Deep Learning for Natural Language Processing

Stephen Clark et al.
University of Cambridge and DeepMind

13. Sentence Representations

Felix Hill
DeepMind

What does a sentence mean?

Classical perspective

- Sentence representations are logical expressions.
- Sentence understanding is parsing and combining constituents to obtain logical form.
- Syntax guides semantics.

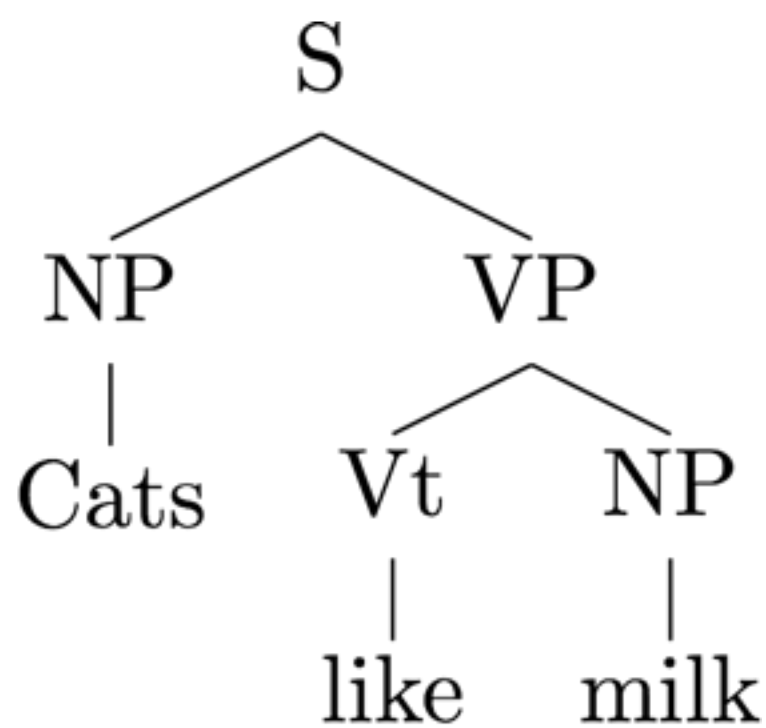
Classical perspective

Syntactic Analysis	Semantic Interpretation
S $\Rightarrow$ NP VP	$[\![VP]\!]([\![NP]\!])$
NP $\Rightarrow$ cats, milk, etc.	$[\![\text{cats}]\!], [\![\text{milk}]\!], \ldots$
VP $\Rightarrow$ Vt NP	$[\![Vt]\!]([\![NP]\!])$
Vt $\Rightarrow$ like, hug, etc.	$\lambda yx.[\![\text{like}]\!](x, y), \ldots$

DeepMind

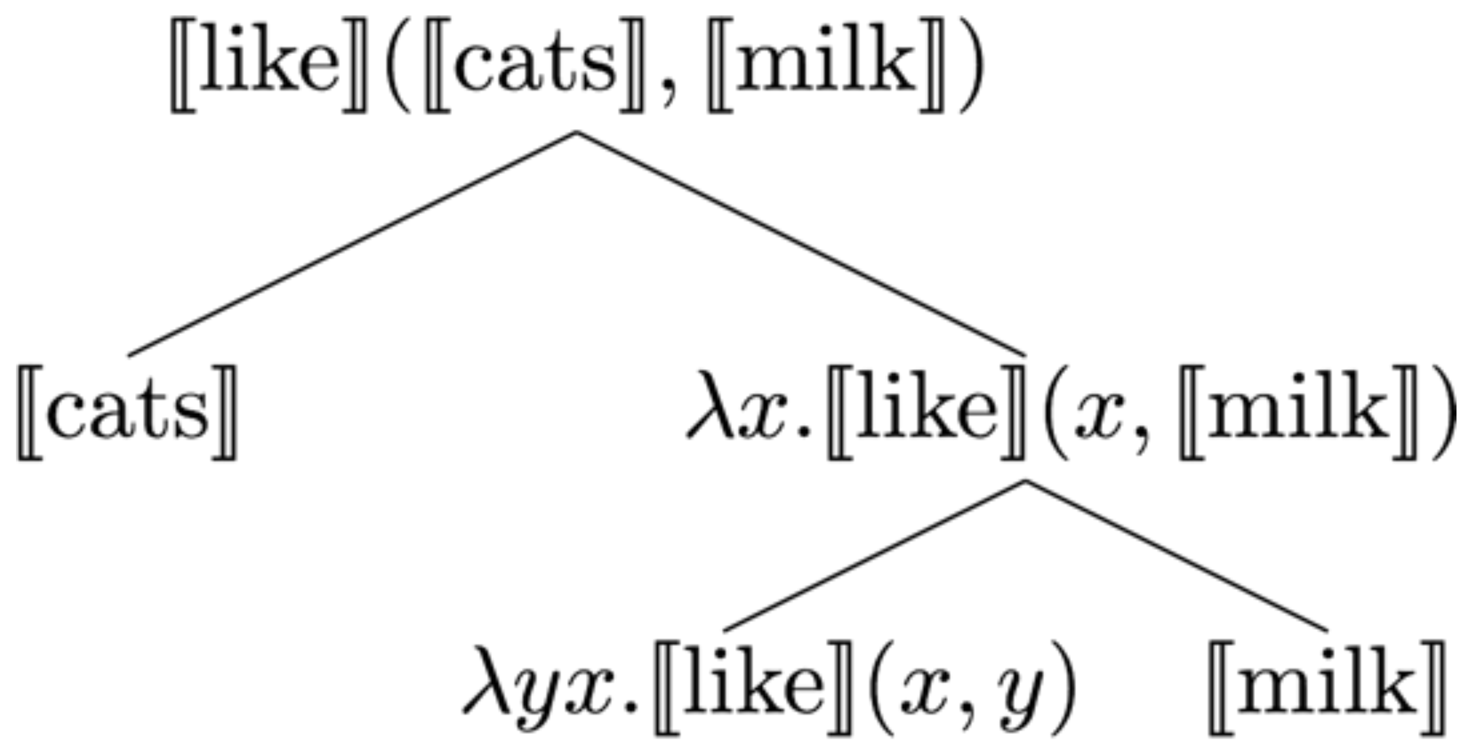
UNIVERSITY OF CAMBRIDGE

Classical perspective



Cats like milk.

Classical perspective



Cats like milk.

