Insight
Centre for Data Analytics

#InsightDL2017

DEEP LEARNING WORKSHOP

Dublin City University
27-28 April 2017

Day 2 Lecture 10

Neural Machine Translation

Xavier Giro-i-Nieto
xavier.giro@upc.edu

Associate Professor
Universitat Politecnica de Catalunya
Technical University of Catalonia

DCU

UPC

Acknowledgments

Antonio Bonafonte

Santiago Pascual

TALP
UPC
TALP Research Center

Acknowledgments

Marta R. Costa-jussà

TALP
UPC
TALP Research Center

Acknowledgments

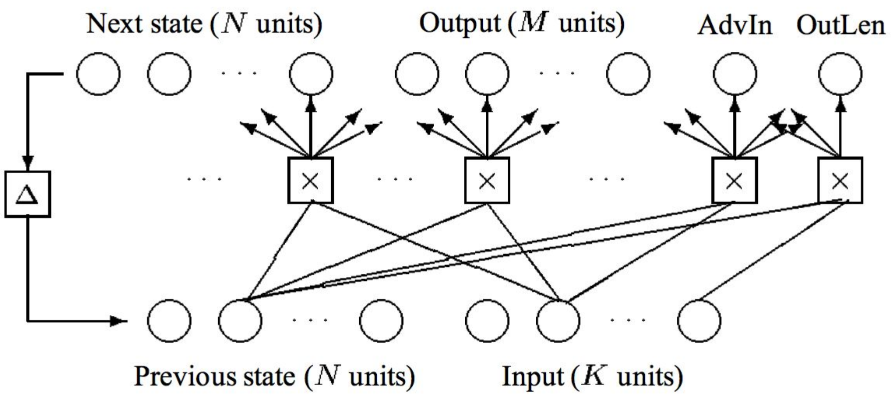
[Kyunghyun Cho]

NYU | COURANT INSTITUTE OF MATHEMATICAL SCIENCES

Predecents of Neural Machine Translation

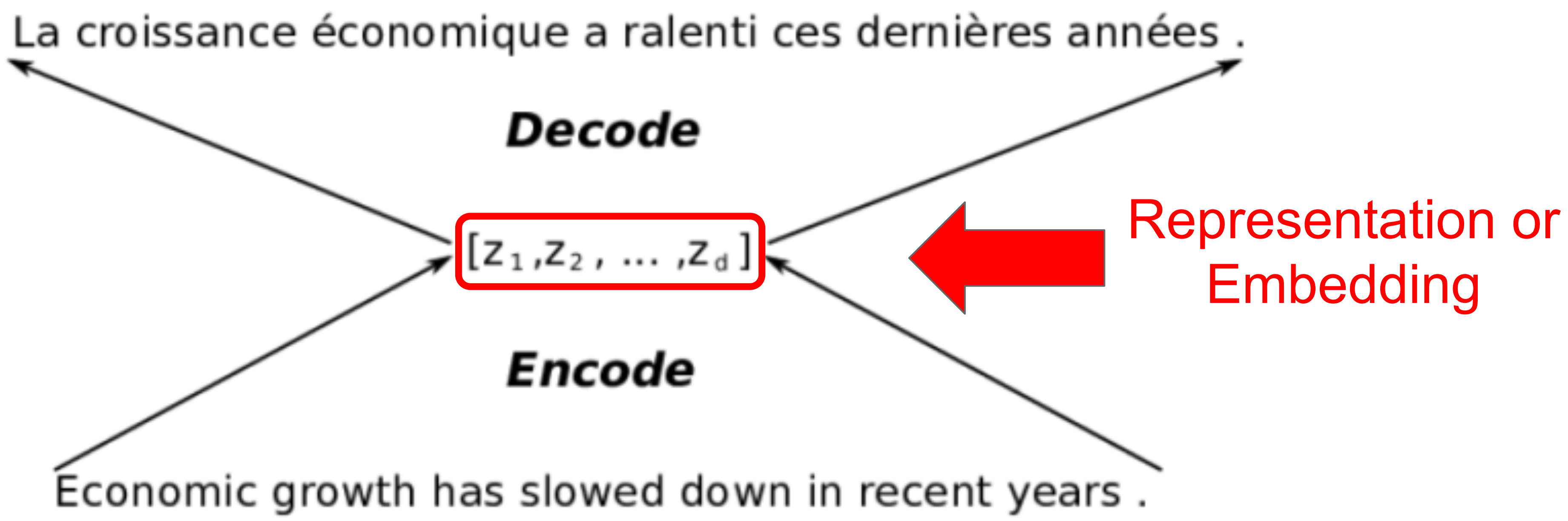
Asynchronous translations with recurrent neural nets

Ramón P. Ñeco, Mikel L. Forcada
Departament de Llenguatges i Sistemes Informàtics,
Universitat d'Alacant, E-03071 Alacant, Spain.
E-mail: {neco,mlf}@dlsi.ua.es



Neco, R.P. and Forcada, M.L., 1997, June. Asynchronous translations with recurrent neural nets. In Neural Networks, 1997., International Conference on (Vol. 4, pp. 2535-2540). IEEE.

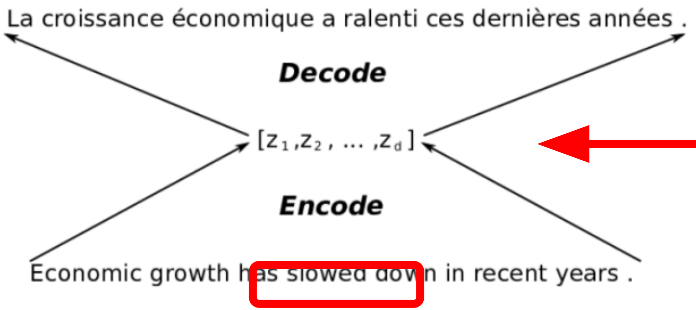
Encoder-Decoder



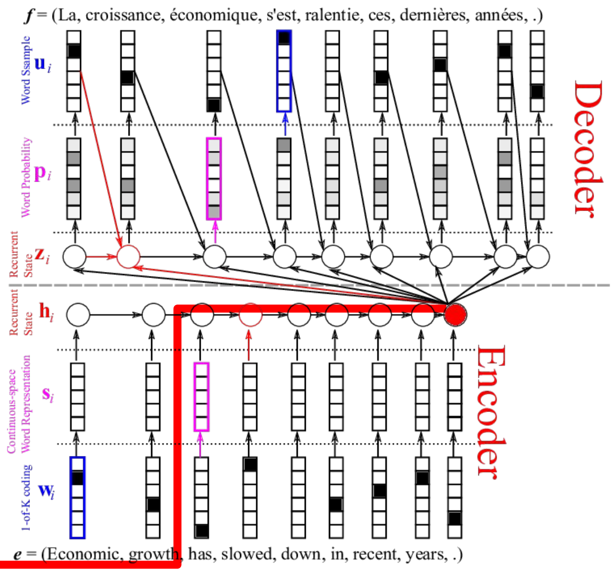
Cho, Kyunghyun, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. "On the properties of neural machine translation: Encoder-decoder approaches." SSST-8 (2014).

