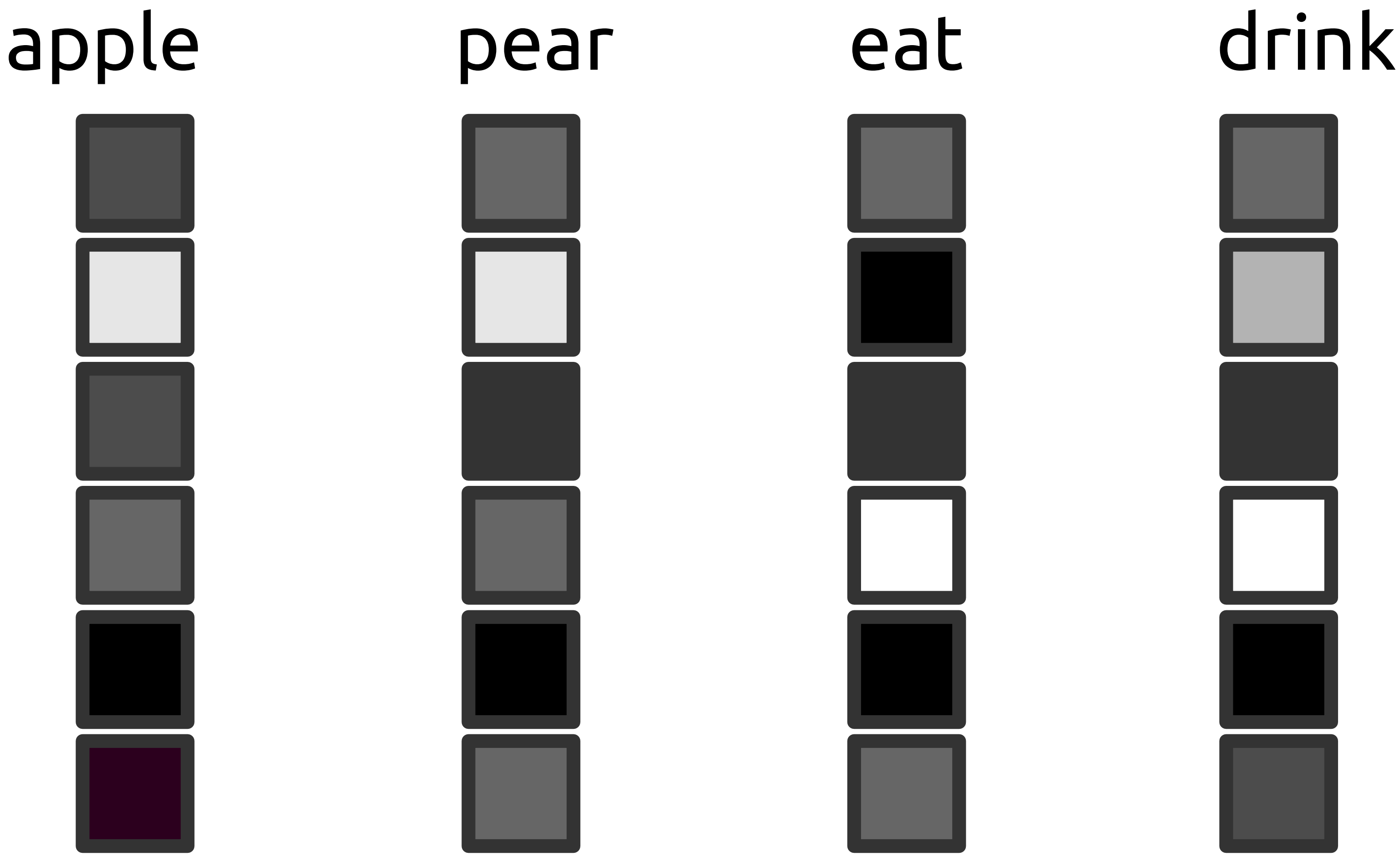
Afsaneh Fazly,[a] Afra Alishahi,[b] Suzanne Stevenson[a]