DeepMind

UNIVERSITY OF CAMBRIDGE

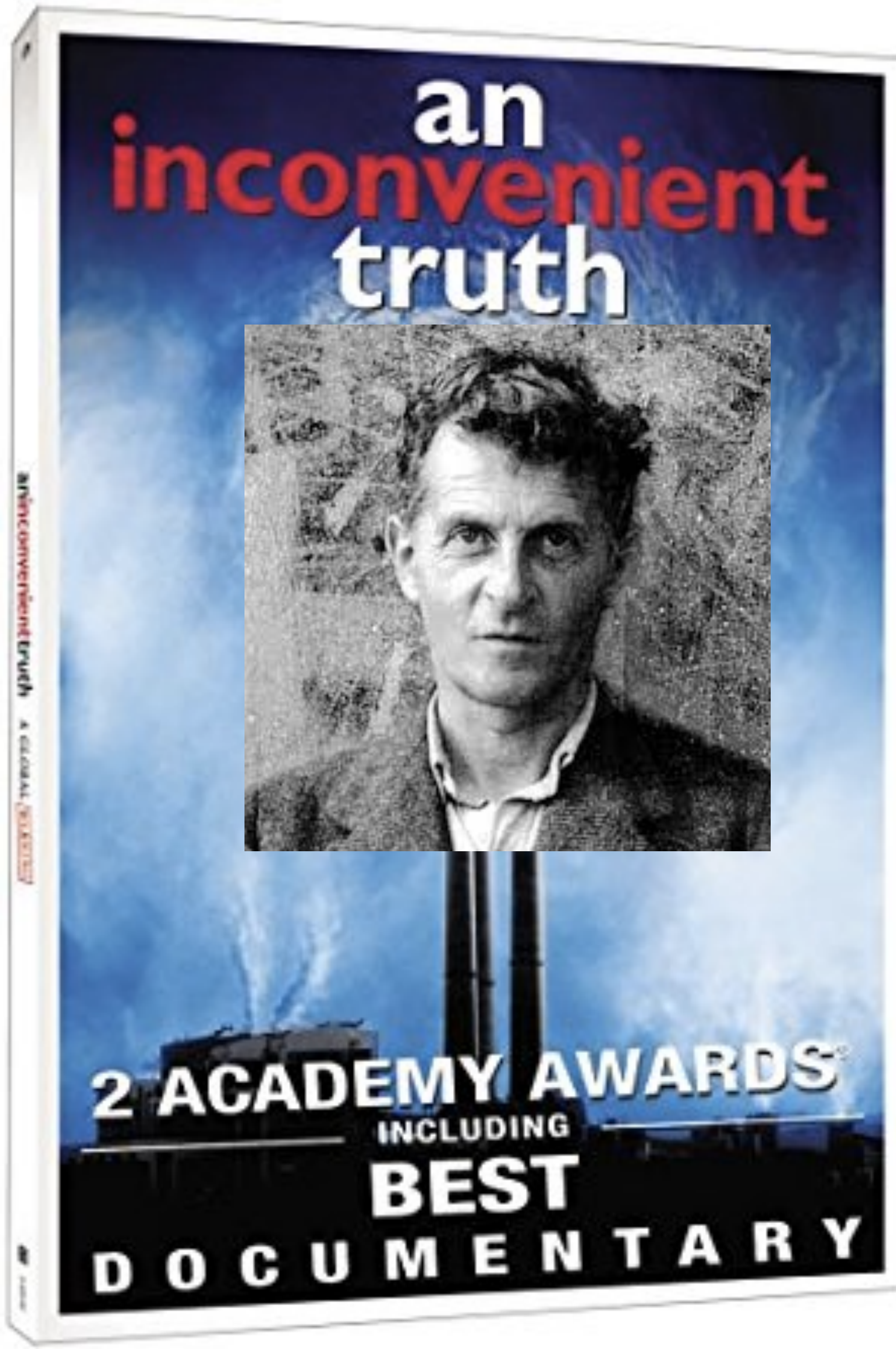
Classical perspective

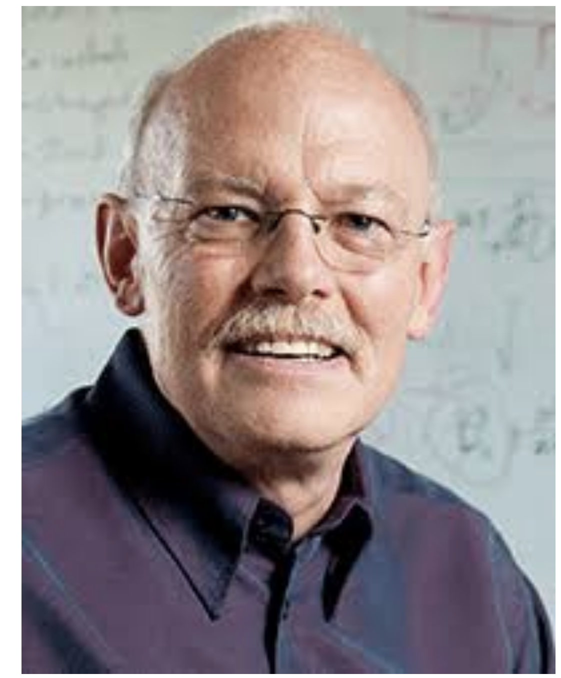
Pros:

- Intuitive and interpretable(?) representations.
- Leverage the power of predicate logic to model semantics
- Evaluate the truth of statements, derive conclusions, etc.

DeepMind

UNIVERSITY OF CAMBRIDGE





Thanks to Jay McClelland for examples

DeepMind

UNIVERSITY OF CAMBRIDGE

(1)

- “*John loves Mary”:*
  `loves(John, Mary)`

- “*John loves ice cream”*
  `loves(John, ice cream)`

(1)

- *"John loves Mary":*
  `loves(John, Mary)`

- *"John loves ice cream"*
  `loves(John, ice cream)`

All meaning is context-dependent

DeepMind

UNIVERSITY OF CAMBRIDGE

(2)

- "*the tiger threatens the giraffe":*
  `threatens(tiger, giraffe)`

- "*the protege threatens the master*"
  `threatens(protege, master)`

- "*the scandal threatens the profits*"
  `threatens(scandal, reputation)`

(2)

- *"Dave pushed the button":*
- *" Dave pushed the trainees":*
- *"Dave pushed the agenda "*
- *"Dave pushed the drugs"*

**Metaphoricity is the rule,
not the exception**

(3)

- “*the apple was in the container*”

- *“the juice was in the container”*

(3)

- *“the apple was in the container”*

- *“the juice was in the container”*

(3)

- *“the man cut his steak”*

DeepMind

UNIVERSITY OF CAMBRIDGE

(3)

- *"the man cut his steak"*

Where did you come from?

How many animals of each kind did Moses take on the Ark?

DeepMind

UNIVERSITY OF CAMBRIDGE

“The haystack was important because the cloth ripped.”

DeepMind

UNIVERSITY OF CAMBRIDGE

(3)

DeepMind

UNIVERSITY OF CAMBRIDGE

(3)

"The haystack was important because the cloth ripped."

Meaning is not in language, language indicates meaning

DeepMind

UNIVERSITY OF CAMBRIDGE

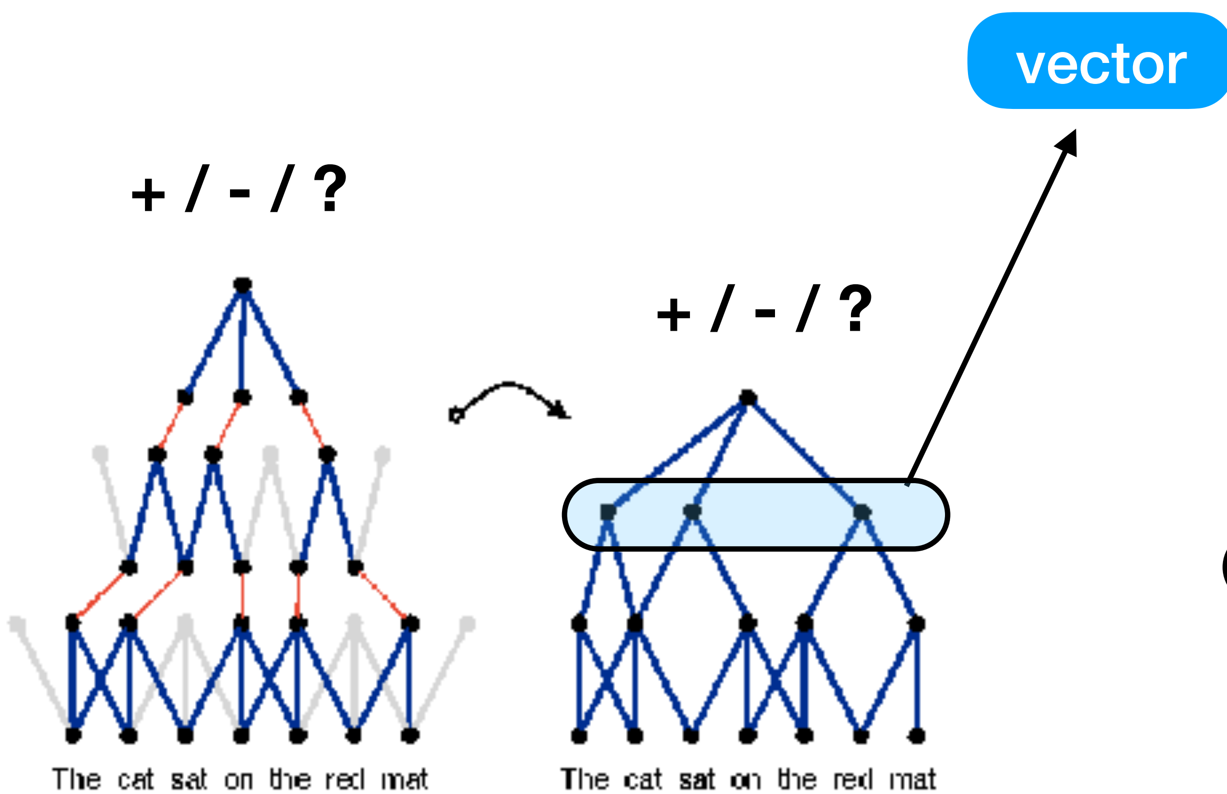
Neural networks to the rescue

- Nothing is an atom, everything a molecule (in theory)
- Linguistic signal (e.g. words), perceptual clues (e.g. vision) and semantic knowledge **all represented similarly**
- Representations of one information type **constrain and interact with** representations of others

