(Neural) Machine Translation
Guest lecture at:
NPFL054 Introduction to ML

Ondřej Bojar
bojar@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

January 3, 2018

Outline

- Non-neural statistical machine translation (SMT):
  - Why is MT hard.
  - Quick description.
  - Unjustified independence assumptions.
- Neural Machine Translation (NMT):
  - Neural networks as universal approximators.
  - Processing text with neural networks.
  - Sequence-to-sequence architecture.
  - Sample outputs, sample errors.

MT Must Guess Word Meaning...

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	kočky	.
	pily	dvě	zelená	pruhovaná	koček	
	...	dvou	zelené	pruhované	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	...		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	...		zelenou	pruhovanou		
	viděl jsem		zelenými	pruhovanými		
	viděla jsem		...	...		

. . . and Pick the Right Translation . . .

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	dvou	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	kočkách	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		zelených	pruhovaných		
	uviděl		zelenému	pruhovanému		
	uviděla		zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

. . . But Context Matters

I	saw	two	green	striped	cats	.
já	pila	dva	zelený	pruhovaný	**kočky**	.
	pily	**dvě**	zelená	pruhovaná	koček	
	. . .	**dvou**	**zelené**	**pruhované**	kočkám	
	viděl	dvěma	zelení	pruhovaní	**kočkách**	
	viděla	dvěmi	zeleného	pruhovaného	kočkami	
	. . .		**zelených**	**pruhovaných**		
zrak mi utkvěl na			zelenému	pruhovanému		
			zeleným	pruhovaným		
	. . .		zelenou	pruhovanou		
viděl jsem			zelenými	pruhovanými		
viděla jsem			. . .	. . .		

Non-Neural Statistical MT (1/2)

Given a source (foreign) language sentence
$f_1^J = f_1 \dots f_j \dots f_J$,
Produce a target language (English) sentence
$e_1^I = e_1 \dots e_j \dots e_I$.
Among all possible e_1^I, choose the most likely one:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(e_1^I | f_1^J) \tag{1}$$

Use Bayes' law to divide the model into components:

$$\hat{e}_1^{\hat{I}} = \underset{I, e_1^I}{\operatorname{argmax}}\, p(f_1^J | e_1^I) p(e_1^I) \tag{2}$$

$p(f_1^J \mid e_1^I)$	Translation model (TM, "reversed", $e_1^I \rightarrow f_1^J$)
$p(e_1^I)$	Language model (LM)

Non-Neural Statistical MT (2/2)

				Total	Weight	Weighted
Phrase log. prob.	0,0	-0,69	-1,39	-2,08	2,0	-4,16
Phrase penalty	1,0	1,0	1,0	3,0	-1,0	-3,0
Word penalty	1,0	2,0	1,0	4,0	-0,5	-2,0
	Peter	left for	home .			
	Petr	odešel	domů .			

	▷ Petr	Petr odešel	odešel domů	domů .	. ◁	Total	Weight	Weighted
Bigram log. prob.	-4,02	-2,50	-3,61	-0,39	-0,08	-10,59	1,0	-10,59
							Total	-19,75

- Output composed of selected translation units ("minimum translation units", MTUs: here phrase pairs, solid)
- Scored with LM *n*-grams (dashed) and other features.
- Scores weighted (log-linear, not pure Bayes actually).

1: SMT Illustrated

Nemám žádného psa.	Viděl kočku.
I have no dog.	He saw a cat.

2: Align Words

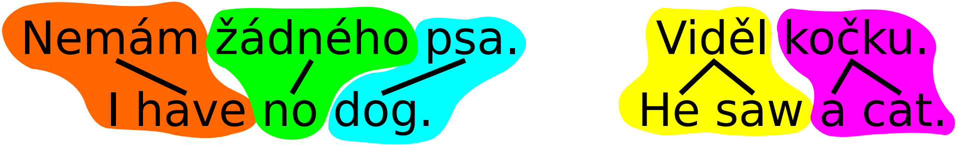
Nemám žádného psa.

I have no dog.

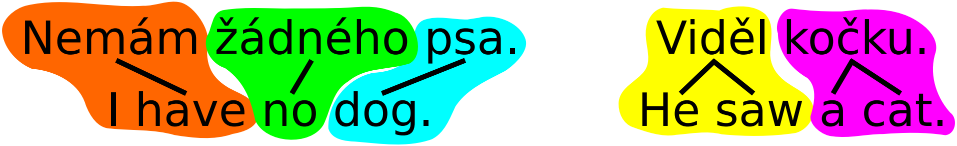
Viděl kočku.