Encoder-Decoder

Front View



Side View



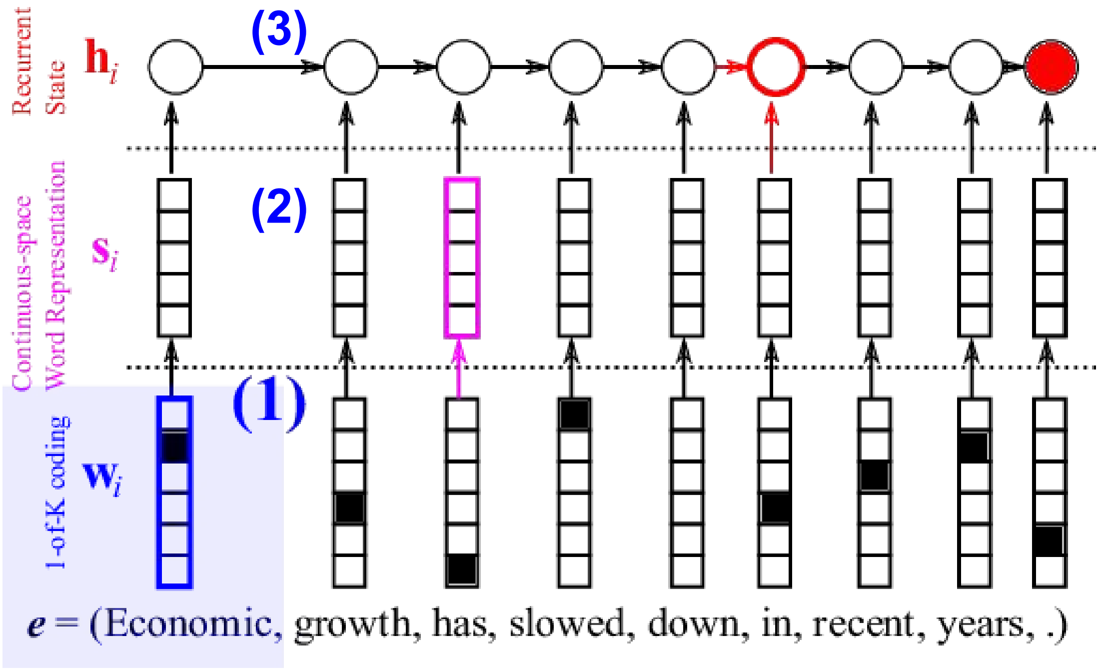
Representation of the sentence

Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. ["Learning phrase representations using RNN encoder-decoder for statistical machine translation."]() arXiv preprint arXiv:1406.1078 (2014).

Encoder

Encoder in three steps



(1) One hot encoding

(2) Continuous-space Word Representation (word embedding)

(3) Sequence summarization

Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

Step 1: One-hot encoding

Example: letters. |V| = 30

```
‘a’: x = 1
‘b’: x = 2
‘c’: x = 3
.
.
.

‘.’: x = 30
```

Step 1: One-hot encoding

Example: letters. |V| = 30

```
'a': x = 1
'b': x = 2
'c': x = 3
.
.
.

'.': x = 30
```

We impose fake range ordering

Step 1: One-hot encoding

Example: letters. |V| = 30

```
'a': x^T = [1,0,0, ..., 0]
'b': x^T = [0,1,0, ..., 0]
'c': x^T = [0,0,1, ..., 0]
.
.
.

'.': x^T = [0,0,0, ..., 1]
```

Step 1: One-hot encoding

Example: words.

```
cat: x^T = [1,0,0, ..., 0]
dog: x^T = [0,1,0, ..., 0]
.
.
house: x^T = [0,0,0, …,0,1,0,...,0]
.
.
.
```

Number of words, |V| ?

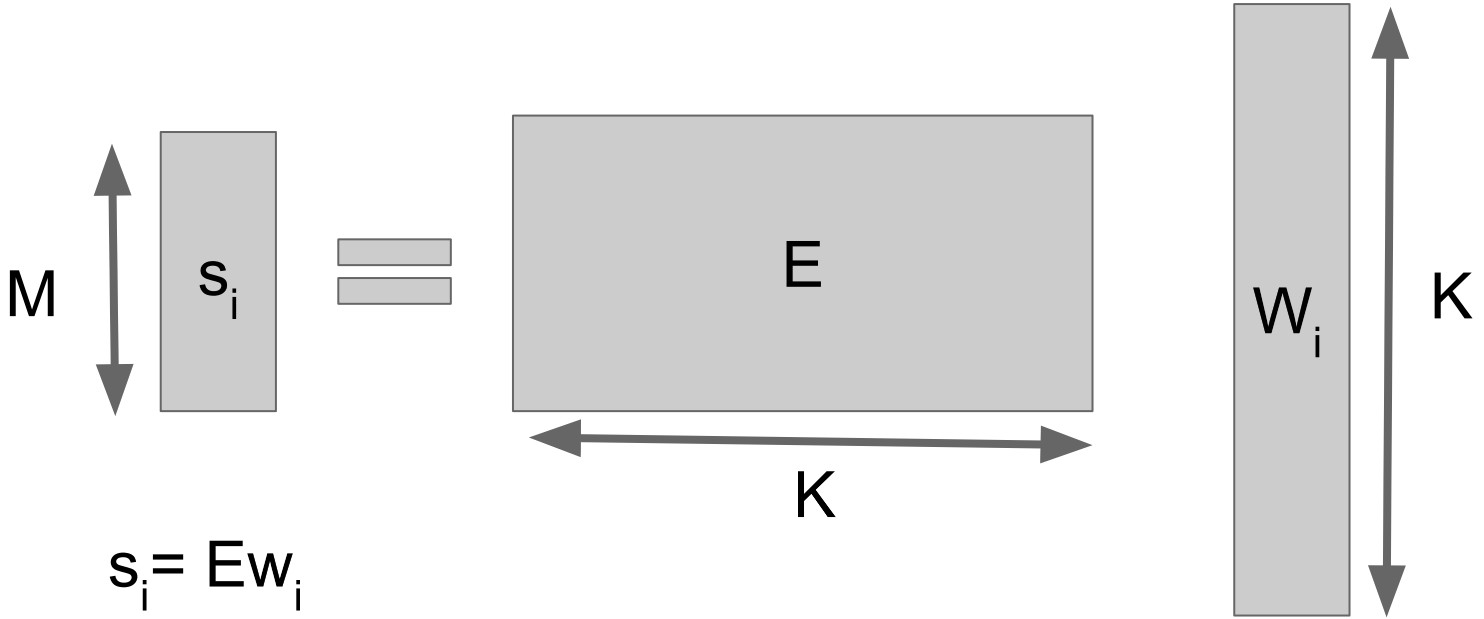
B2: 5K
C2: 18K
LVSR: 50-100K
Wikipedia (1.6B): 400K
Crawl data (42B): 2M

Step 1: One-hot encoding

- Large dimensionality
- Sparse representation (mostly zeros)
- Blind representation
  - Only operators: ‘!=’ and ‘==’

Step 2: Projection to word embedding

The one-hot is linearly projected to a embedded space of lower dimension with matrix E for learned weights (=fully connected).



$s_i = Ew_i$

Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”] (2015)