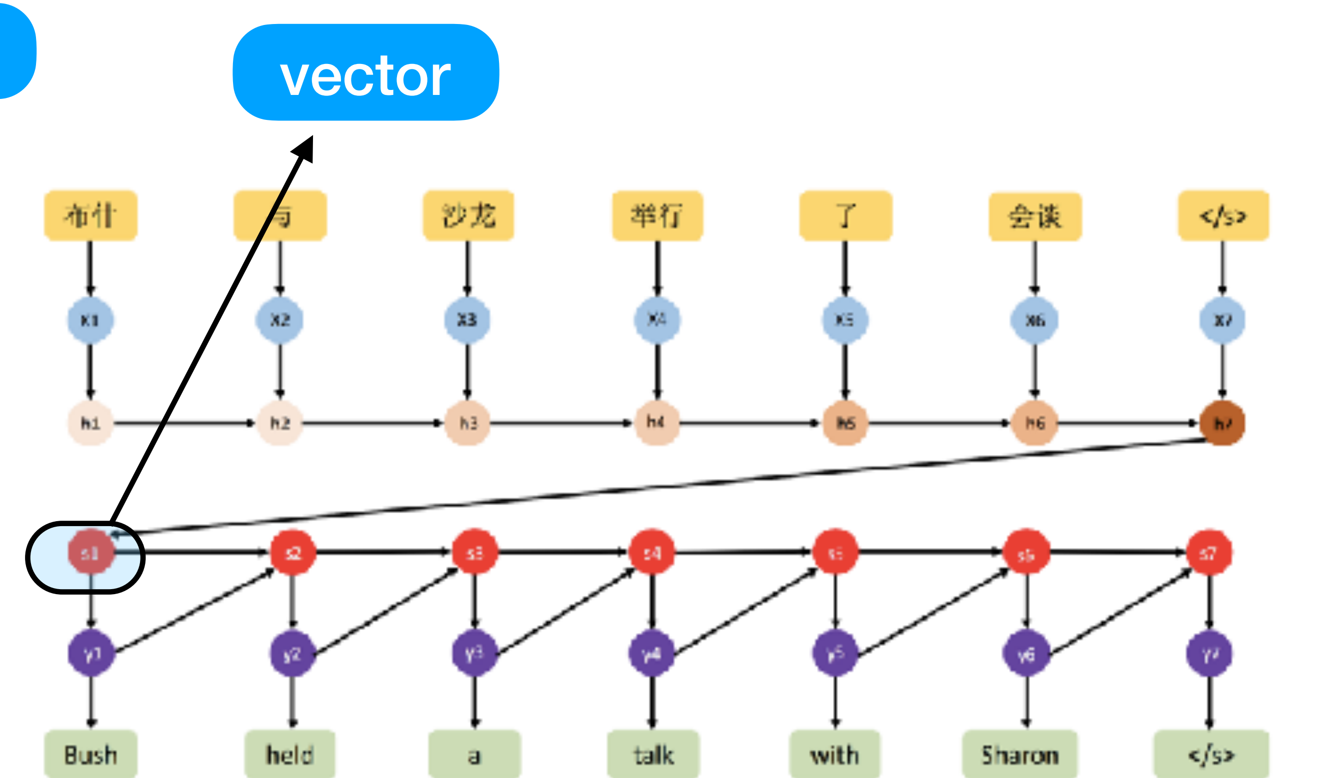
DeepMind

UNIVERSITY OF CAMBRIDGE

Sentence representations in neural nets



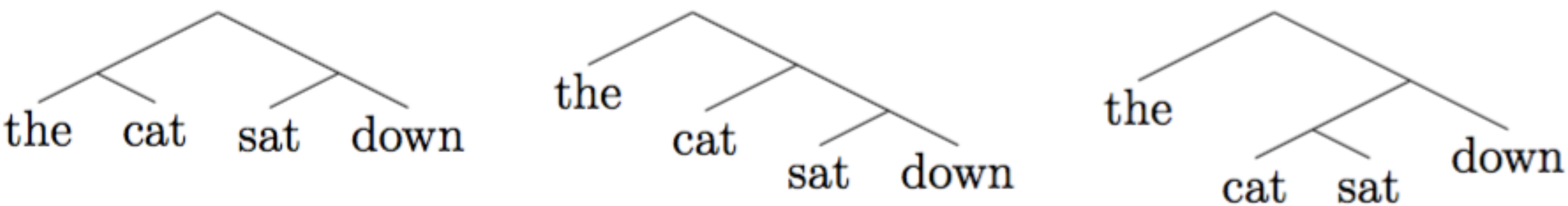
Kalchbrenner & Blunsom, 2014



[Sutskever et al., 2014]

Can we improve on this?

One approach..



[SHIFT, SHIFT, REDUCE, SHIFT, SHIFT, REDUCE, REDUCE]

[SHIFT, SHIFT, SHIFT, SHIFT, REDUCE, REDUCE, REDUCE]

[SHIFT, SHIFT, SHIFT, REDUCE, SHIFT, REDUCE, REDUCE]

Figure due to Sam Bowman.
Reproduced with author's permission.

Stanford NLI task

(1) *the man inspects a painting in a museum*

(2) *the man is sleeping*

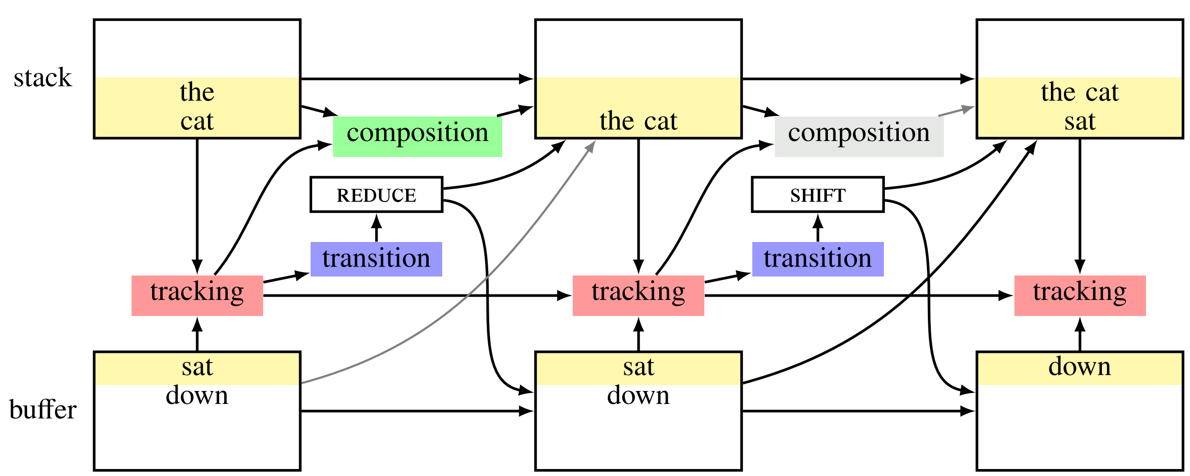
CONTRADICTION

DeepMind

UNIVERSITY OF CAMBRIDGE

Bowman et al. 2015 (see also Socher et al. 2013)