He saw a cat.

3: Extract Phrase Pairs (MTUs)

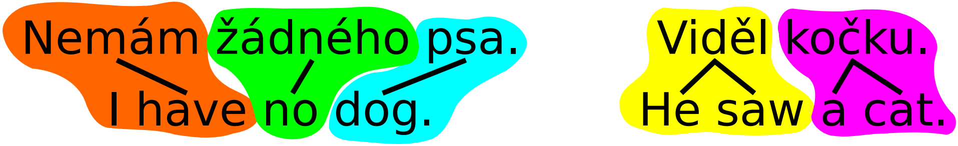


4: New Input



New input: Nemám kočku.

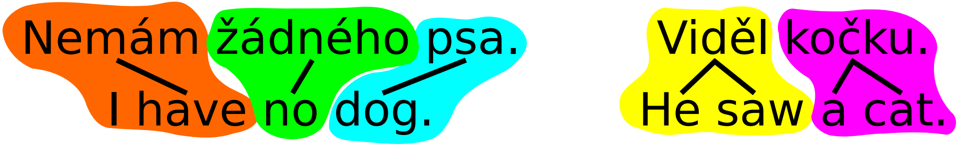
4: New Input



... I don't have cat.

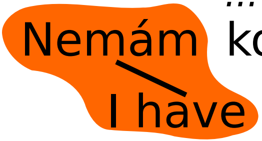
New input: Nemám kočku.

5: Pick Probable Phrase Pairs (TM)



... I don't have cat.

New input:



kočku.

6: So That *n*-Grams Probable (LM)

Meaning Got Reversed!

What Went Wrong?

$$\hat{e}_1^{\hat{I}} = \operatorname*{argmax}_{I,e_1^I} p(f_1^J|e_1^I)p(e_1^I) = \operatorname*{argmax}_{I,e_1^I} \prod_{(\hat{f},\hat{e})\in\text{phrase pairs of } f_1^J,e_1^I} p(\hat{f}|\hat{e})p(e_1^I) \tag{3}$$

- Too strong phrase-independence assumption.
  - Phrases do depend on each other.
    Here "nemám" and "žádného" jointly express one negation.
  - Word alignments ignored that dependence.
    But adding it would increase data sparseness.
- Language model is separate from translation model.
  - $p(e_1^I)$ models the target sentence independently of f_1^J.

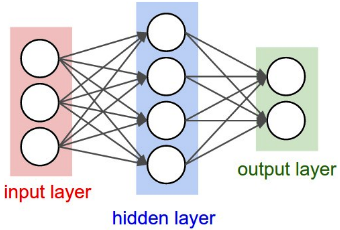
Redefining $p(e_1^I|f_1^J)$

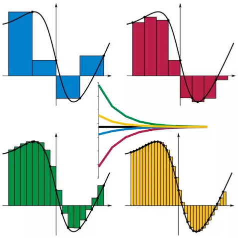
What if we modelled $p(e_1^I|f_1^J)$ directly, word by word:

$$
\begin{aligned}
p(e_1^I|f_1^J) &= p(e_1, e_2, \ldots e_I|f_1^J) \\
&= p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \ldots \\
&= \prod_{i=1}^{I} p(e_i|e_1, \ldots e_{i-1}, f_1^J)
\end{aligned}
\tag{4}
$$

...this is "just a cleverer language model:"
$p(e_1^I) = \prod_{i=1}^{I} p(e_i|e_1, \ldots e_{i-1})$
Main Benefit: All dependencies available.
But what technical device can learn this?