Step 2: Projection to word embedding

Word embeddings

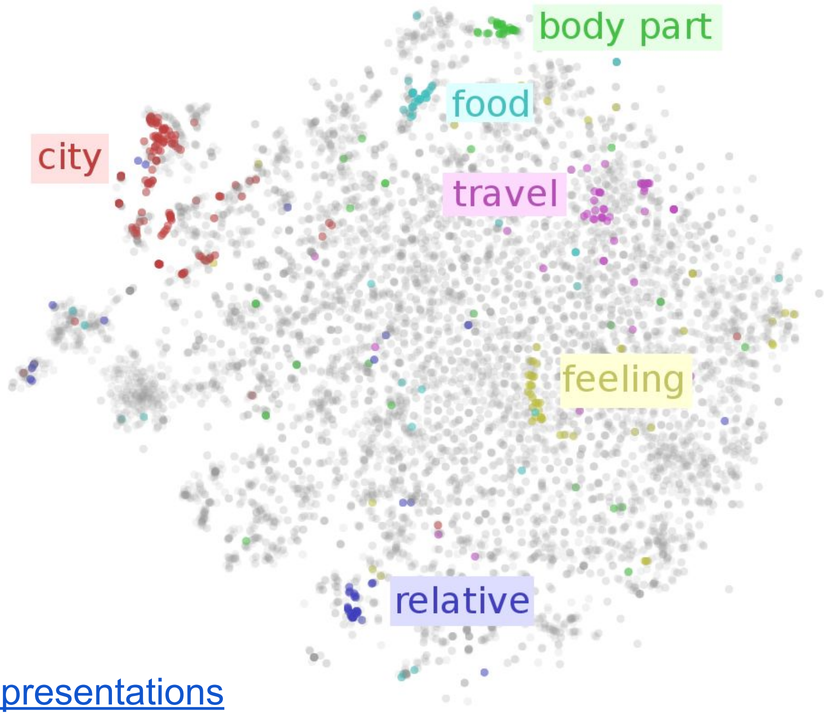
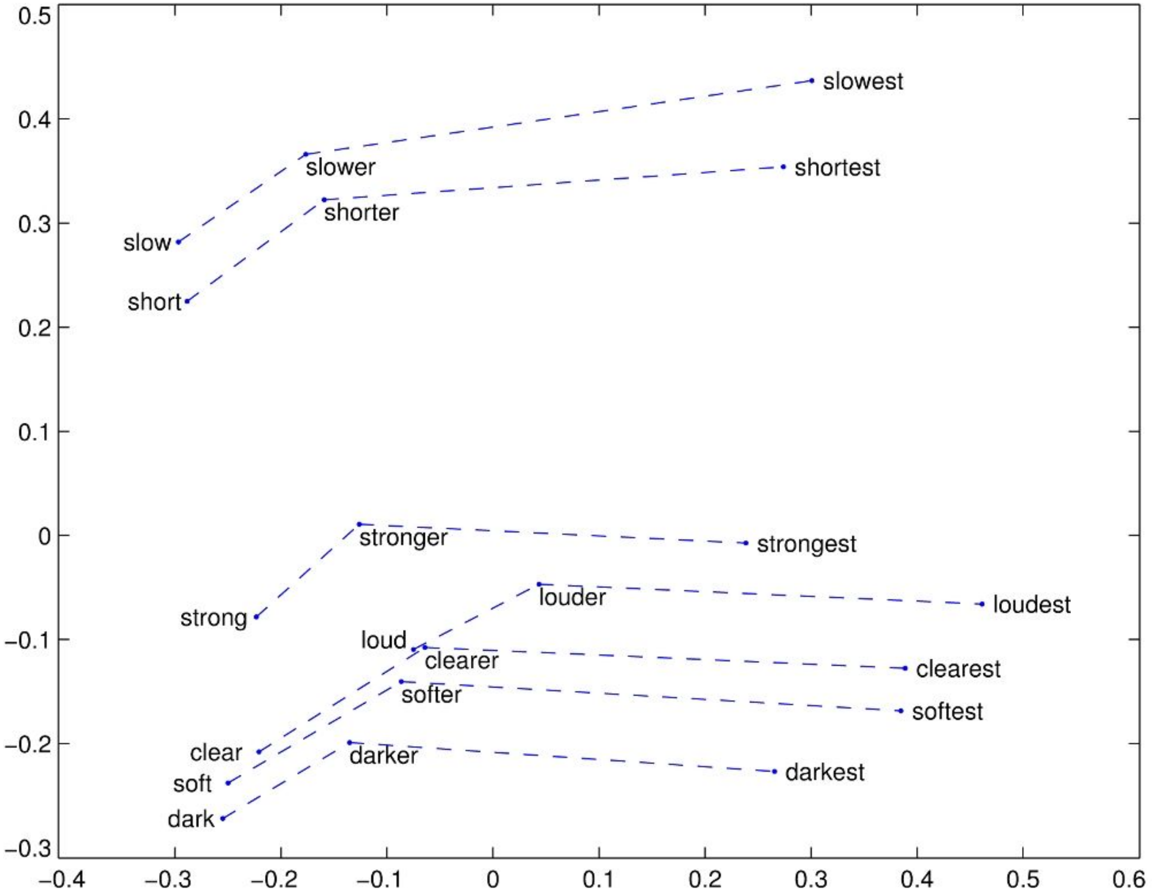


Figure: Christopher Olah, [Visualizing Representations]

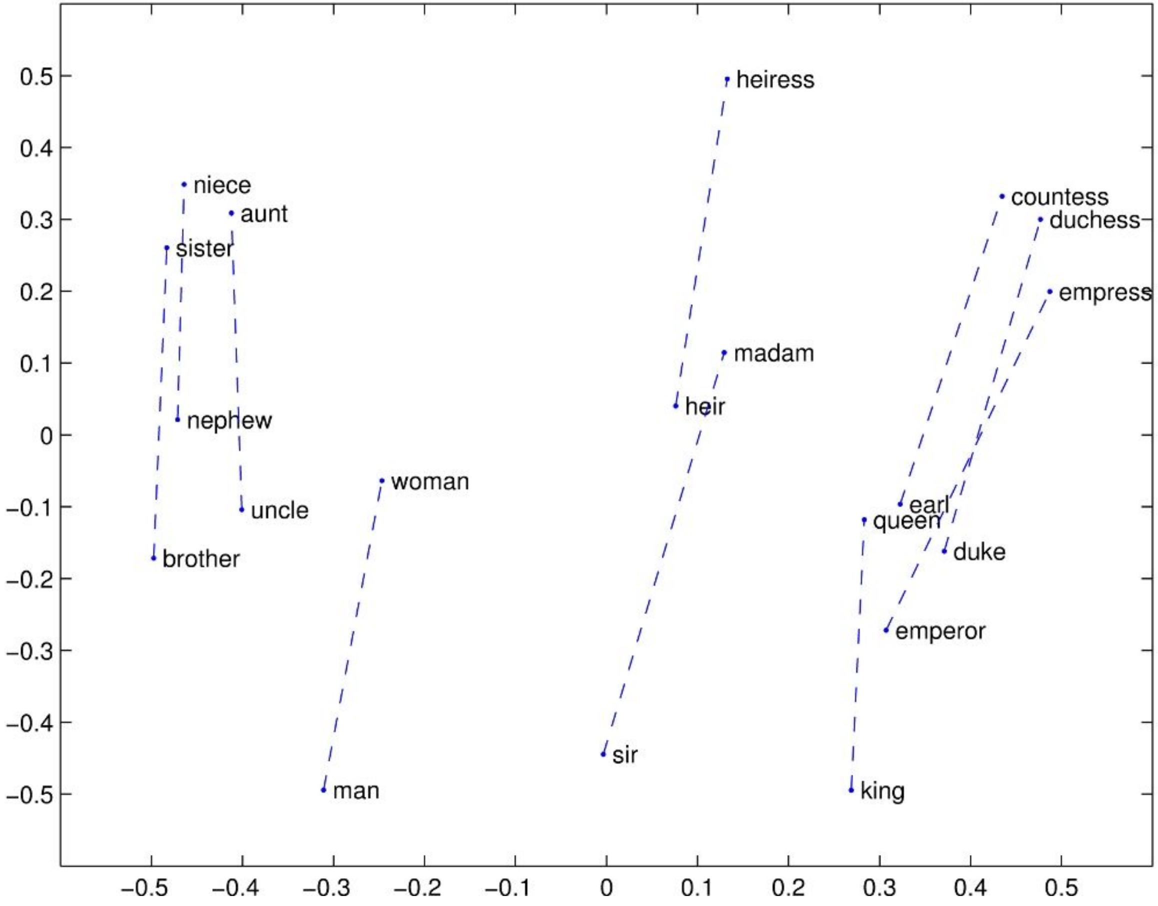
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. ["Distributed representations of words and phrases and their compositionality."] NIPS 2013

Step 2: Projection to word embedding



GloVe (Stanford)

Step 2: Projection to word embedding



GloVe (Stanford)