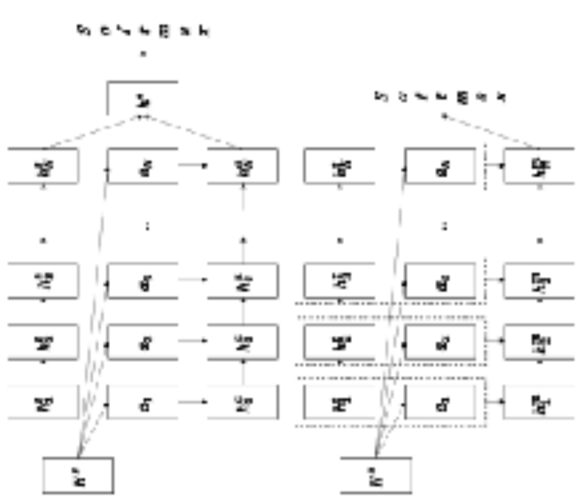
Model	Params.	Trans. acc. (%)	Train acc. (%)	Test acc. (%)
Previous non-NN results				
Lexicalized classifier (Bowman et al., 2015a)	—	—	99.7	78.2
Previous sentence encoder-based NN results				
100D LSTM encoders (Bowman et al., 2015a)	221k	—	84.8	77.6
1024D pretrained GRU encoders (Vendrov et al., 2016)	15m	—	98.8	81.4
300D Tree-based CNN encoders (Mou et al., 2016)	3.5m	—	83.4	82.1
Our results				
300D LSTM RNN encoders	3.0m	—	83.9	80.6
300D SPINN-PI-NT (*parsed input, no tracking*) encoders	3.4m	—	84.4	80.9
300D SPINN-PI (*parsed input*) encoders	3.7m	—	89.2	**83.2**
300D SPINN (unparsed input) encoders	2.7m	92.4	87.2	82.6



+ $$$$$$$$$$

Wang and Jiang (2015)

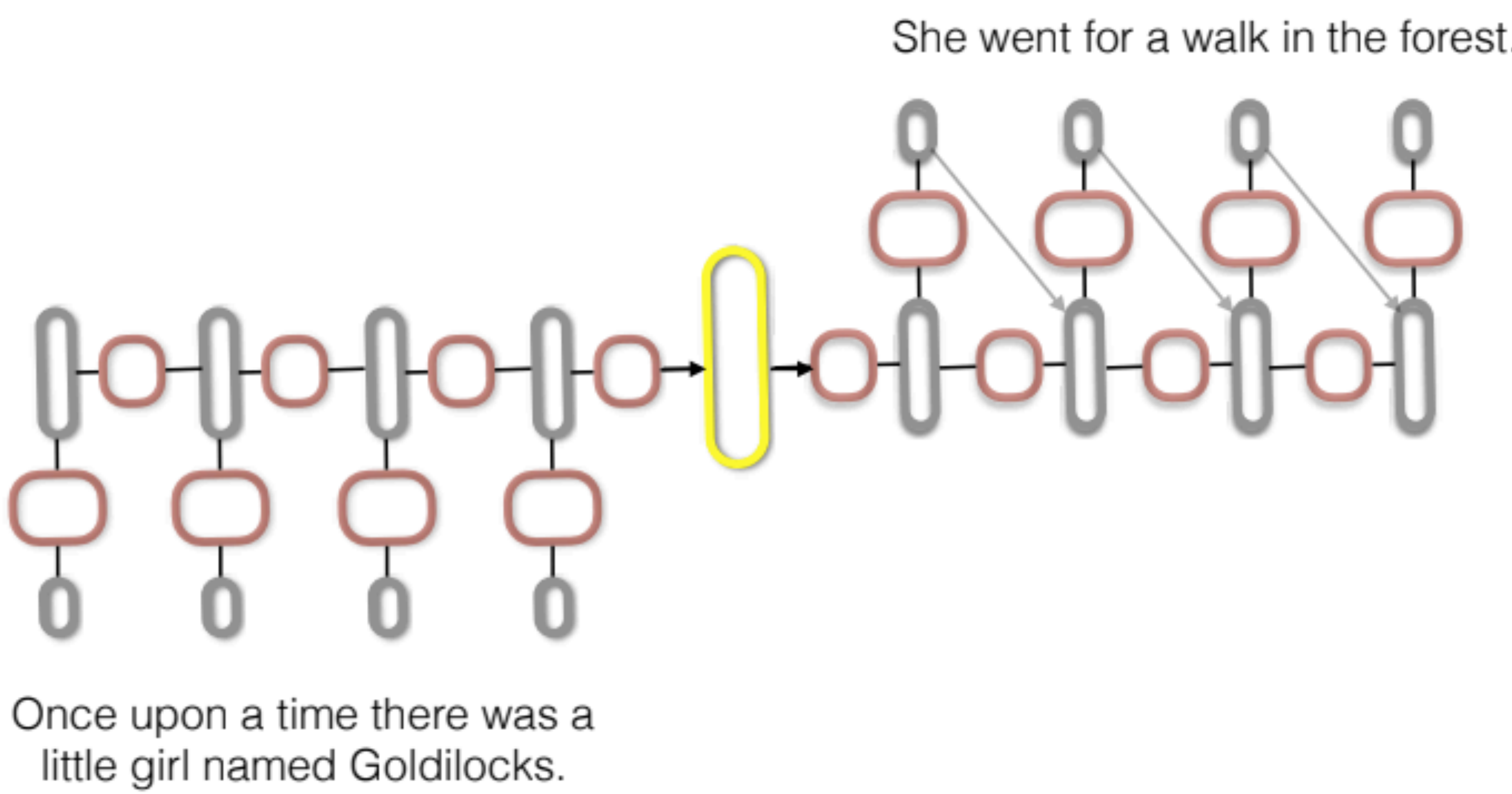
Model	d	$\|\theta\|_{W+M}$	$\|\theta\|_M$	Train	Dev	Test
LSTM [Bowman et al. (2015)]	100	10M	221K	84.4	-	77.6
Classifier [Bowman et al. (2015)]	-	-	-	99.7	-	78.2
LSTM shared [Rocktäschel et al. (2015)]	159	3.9M	252K	84.4	83.0	81.4
Word-by-word attention [Rocktäschel et al. (2015)]	100	3.9M	252K	85.3	83.7	83.5
Word-by-word attention (our implementation)	150	340K	340K	85.5	83.3	82.6
*m*LSTM	150	544K	544K	91.0	86.2	85.7
*m*LSTM with bi-LSTM sentence modeling	150	1.4M	1.4M	91.3	86.6	86.0
*m*LSTM	300	1.9M	1.9M	92.0	**86.9**	**86.1**
*m*LSTM with word embedding	300	1.3M	1.3M	88.6	85.4	85.3