NNs: Universal Approximators

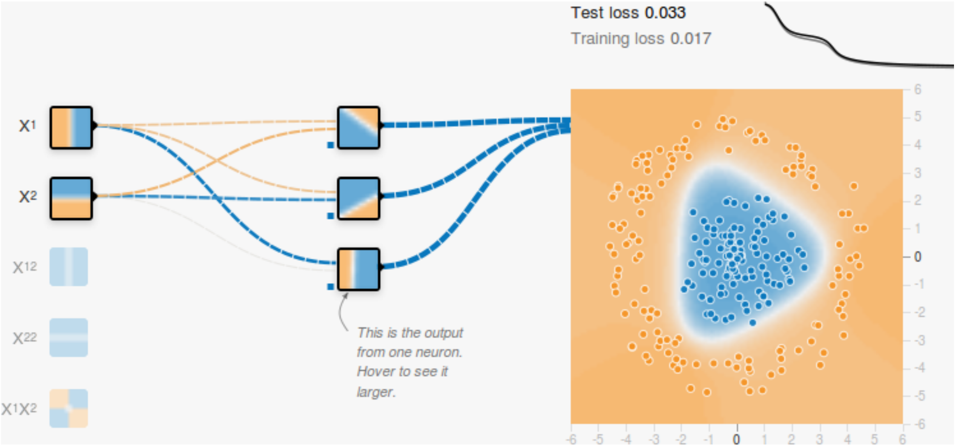




- A neural network with a single hidden layer (possibly huge) can approximate any continuous function to any precision.
  - (Nothing claimed about learnability.)
- There exists a two-layer neural network (ReLU activations) and $2n + d$ weights that can represent any function on a sample of size n in d dimensions. (Zhang et al., 2016)

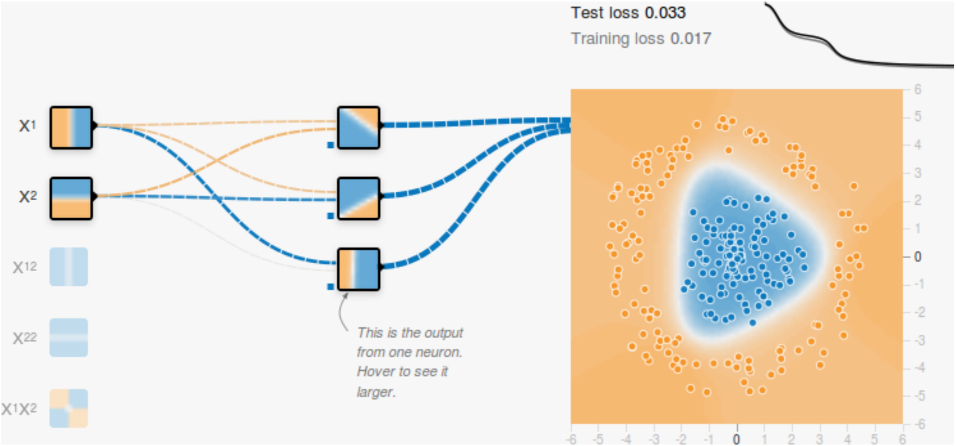
$\Rightarrow$ Big risks of overfitting.

playground.tensorflow.org



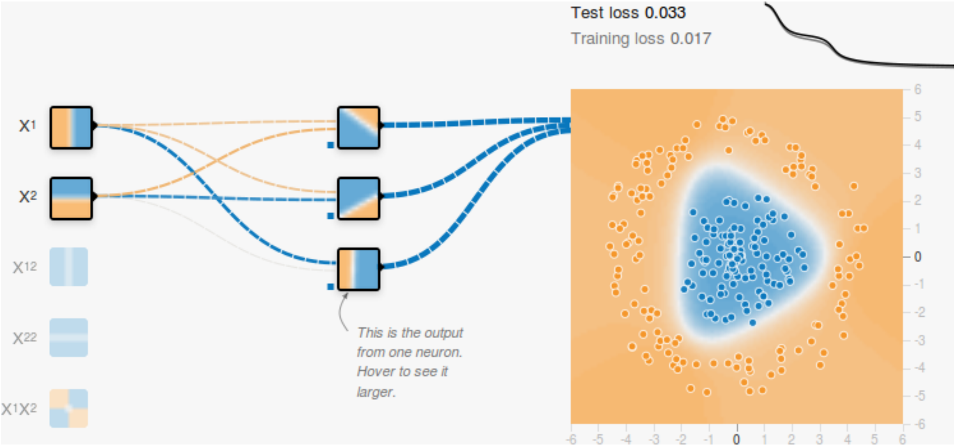
$$-0.43x_1 - 0.89x_2 + 2.0 > 0$$
$$\text{and } -0.67x_1 + 0.89x_2 + 2.1 > 0$$
$$\text{and } 1.4x_1 - 0.067x_2 + 2.3 > 0$$