Cognitive Science, 2010

Feature-word alignment



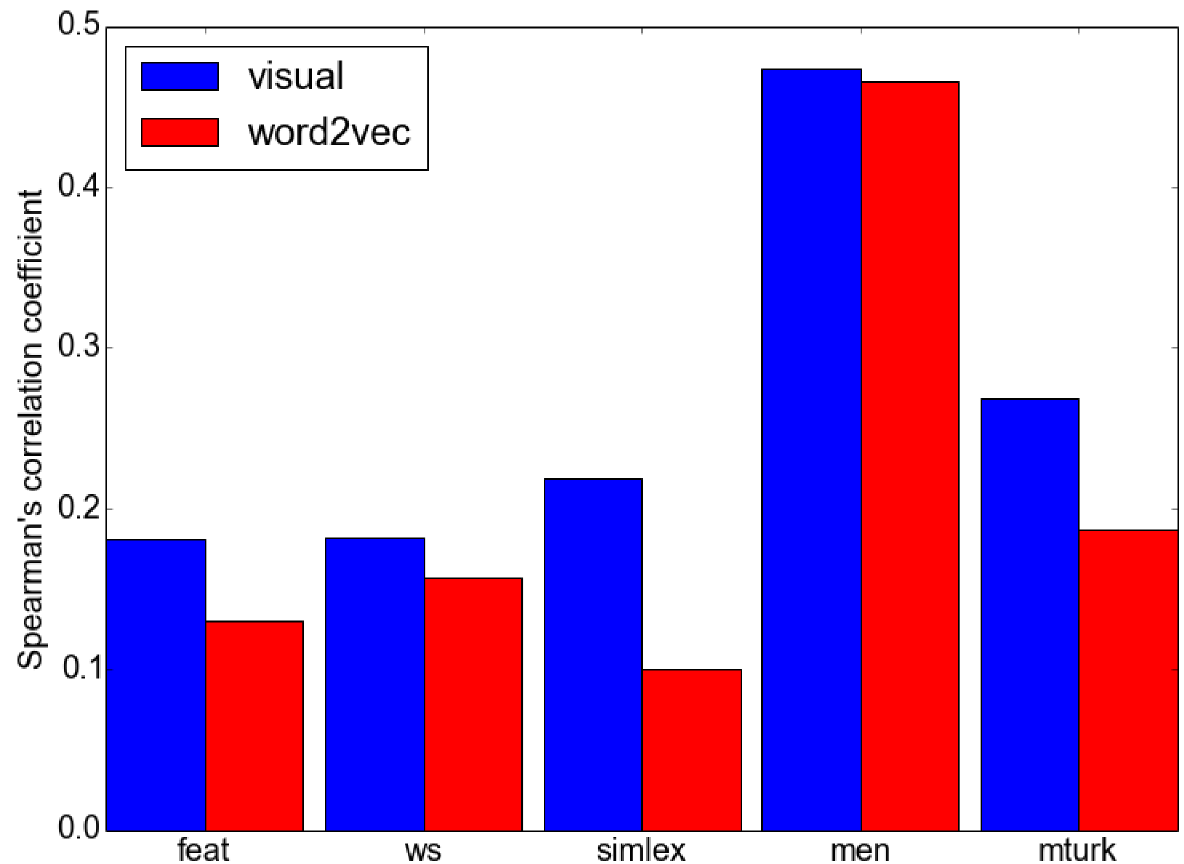
Visual word vectors (4096 dimensions)



Evaluation

- Unlike for synthetic data – no ground truth
- Indirect evaluation
  - correlation with human similarity judgments
    - what exactly do we get when we ask for these judgments?
  - search images based on captions
  - generate captions for images
  - paraphrase captions
  - ….

Correlations with human judgments (Flickr30K)