Parallel interactive processing wins

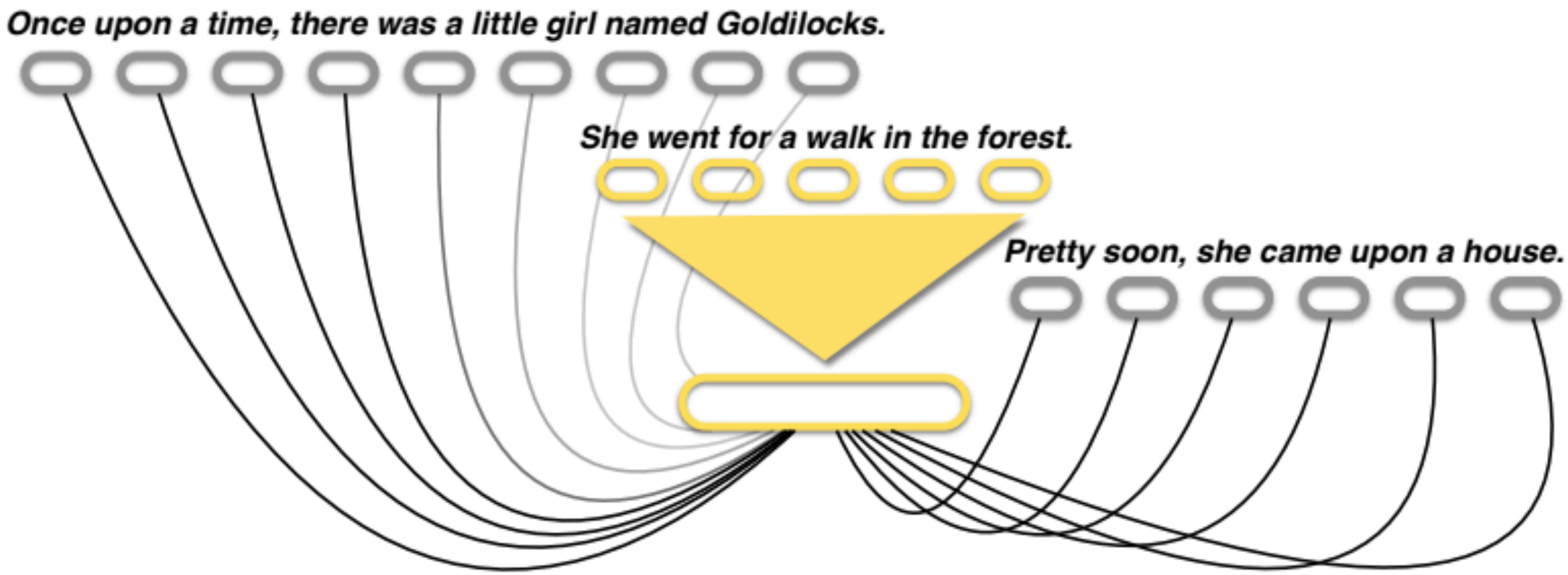
perception and
conceptual knowledge?

Knowledge from stories



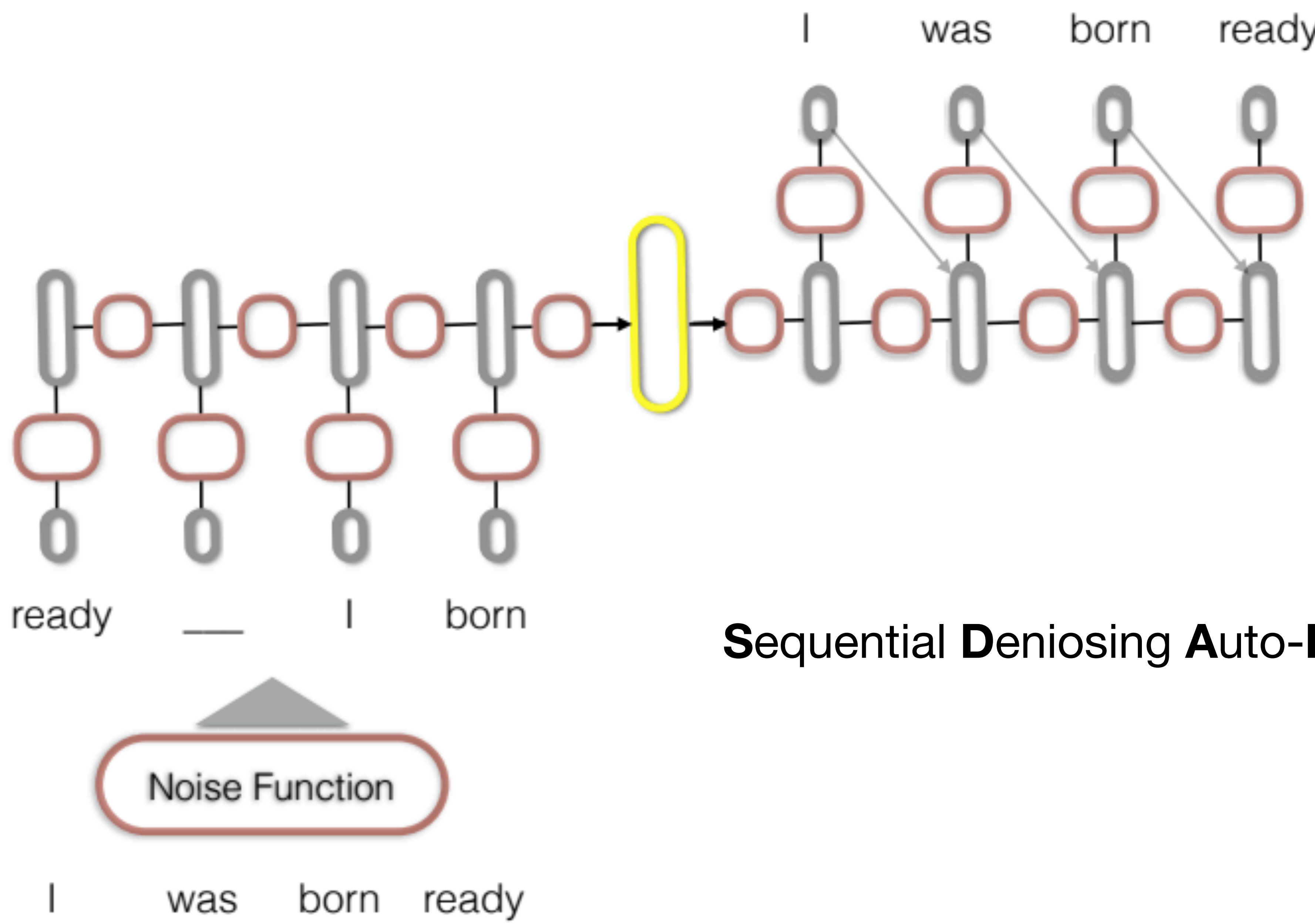
"Skip-Thought Vectors"
Kiros et al. 2015

Fast knowledge from stories



"Learning distributed representations of sentences from unlabelled data"
Hill et al. 2015