playground.tensorflow.org



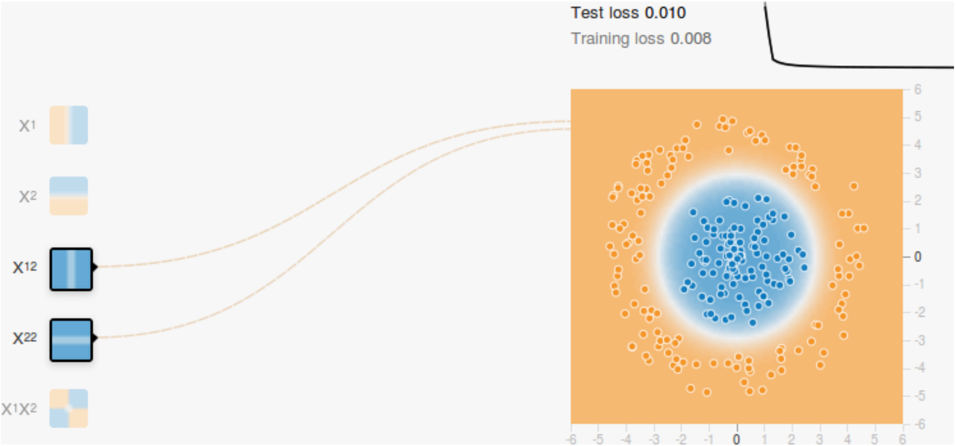
$$\text{In fact: } 1\tanh(-0.43x_1 - 0.89x_2 + 2.0)$$
$$+1\tanh(-0.67x_1 + 0.89x_2 + 2.1)$$
$$+1\tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0$$

playground.tensorflow.org



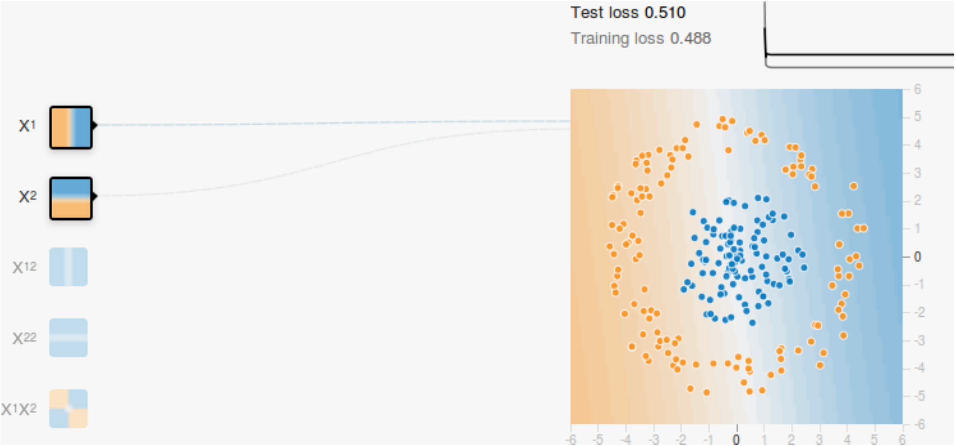
$$
\begin{aligned}
\text{In fact: } & 1\tanh(-0.43x_1 - 0.89x_2 + 2.0) \\
& + 1\tanh(-0.67x_1 + 0.89x_2 + 2.1) \\
& + 1\tanh(1.4x_1 - 0.067x_2 + 2.3) - \pi/2 > 0
\end{aligned}
$$