Step 2: Projection to word embedding

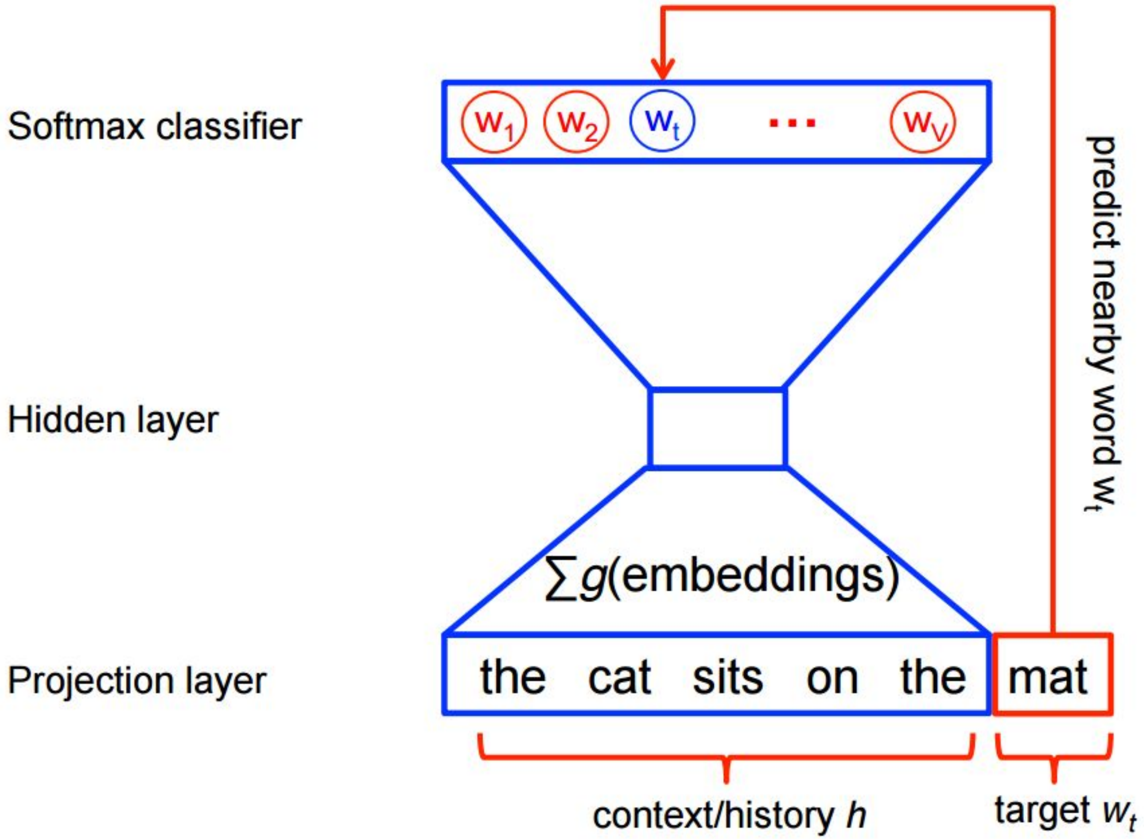
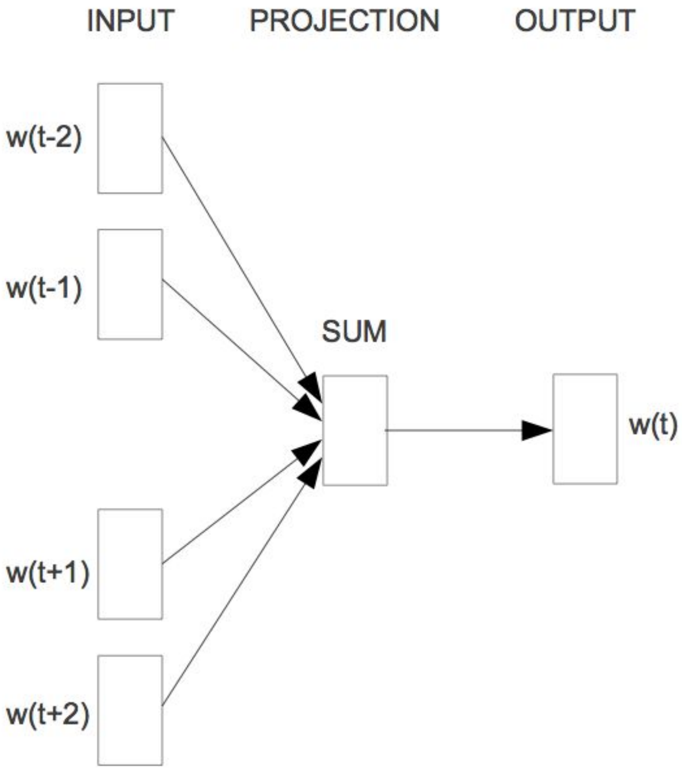


Figure:
[TensorFlow tutorial]

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. ["A neural probabilistic language model."] Journal of machine learning research 3, no. Feb (2003): 1137-1155.

Step 2: Projection to word embedding

Word2Vec: Continuous Bag of Words



`the cat` $\boxed{\texttt{climbed}}$ `a tree`

Given context:

`a, cat, the, tree`

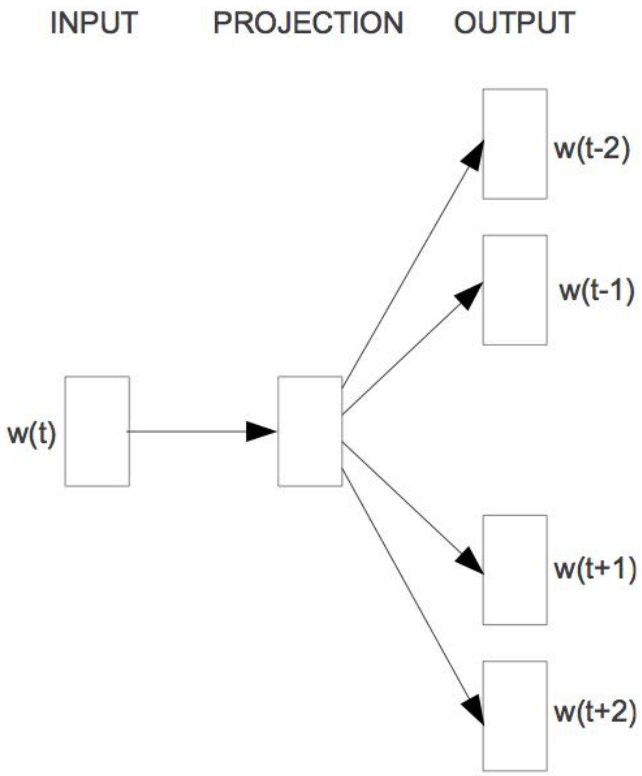
Estimate prob. of

`climbed`

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. ["Distributed representations of words and phrases and their compositionality."]() NIPS 2013

Step 2: Projection to word embedding

Word2Vec:
Skip-gram



the cat climbed a tree

Given word:

climbed

Estimate prob. of context words:

a, cat, the, tree

(It selects randomly the context length, till max of 10 left + 10 right)