Predicting ImageNet labels from word representations

Label: aircraft carrier, carrier, flattop

Hypernym: vehicle

Predicted: distant, boats, ship, houses

Label: stove

Hypernym: device

Predicted: fire, candles, taken, lit

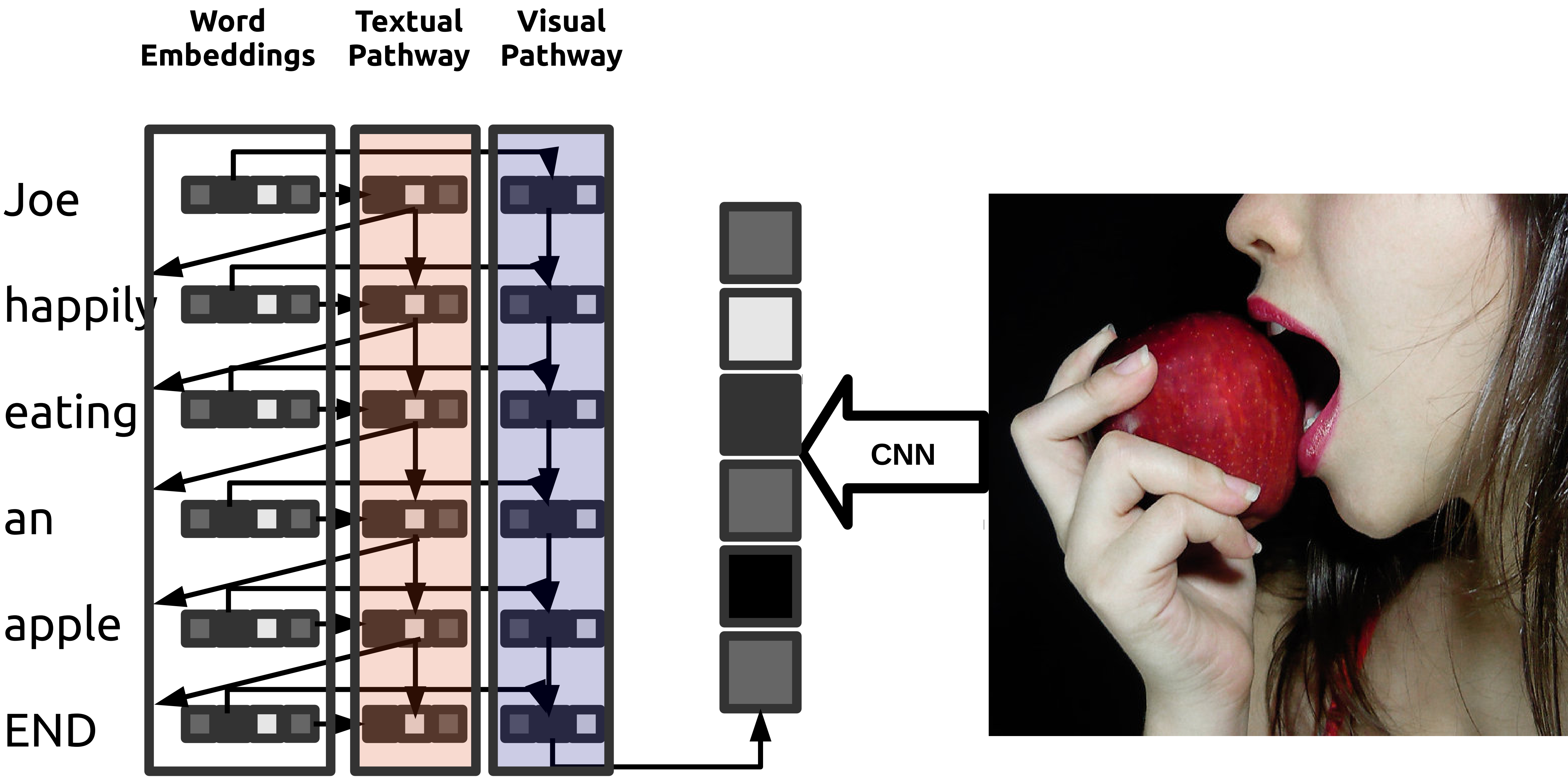
But need more

- Integrate linguistic and visual context
- Representations of phrases and complete sentences

- Start from scratch

IMAGINET
Multi-task language/image model

- Neural network model
  - Generality
  - Separation of modeling and learning algorithm
  - Reusable building blocks
  - Successful in a variety of tasks including captioning
- But, opaque internal states
  - Need techniques to help interpretability


Word
Embeddings
Textual
Pathway
Visual
Pathway
Joe
happily
eating
an
apple
END
CNN

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

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

Some details

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

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

Multi-task objective

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

- L^T – cross-entropy loss
  (mean negative log probability of next word)
- L^V – mean squared error
- $a = 0$ – purely visual model
- $a = 1$ – purely textual model
- $0 < a < 1$ – multi-task model

Bag-of-words linear regression as a baseline

- How much do embeddings and recurrent nets contribute?
- Baseline
  - Input: word-count vector
  - Output: image vector
  - L2-penalized sum-of-squared errors regression

Dataset

Microsoft COCO
Common Objects in Context

What is Microsoft COCO?