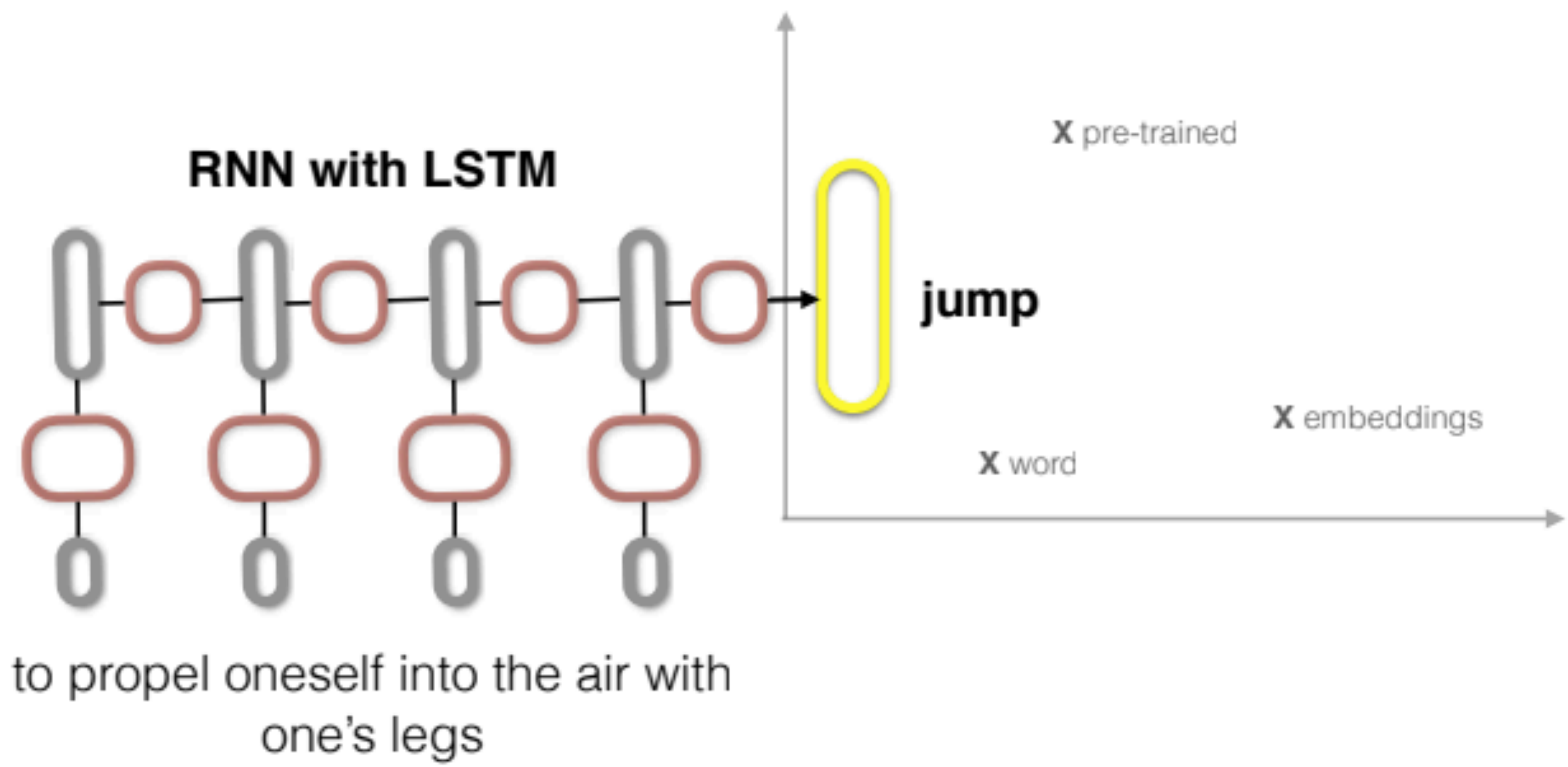
Knowledge from raw text



Sequential **D**eniosing **A**uto-**E**ncoder

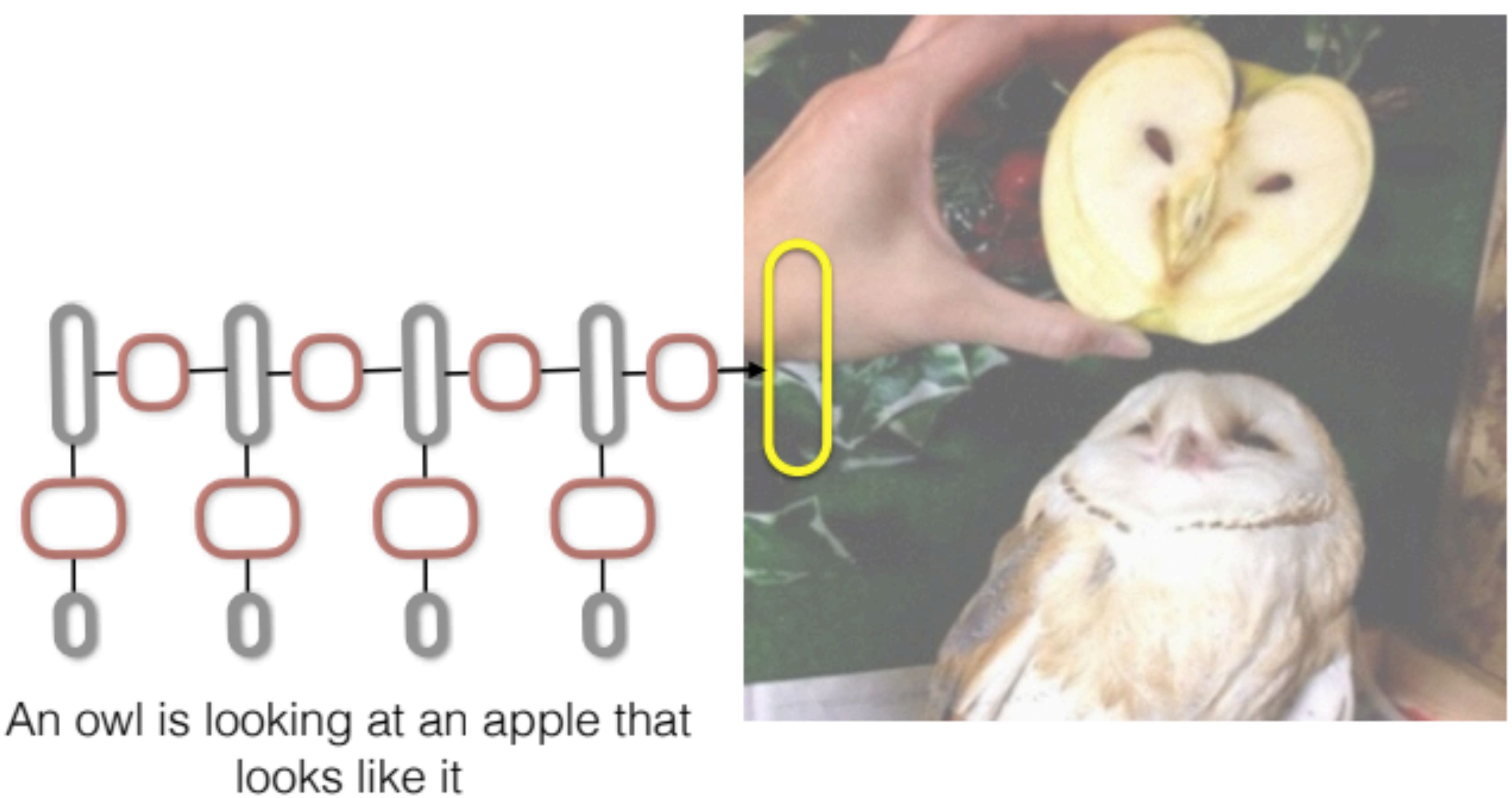
"Learning distributed representations of sentences from unlabelled data"
Hill et al. 2015