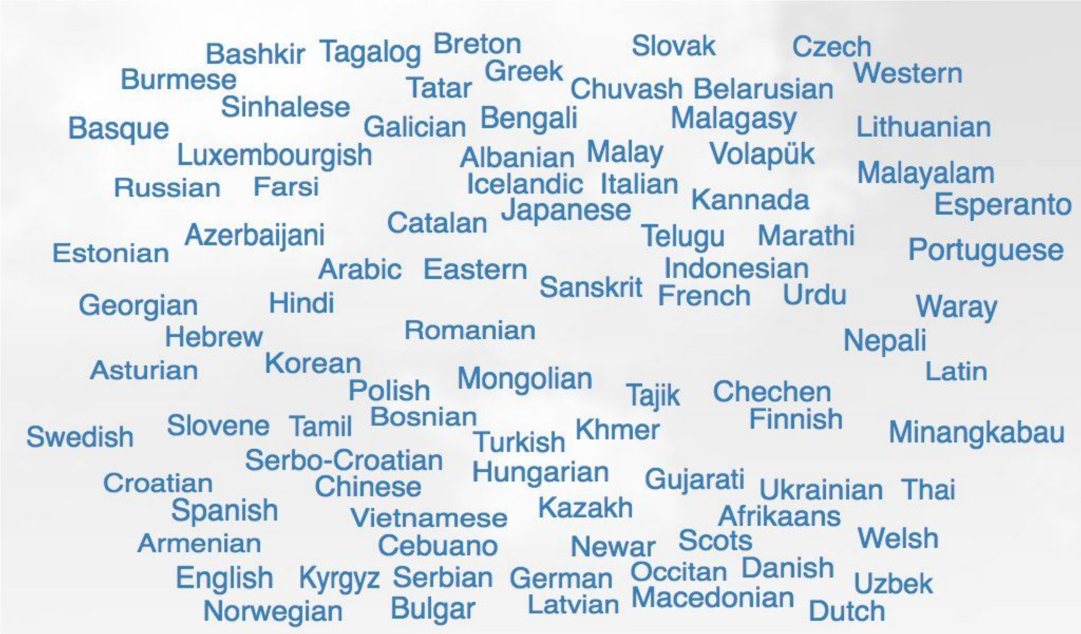
Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." NIPS 2013

Step 2: Projection to word embedding

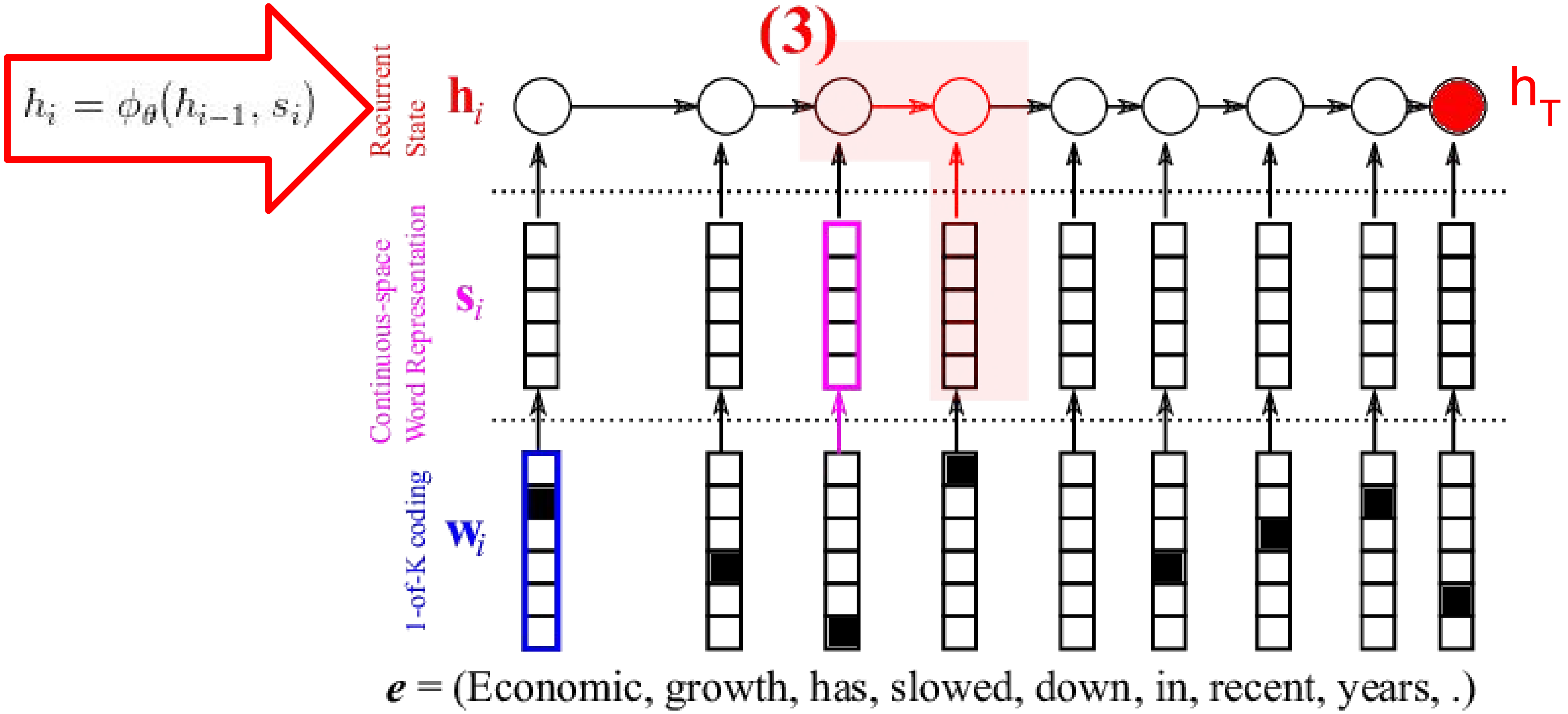
- Represent words using vectors of dimension *d* (~100 - 500)
- Meaningful (semantic, syntactic) distances
- Dominant research topic in last years in NLP conferences (EMNLP)
- *Good* embeddings are useful for *many* other tasks

Step 2: Projection to word embedding

Pre-trained Word Embeddings for 90 languages trained using FastText, on Wikipedia.

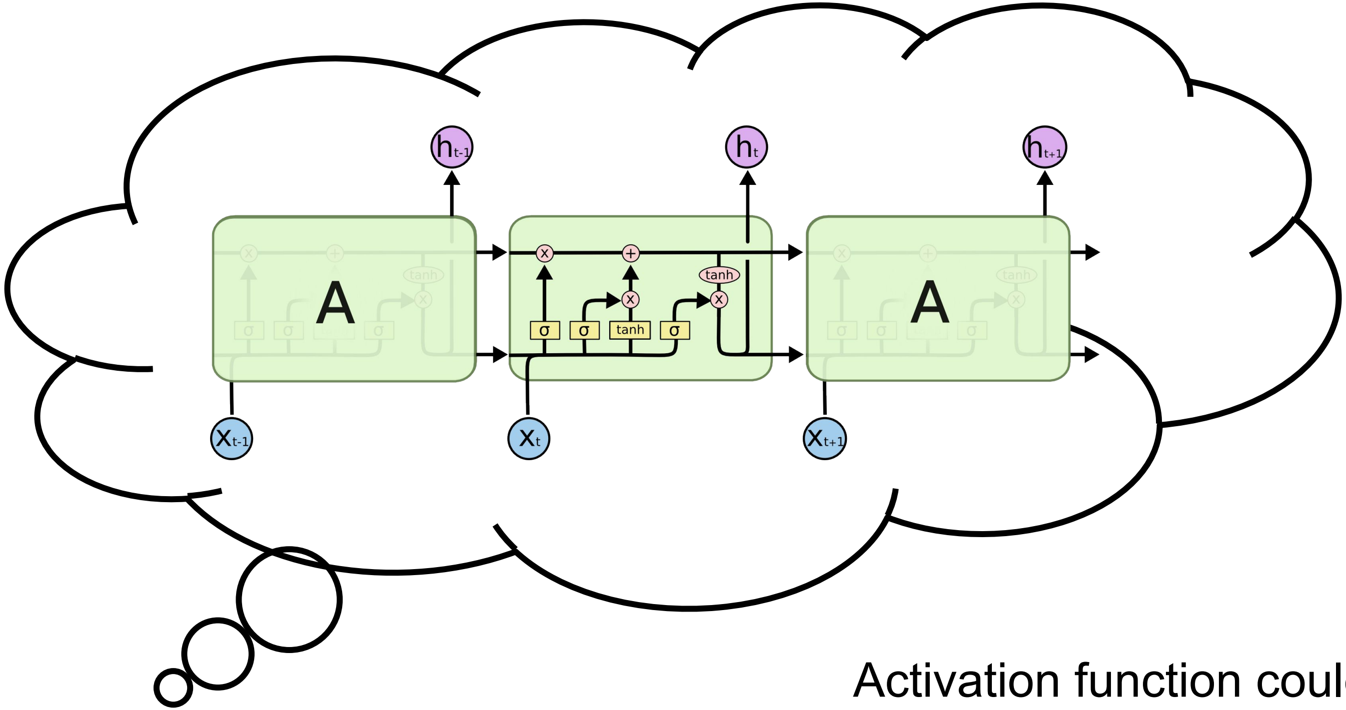


Step 3: Recurrence



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”] (2015)

Step 3: Recurrence



Sequence

Activation function could be LSTM, GRU, QRNN, pLSTM...