Microsoft COCO is a new image recognition, segmentation, and captioning dataset. Microsoft COCO has several features:

- **Object segmentation**
- **Recognition in Context**
- **Multiple objects per image**
- **More than 300,000 images**
- **More than 2 Million instances**
- **80 object categories**
- **5 captions per image**

Correlations with human jugdments

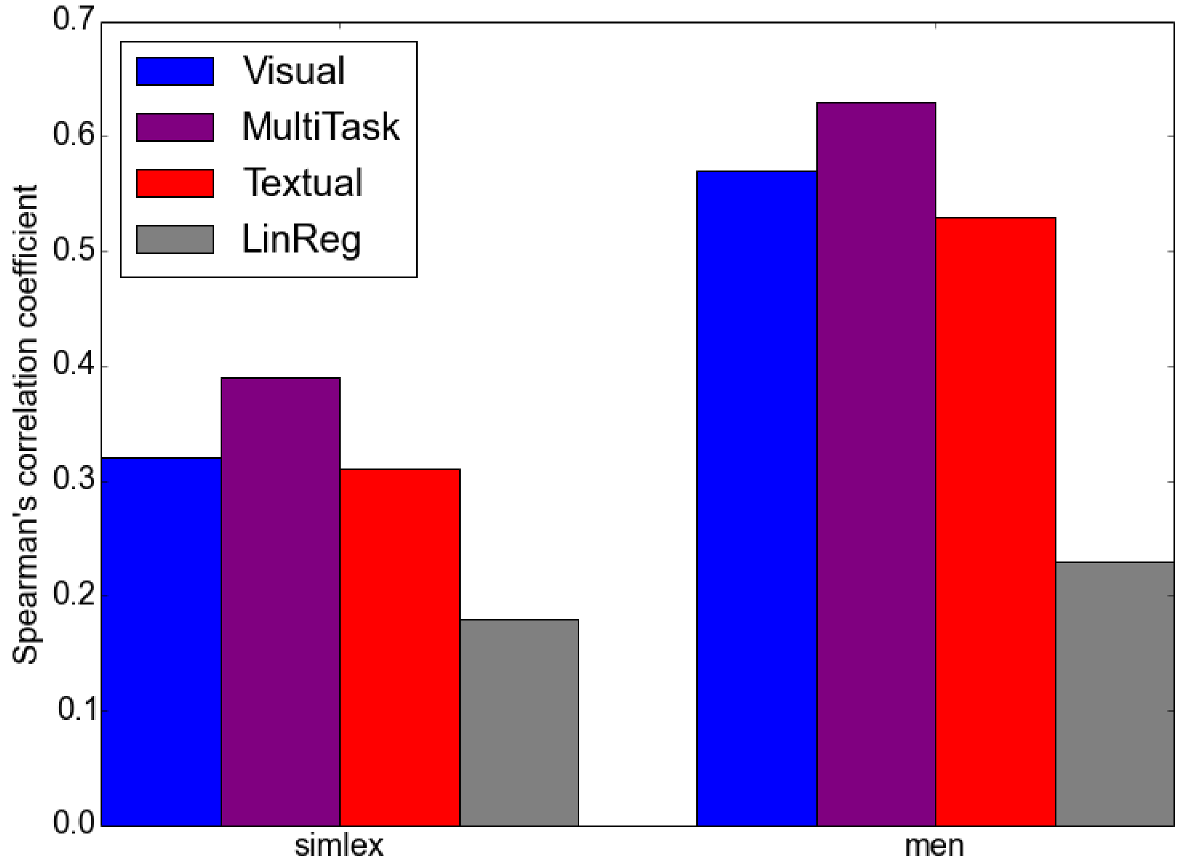
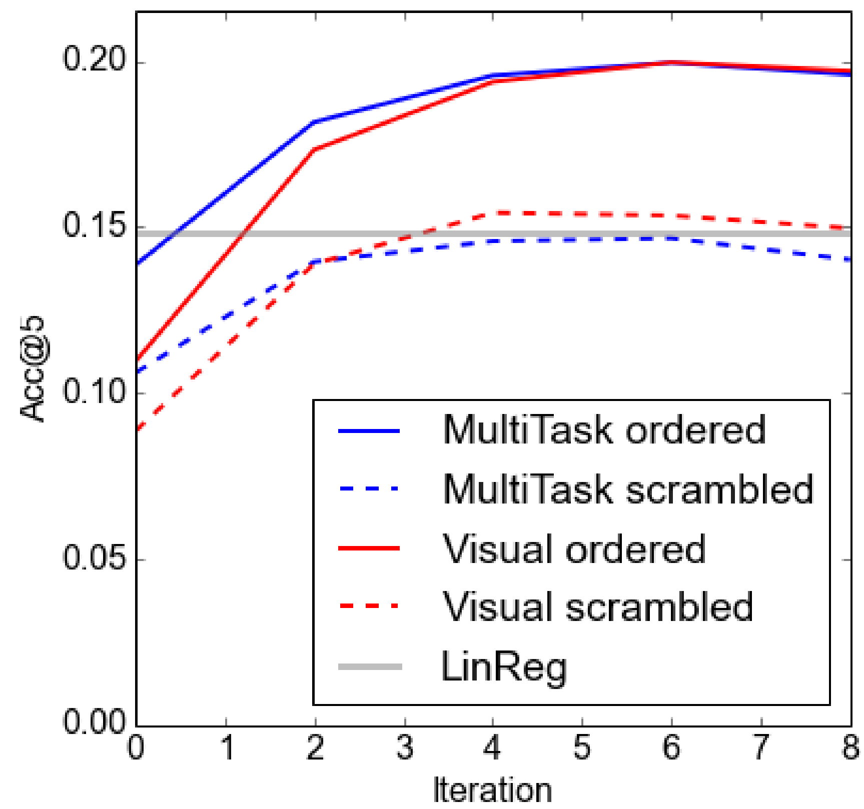