Knowledge from dictionaries



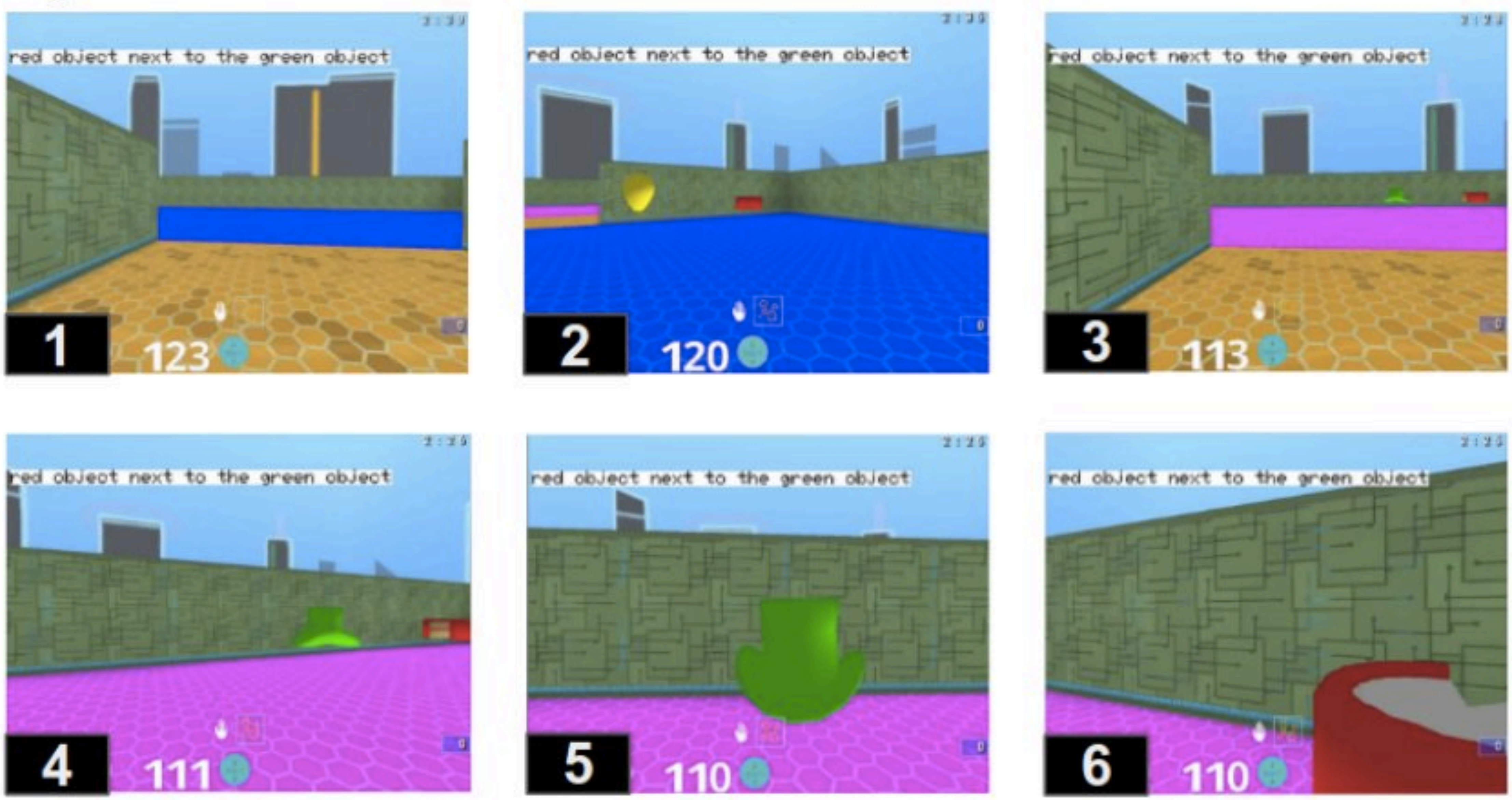
"Learning distributed representations of sentences from unlabelled data"
Hill et al. 2015

Knowledge from images?



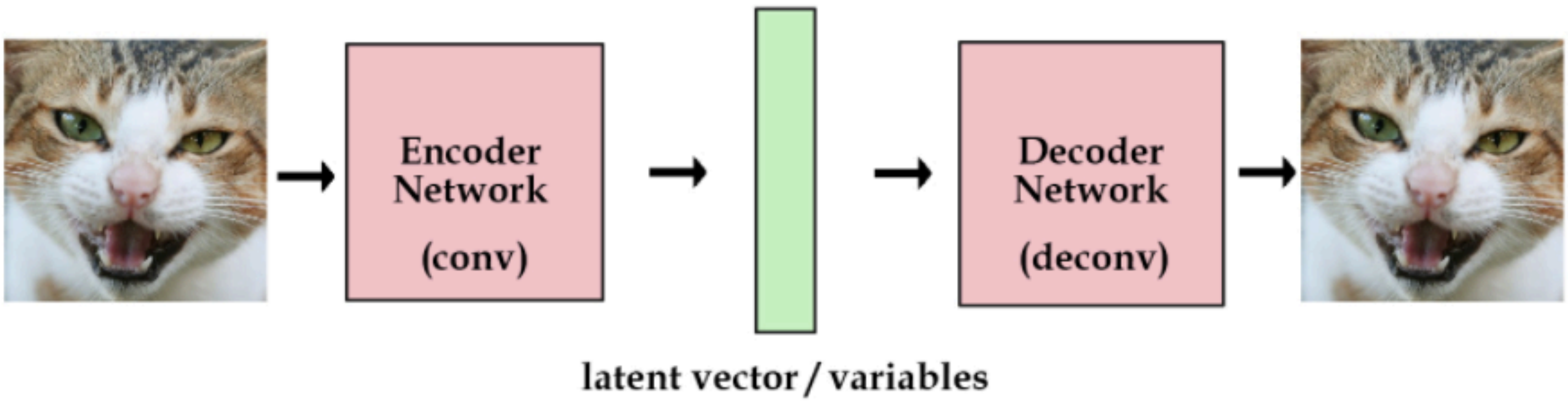
"Learning distributed representations of sentences from unlabelled data"
Hill et al. 2015

Next time: full “embodiment”



Richer representation spaces

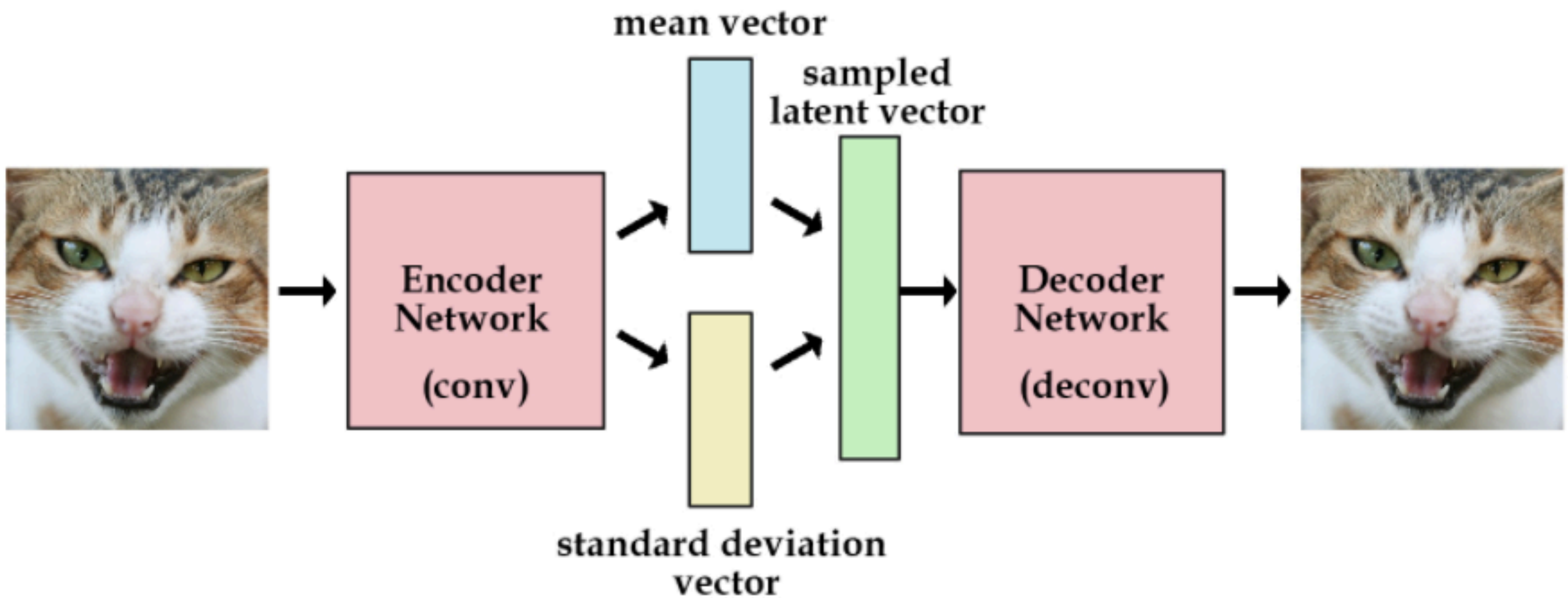
Auto-encoding for representation learning