Figure: Cristopher Olah, [“Understanding LSTM Networks”]() (2015)

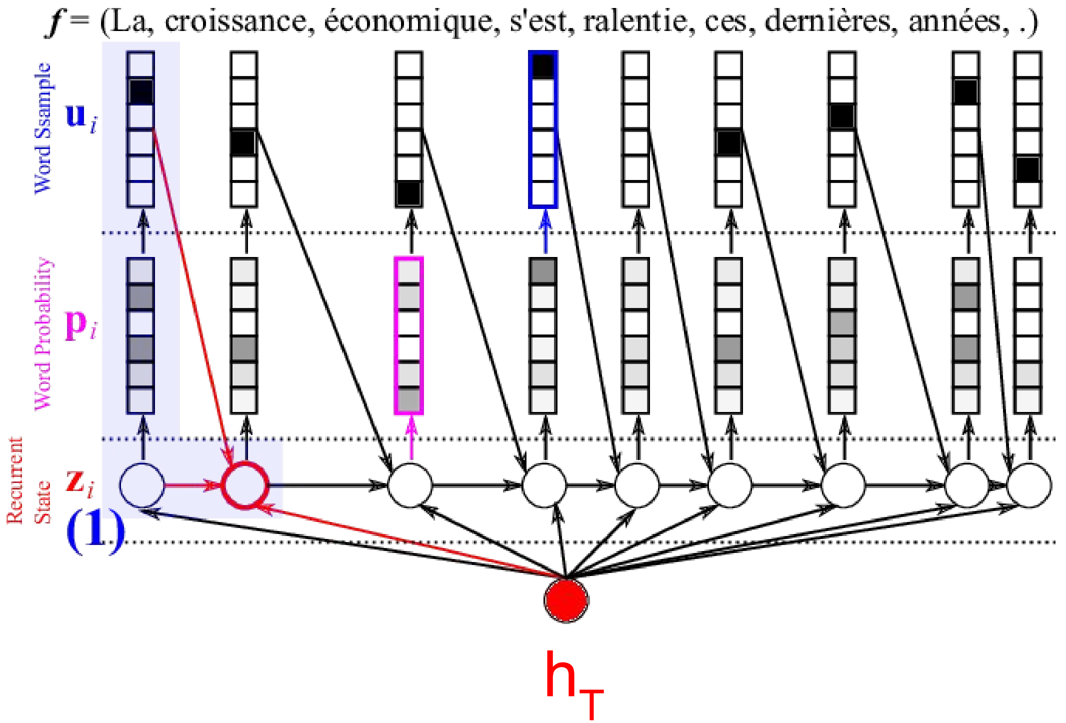
Decoder

Decoder

The Recurrent State (z_i) of the decoder is determined by:

1) summary vector h_T
2) previous output word u_{i-1}
3) previous state z_{i-1}

$$z_i = \phi_{\theta'}(h_T, u_{i-1}, z_{i-1}).$$



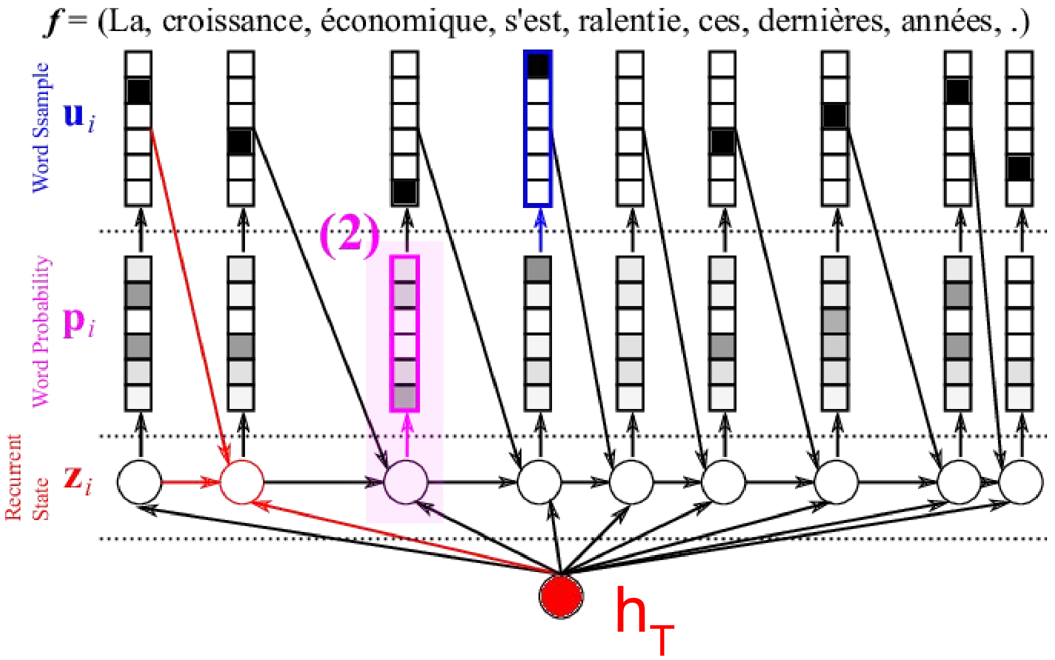
Decoder

With z_i ready, we can compute a **<u>score e(k)</u>** for each word *k* in the vocabulary with a dot product with the Recurrent State (z_i):

$$e(k) = w_k^\top z_i + b_k,$$

Neuron weights for word k (output layer)

RNN internal state



Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

Decoder

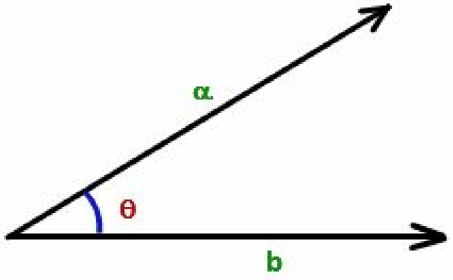
A score e(k) is higher if neuron weights for word k (w_k) and the decoder's internal state z_i are similar to each other.

$$e(k) = w_k^\top z_i + b_k,$$

Neuron weights for word k (output layer)

RNN internal state

<u>Reminder</u>:
a **dot product** computes the length of the projection of one vector onto another. Similar vectors (nearly parallel) the projection is longer than if they are very different (nearly perpendicular)



$$a.b = |a|\ |b|\ \text{Cos}\theta$$

Decoder

Given the score for word k...

$$e(k) = w_k^\top z_i + b_k,$$

...we can finally normalize to word probabilities with a softmax.