Image retrieval and sentence structure

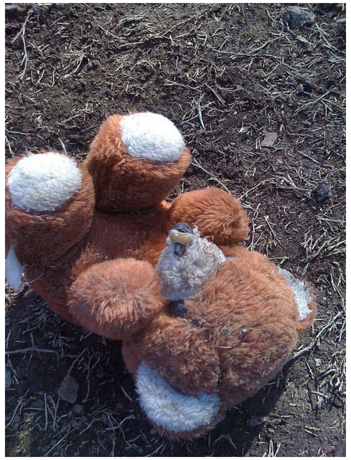
- Project **original** and **scrambled** caption to visual space
- Rank images according to cosine similarity to caption



a pigeon with red feet perched on a wall .

feet on wall . pigeon a red with a perched

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

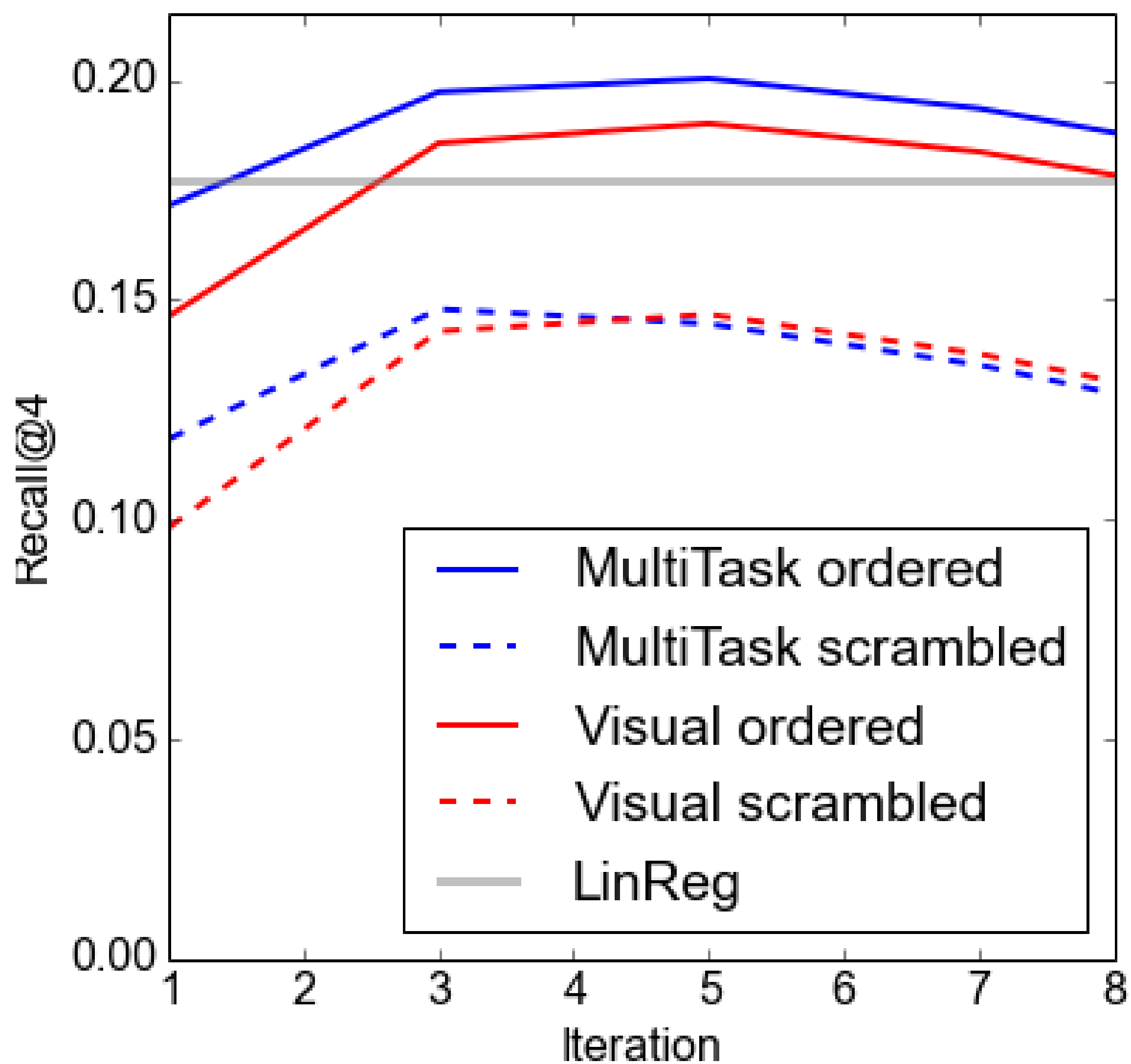
a variety of kitchen utensils hanging from a UNK board .

kitchen of from hanging UNK variety a board utensils a .

Paraphrase retrieval

- Record the final state along the visual pathway for a (maybe scrambled) 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

a couple of horses UNK their head over a rock pile

- two brown horses hold their heads above a rocky wall .
- two horses looking over a short stone wall .

rock couple their head pile a a UNK over of horses

- an image of a man on a couple of horses
- looking in to a straw lined pen of cows

Currently working on

- Encourage complete sentence representations along textual pathway
  - longer-range predictions
  - caption reconstruction
- Disentangle relative contribution of
  - word embeddings
  - recurrent state
- Controlled manipulation of inputs

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

Thanks!