Perfect Features

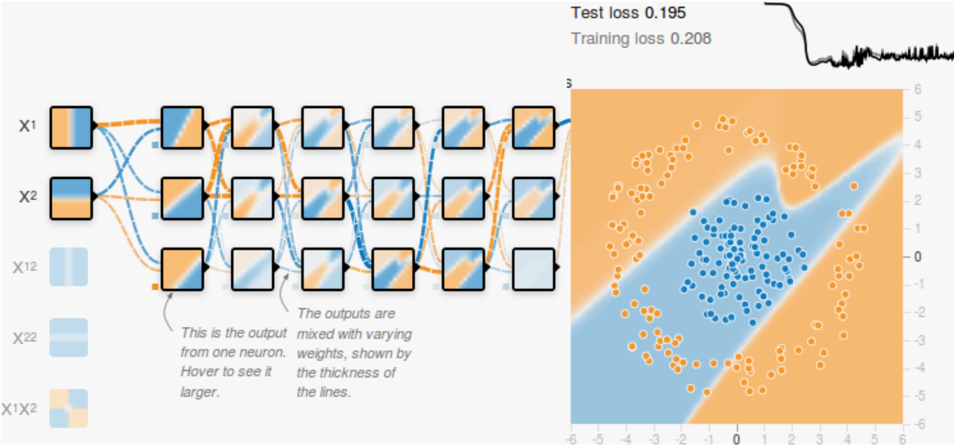


$$1x_1^2 + 1x_2^2 - 1 < 0$$

Bad Features & Low Depth

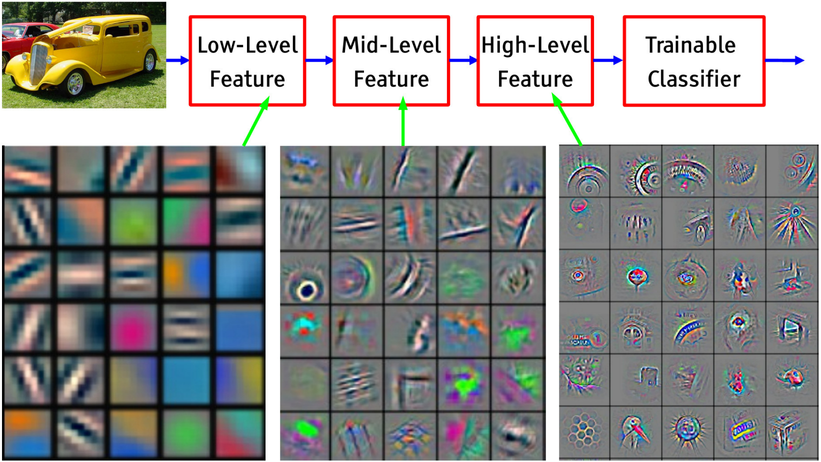


Too Complex NN Fails to Learn



Deep NNs for Image Classification

- It's deep if it has more than one stage of non-linear feature transformation



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

Representation Learning

Representation Learning

- We saw DL finding useful features.
- We can think of these features as *new coordinates*.

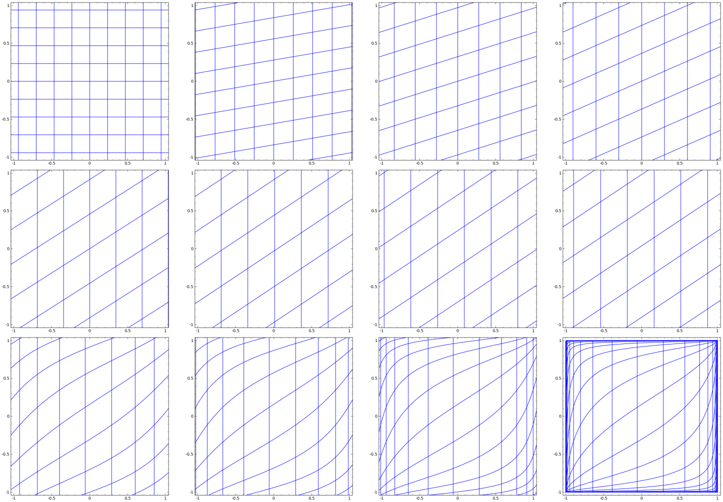
$\Rightarrow$ NNs are learning how to represent the input
to make it linearly separable.

One Layer $\tanh(Wx + b)$, 2D→2D

Skew:
W

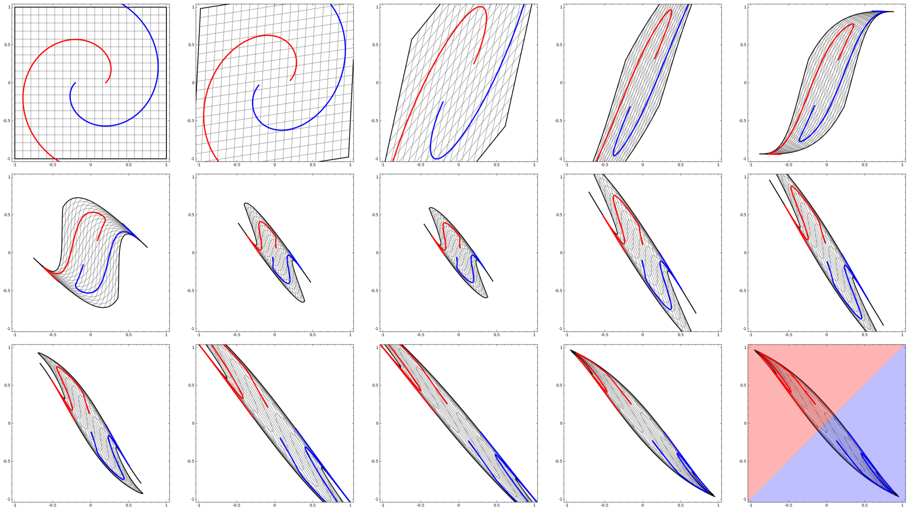
Transpose:
b

Non-lin.:
tanh



Animation by `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Four Layers, Disentagling Spirals