Probability that the output word at timestep i (w_i) is word k

$$p(w_i = k | \underbrace{w_1, w_2, \ldots, w_{i-1}}_{\text{Previous words}}, \underbrace{h_T}_{\text{Hidden state}}) = \frac{\exp(e(k))}{\sum_j \exp(e(j))}.$$

Bridle, John S. "Training Stochastic Model Recognition Algorithms as Networks can Lead to Maximum Mutual Information Estimation of Parameters." NIPS 1989

Decoder

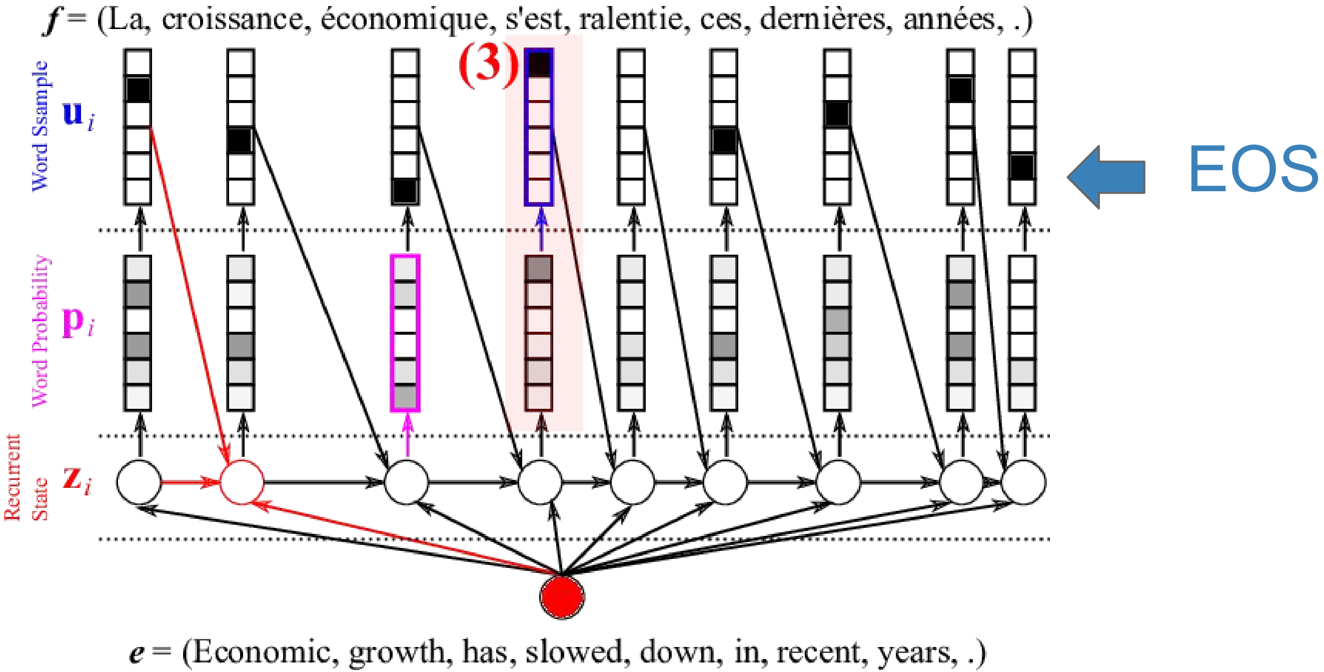
Once an output word sample u_i is predicted at timestep i, the process is iterated…

(1) update the decoder's internal state z_{i+1}
(2) compute scores and probabilities p_{i+1} for all possible target words
(3) predict the word sample u_{i+1}…

Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)

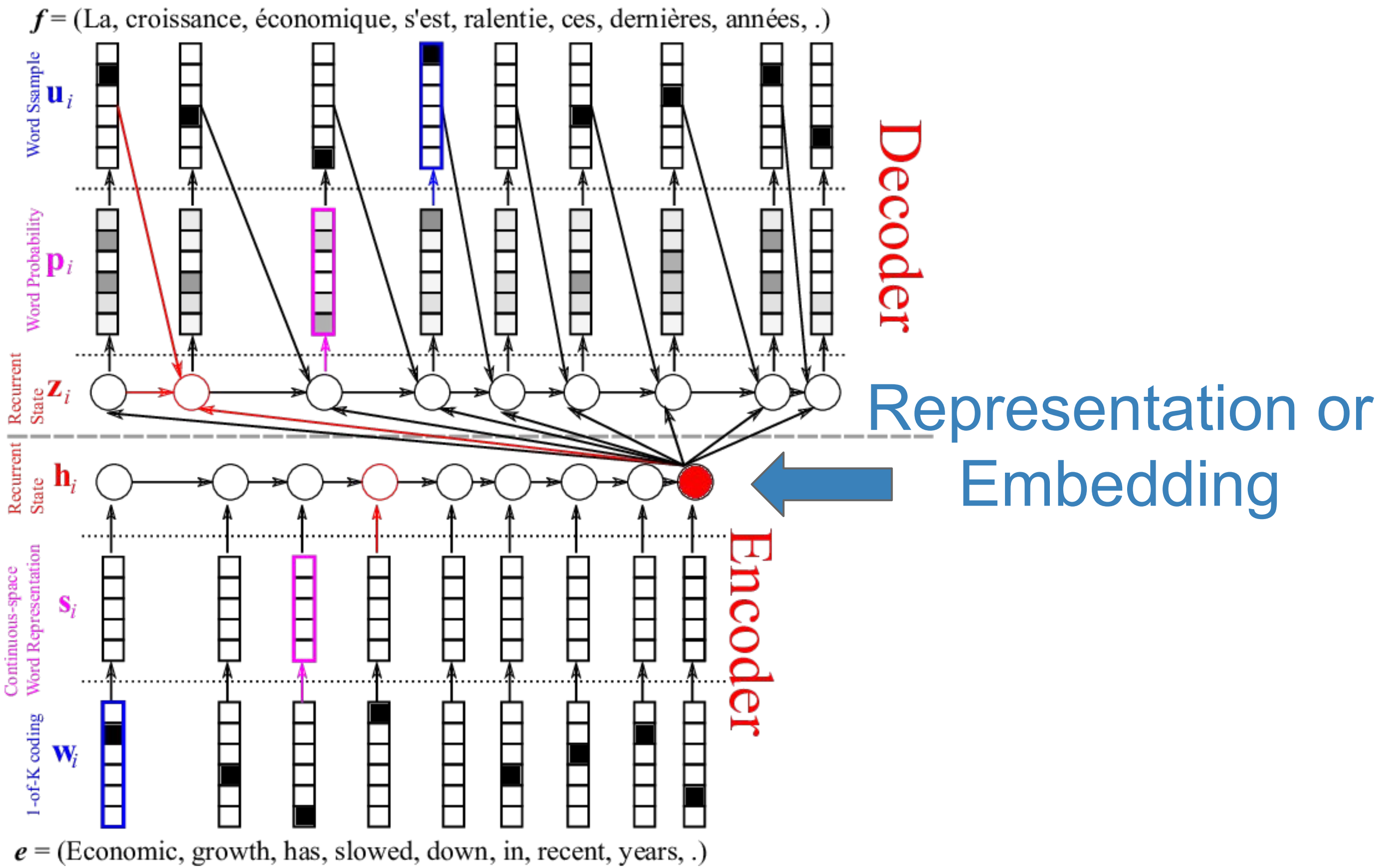
Decoder

More words for the decoded sentence are generated until a <EOS> (End Of Sentence) “word” is predicted.