loss = pixel reconstruction loss

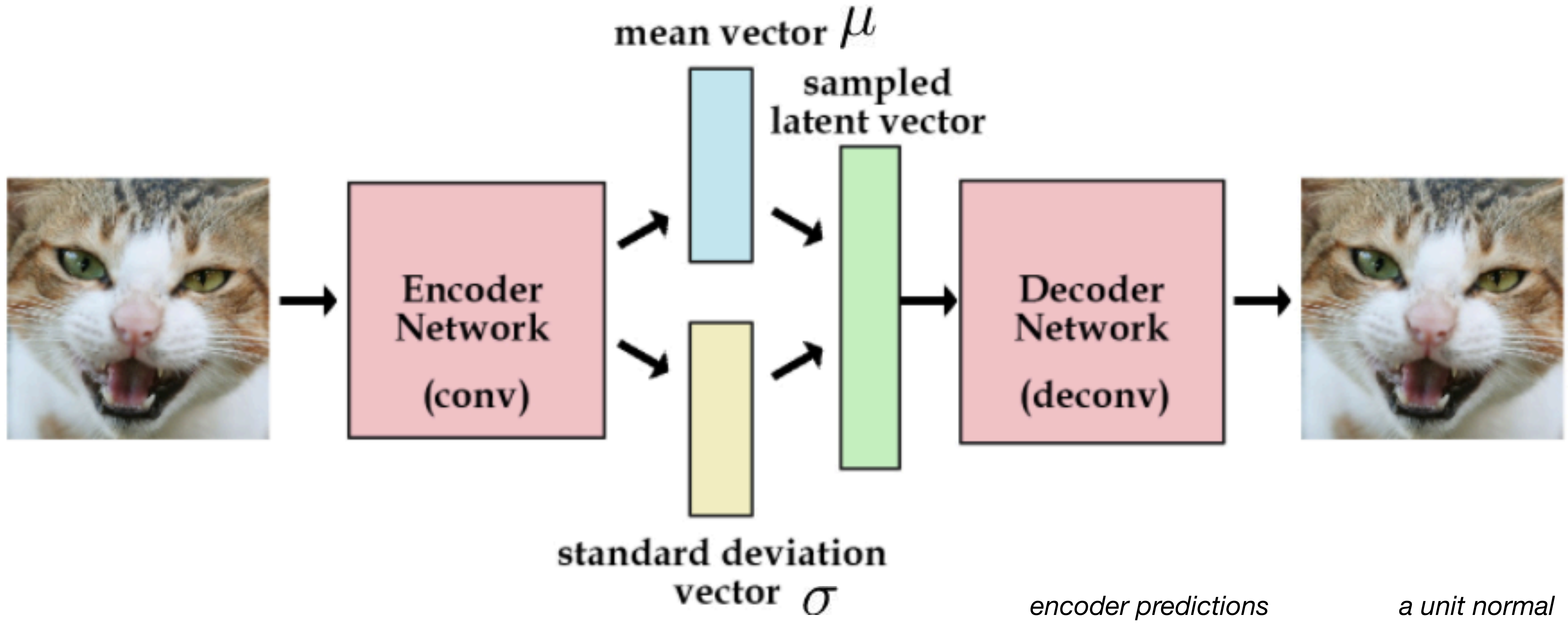
http://kvfrans.com/variational-autoencoders-explained/

Auto-encoding via a richer latent space



http://kvfrans.com/variational-autoencoders-explained/

VAE: variational auto-encoder

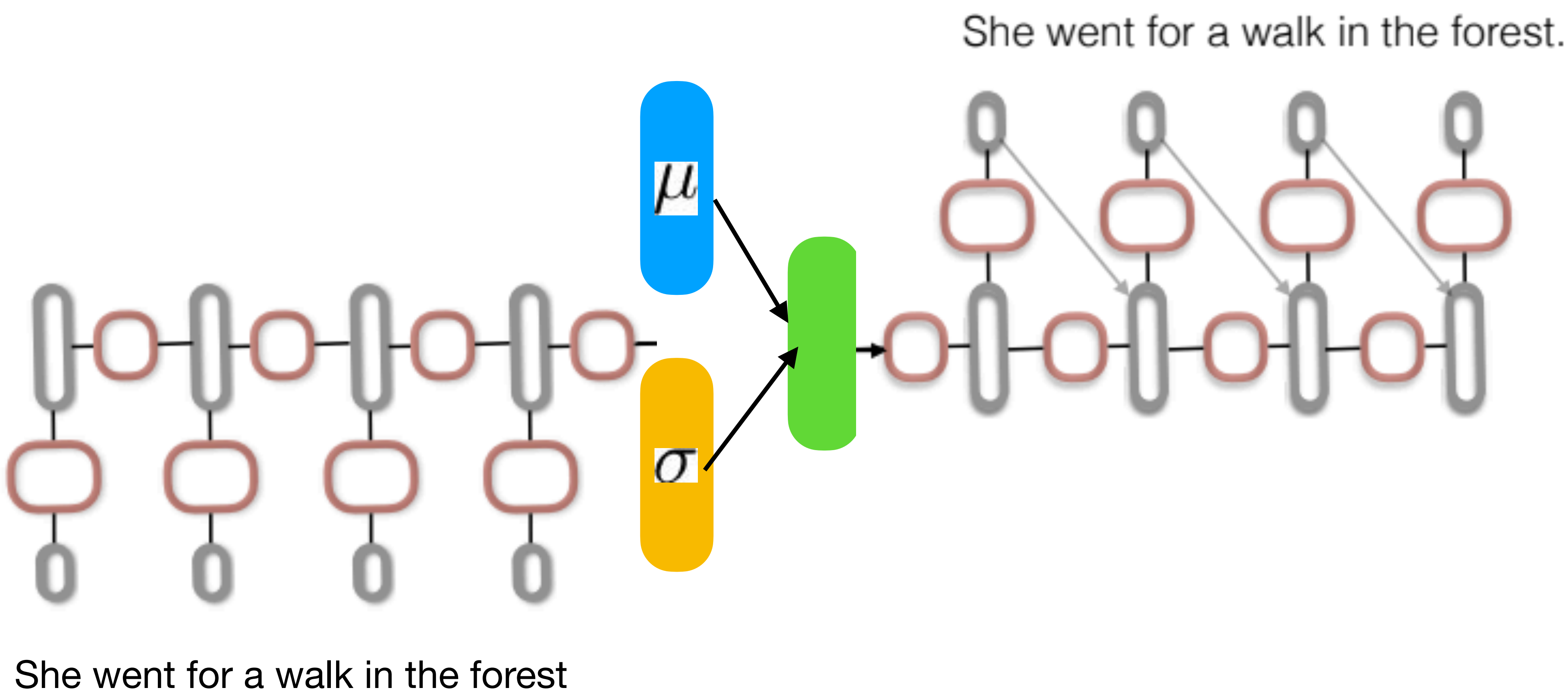


encoder predictions of mean and variance

a unit normal distribution

loss = pixel reconstruction loss + $\mathbf{KL}(\mathcal{N}(\mu,\ \sigma^2), \mathcal{N}(0,\ 1))$

VAE for text



She went for a walk in the forest.
μ
σ
She went for a walk in the forest

Benefits of VAE

1. Smooth(er) latent space of representations

i went to the store to buy some groceries .
i store to buy some groceries .
i were to buy any groceries .
horses are to buy any groceries .
horses are to buy any animal .
horses the favorite any animal .
horses the favorite favorite animal .
horses are my favorite animal .

2. Generate from the model

Conclusions

- The meaning of language is not in the language itself
- Neural networks provide a model for combining the necessary information sources
- Finding and using the right information is just as important as elaborate modelling

Reading

Formal semantics: Montague, R. (1970). English as a formal language.

Meaning in context: McClelland, J. L. (1992). Can connectionist models discover the structure of natural language?

Dictionary definitions to guide meaning: Hill, F, Cho, K and Korhonen, A. Learning to Understand Phrases by Embedding the Dictionary *TACL*. (2015).

Skip-Thought Vectors: Kiros, R. et al. (NIPS 2015)

Comparison of sentence representations (SDAE, FastSent): Hill, F, Cho, K and Korhonen, A. Learning Distributed Representations of Sentences from Unlabelled Data, *NAACL*. (2015).

Variational AutoEncoder: Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Variational AutoEncoder for sentences: Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., & Bengio, S. (2015). Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349.

DeepMind

UNIVERSITY OF CAMBRIDGE