Animation by `http://colah.github.io/posts/2014-03-NN-Manifolds-Topology/`

Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Sentence is then a matrix:

		the	cat	is	on	the	mat
$\uparrow$	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
	...	...	...	...	...	...	...
	is	0	0	**1**	0	0	0
	...	...	...	...	...	...	...
	on	0	0	0	**1**	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
$\downarrow$	zebra	0	0	0	0	0	0

Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

- Sentence is then a matrix:

		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
Vocabulary size:	...	...	...	...	...	...	...
1.3M English	is	0	0	**1**	0	0	0
2.2M Czech	...	...	...	...	...	...	...
	on	0	0	0	**1**	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Processing Text with NNs

- Map each word to a vector of 0s and 1s ("1-hot repr."):

$$\text{cat} \mapsto (0, 0, \ldots, 0, 1, 0, \ldots, 0)$$

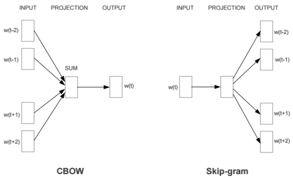
- Sentence is then a matrix:

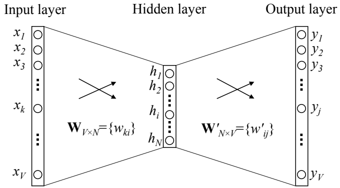
		the	cat	is	on	the	mat
↑	a	0	0	0	0	0	0
	about	0	0	0	0	0	0
	...	...	...	...	...	...	...
	cat	0	**1**	0	0	0	0
Vocabulary size:	...	...	...	...	...	...	...
1.3M English	is	0	0	**1**	0	0	0
2.2M Czech	...	...	...	...	...	...	...
	on	0	0	0	**1**	0	0
	...	...	...	...	...	...	...
	the	**1**	0	0	0	**1**	0
	...	...	...	...	...	...	...
↓	zebra	0	0	0	0	0	0

Main drawback: No relations, all words equally close/far.

Word Embeddings

- Map each word to a dense vector.
- In practice 300–2000 dimensions are used, not 1–2M.
  - The dimensions have no clear interpretation.
- Embeddings are trained for each particular task.
  - NNs: The matrix that maps 1-hot input to the first layer.
- The famous word2vec (Mikolov et al., 2013):
  - CBOW: Predict the word from its four neighbours.
  - Skip-gram: Predict likely neighbours given the word.



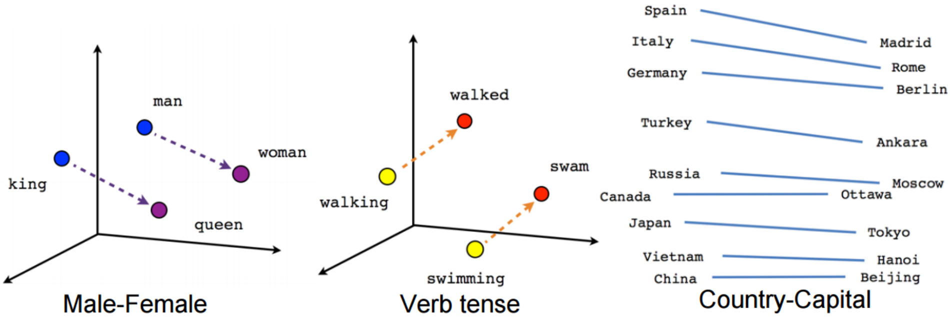


Right: CBOW with just a single-word context
(`http://www-personal.umich.edu/~ronxin/pdf/w2vexp.pdf`)

Continuous Space of Words

Word2vec embeddings show interesting properties:

$$v(\text{king}) - v(\text{man}) + v(\text{woman}) \approx v(\text{queen}) \tag{5}$$

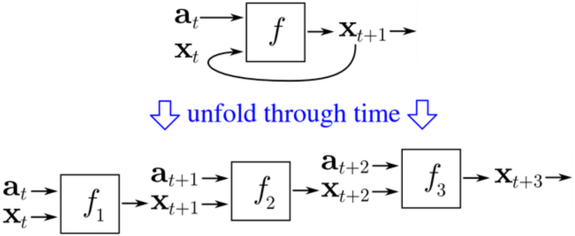


Illustrations from `https://www.tensorflow.org/tutorials/word2vec`

Variable-Length Inputs

Variable-length input can be handled by recurrent NNs:

- Reading one input symbol at a time.
  - The same (trained) transformation used repeatedly.
- Unroll in time (up to a fixed length limit).
- "Back-propagation through time".



$$
\begin{aligned}
x_{t+1} &= f(\text{concat}(a_t, x_t)) \\
&= \tanh(W \text{concat}(a_t, x_t) + b)
\end{aligned}
\tag{6}
$$

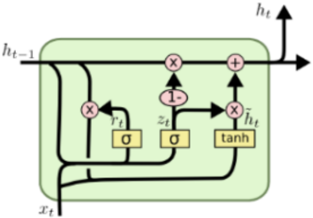
Better RNN Cells

Reusing the same weights easily leads to:

- exploding gradients (solve by clipping)
- vanishing gradients

Gating: only part of input/output used/updated:

- LSTM (Long Short-Term Memory Cells, Hochreiter and Schmidhuber (1997)).
- GRU (Gated Recurrent Unit Cells, Chung et al. (2014)).



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

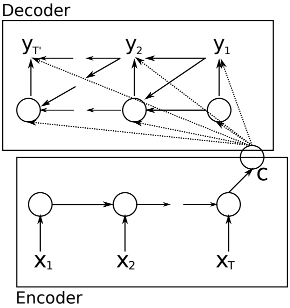
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Important: h_{t-1} and h_t live in the same vector space.

NNs as Translation Model in SMT

Cho et al. (2014) propose:

- encoder-decoder architecture and
- GRU unit (name given later by Chung et al. (2014))
- to score variable-length phrase pairs in PBMT.

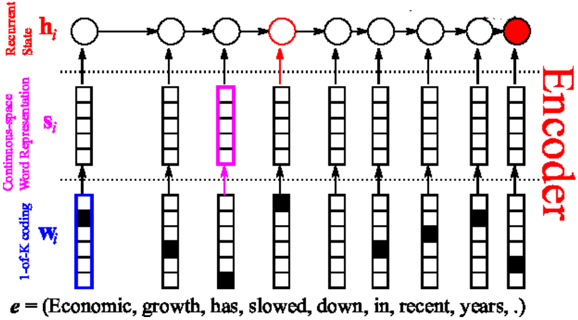


NMT: Sequence to Sequence

Sutskever et al. (2014) use:

- LSTM RNN encoder-decoder
- to consume
  and *produce* variable-length sentences.

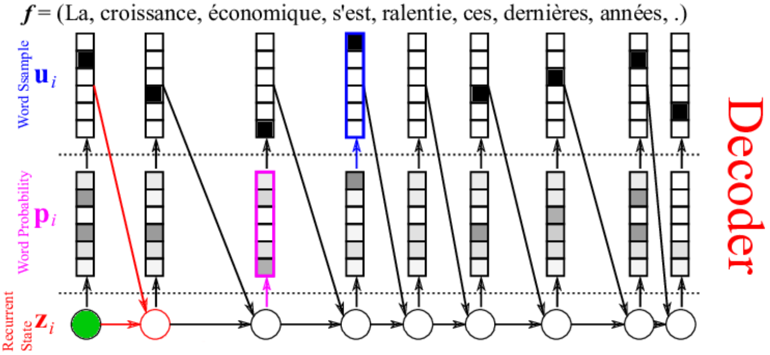
First the Encoder:



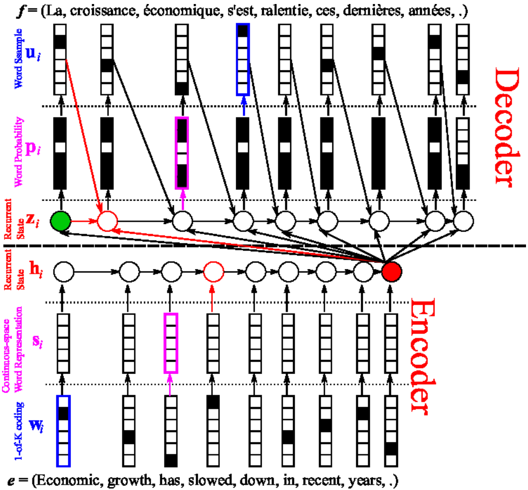
Then the Decoder

Remember: $p(e_1^I|f_1^J) = p(e_1|f_1^J) \cdot p(e_2|e_1, f_1^J) \cdot p(e_3|e_2, e_1, f_1^J) \ldots$

- Again RNN, producing one word at a time.
- The produced word fed back into the network.
  - (Word embeddings in the target language used here.)