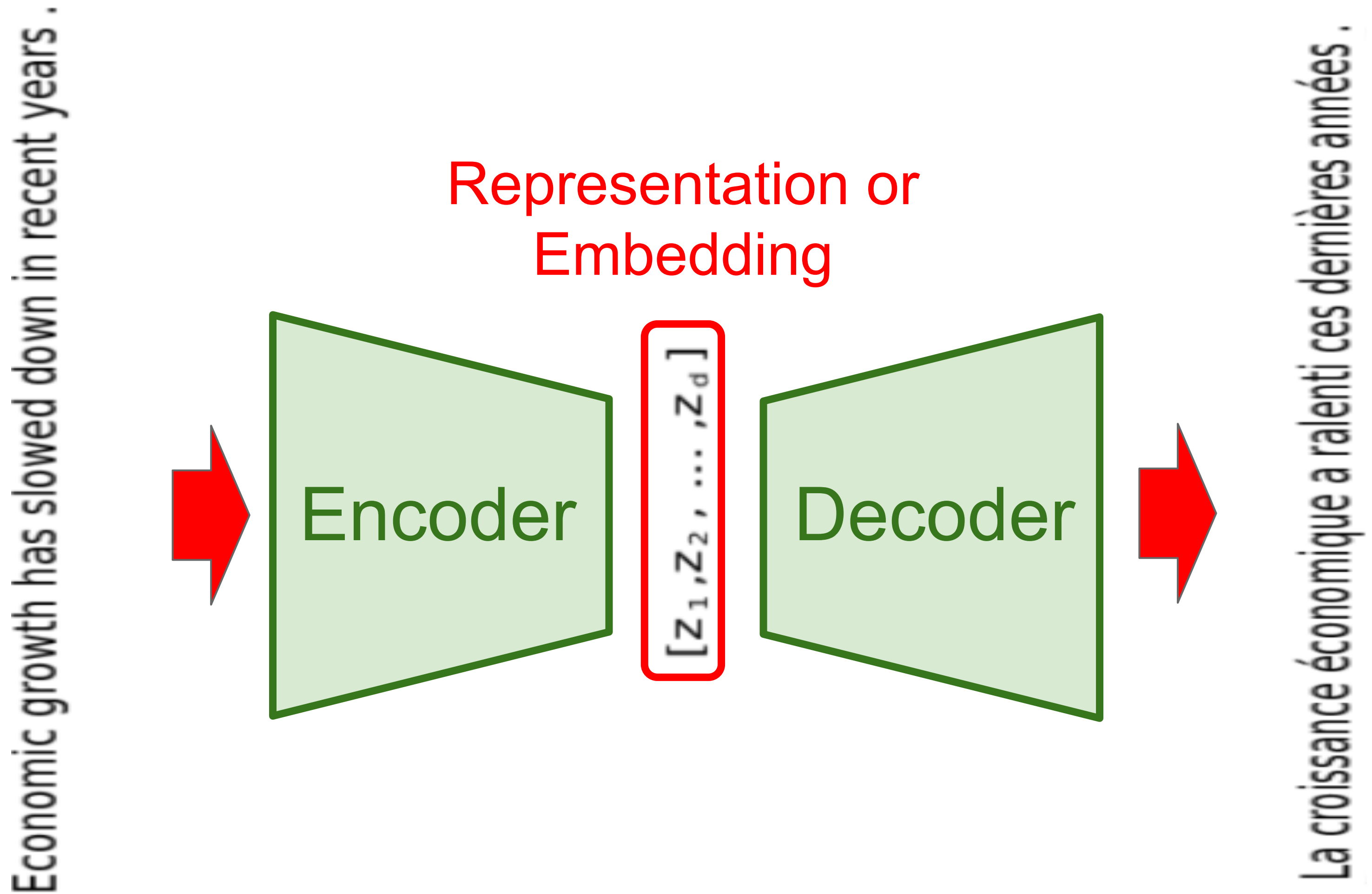


Kyunghyun Cho, [“Introduction to Neural Machine Translation with GPUs”]() (2015)

Encoder-Decoder



Kyunghyun Cho, "Introduction to Neural Machine Translation with GPUs" (2015)


Economic growth has slowed down in recent years .
Representation or Embedding
Encoder
$[z_1, z_2, \dots, z_d]$
Decoder
La croissance économique a ralenti ces dernières années .

Encoder-Decoder: Training

Training requires a large dataset of pairs of sentences in the two languages to translate.

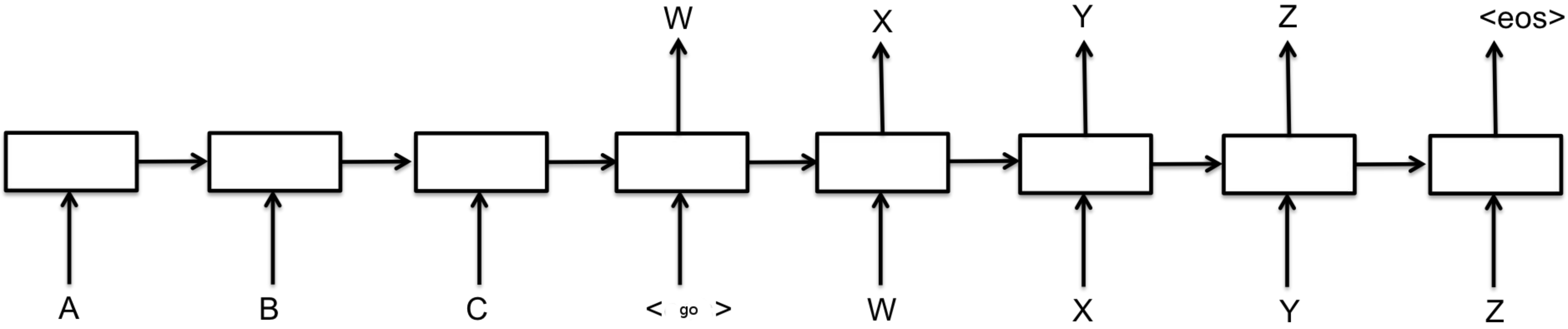
Source	Translation Model
at the end of the	[a la fin de la] [f la fin des années] [être supprimés à la fin de la]
for the first time	[r © pour la premirëre fois] [été donnés pour la première fois] [été commémorée pour la première fois]
in the United States and	[? aux ?tats-Unis et] [été ouvertes aux États-Unis et] [été constatées aux États-Unis et]
, as well as	[?s , qu'] [?s , ainsi que] [?re aussi bien que]
one of the most	[?t ?l' un des plus] [?l' un des plus] [être retenue comme un de ses plus]

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. ["Learning phrase representations using RNN encoder-decoder for statistical machine translation."]() AMNLP 2014.

Encoder-Decoder: Seq2Seq

The **Seq2Seq** variation:

- trigger the output generation with an input <go> symbol.
- the predicted word at timestep *t*, becomes the input at *t+1*.

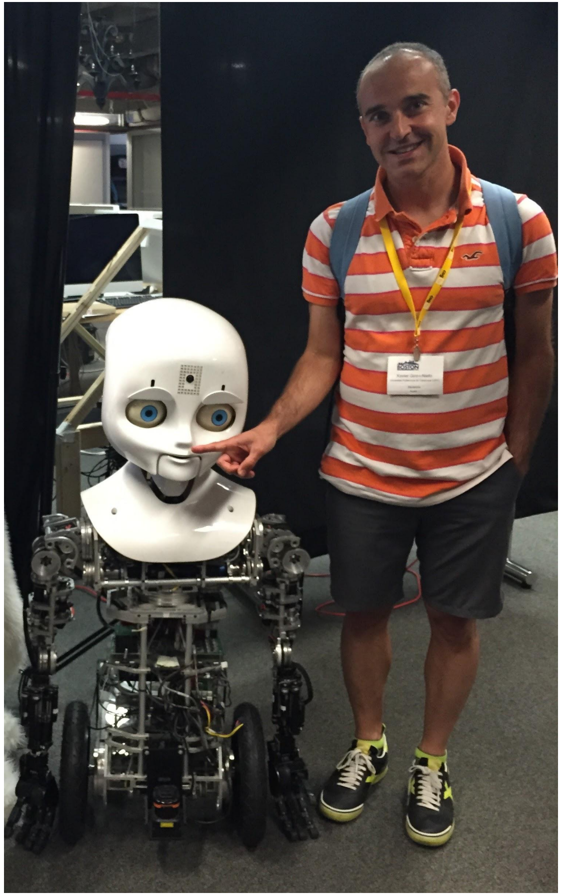


Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. ["Sequence to sequence learning with neural networks."]() NIPS 2014.

Encoder-Decoder: Seq2Seq

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. ["Sequence to sequence learning with neural networks."]() NIPS 2014.

Thanks ! Q&A ?



Follow me at

/ProfessorXavi

@DocXavi

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Department of Signal Theory
and Communications

Image Processing Group

https://imatge.upc.edu/web/people/xavier-giro