Gated recurrent units

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

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

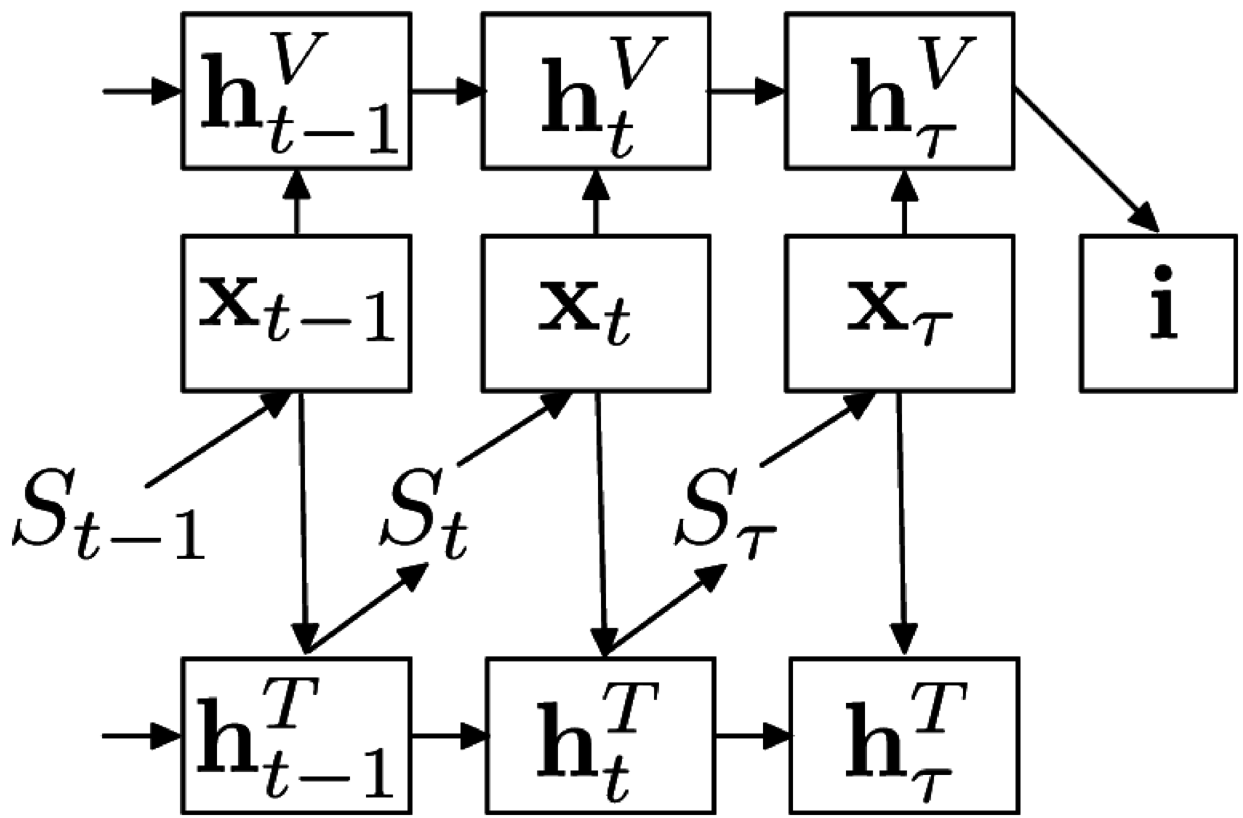
$$\tilde{h}_t^j = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1}))^j$$

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

Character level

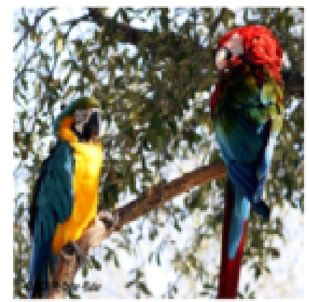
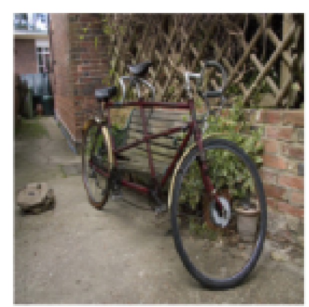
- character embeddings: 128 units
- GRUs: 1024
- Accuracy@5: 15%

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}
$\mathbf{h}_{t-1}^{T}$
$\mathbf{h}_{t}^{T}$
$\mathbf{h}_{\tau}^{T}$

Retrieving ImageNet pictures

		
Keyword:	*dessert*	*parrot*
Original label:	*ice cream*	*macaw*
Hypernym:	*dessert*	*parrot*
		
Keyword:	*locomotive*	*bicycle*
Original label:	*steam locomotive*	*bicycle-built-for-two*
Hypernym:	*locomotive*	*bicycle*

Model	Accuracy@5
Visual	0.38
MultiTask	0.38
LinReg	0.33

Cosine and standardization