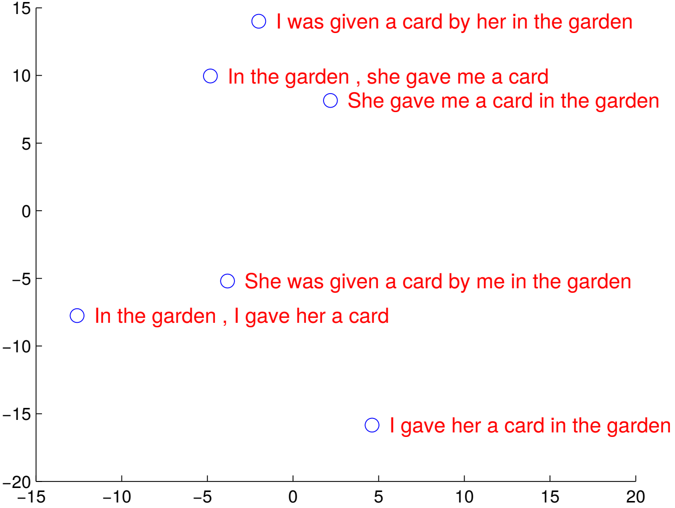


Encoder-Decoder Architecture



https://devblogs.nvidia.com/parallelforall/introduction-neural-machine-translation-gpus-part-2/

Continuous Space of Sentences



2-D PCA projection of 8000-D space representing sentences (Sutskever et al., 2014).

Ultimate Goal of SMT vs. NMT

Goal of "classical" SMT:

Find **minimum translation units** $\sim$ graph partitions:

- such that they are frequent across many sentence pairs.
- without imposing (too hard) constraints on reordering.
- in an unsupervised fashion.

Goal of neural MT:

Avoid minimum translation units.

Come up with a network architecture that:

- Reads input in as original form as possible.
- Produces output in as final form as possible.
- Can be optimized end-to-end *in practice*.

Sample Outputs: NMT or Humans?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
	28letý šéfkuchař Found Dead v San Francisco Mall
	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

Sample Outputs: NMT or Humans?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

Sample Outputs: NMT or Humans?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden $\emptyset$ schodech místního obchodu.

Sample Outputs: NMT or Humans?

SRC	28-Year-Old Chef Found Dead at San Francisco Mall
MT	28letý šéfkuchař Found Dead v San Francisco Mall
REF	Osmadvacetiletý šéfkuchař nalezen mrtev v obchodě v San Francisku

SRC	A 28-year-old chef who had recently moved to San Francisco was found dead in the stairwell of a local mall this week.
MT	Osmadvacetiletý kuchař, který se nedávno přestěhoval do San Francisca, byl tento týden nalezen mrtvý na schodišti místního obchodního centra.
REF	Osmadvacetiletý šéfkuchař, který se nedávno přistěhoval do San Franciska, byl tento týden ∅ schodech místního obchodu.

Sample Outputs: NMT or Humans?

SRC	A spokesperson for Sons & Daughters **said** they were "shocked and devastated" by his death.
	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí šokováni a zničeni.

Sample Outputs: NMT or Humans?

SRC	A spokesperson for Sons & Daughters **said** they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí šokováni a zničeni.

Sample Outputs: NMT or Humans?

SRC	A spokesperson for Sons & Daughters **said** they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí šokováni a zničeni.

SRC	"He found an apartment, he was dating a girl," Louis Galici**a** told KGO.
	Našel si byt, chodil s dívkou, řekl Louis Galici**a** pro KGO.
	"Našel si byt, chodil s holkou," řekl Louis Galici**e** KGO.

Sample Outputs: NMT or Humans?

SRC	A spokesperson for Sons & Daughters **said** they were "shocked and devastated" by his death.
MT	Mluvčí společnosti Sons & Daughters **uvedla**, že jsou jeho smrtí "šokováni a zdrceni".
REF	Mluvčí restaurace Sons & Daughters **řekl**, že jsou jeho smrtí šokováni a zničeni.

SRC	"He found an apartment, he was dating a girl," Louis Galici**a** told KGO.
REF	Našel si byt, chodil s dívkou, řekl Louis Galici**a** pro KGO.
MT	"Našel si byt, chodil s holkou," řekl Louis Galici**e** KGO.

Sample Outputs: NMT or Humans?

SRC The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.

Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.

Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

Sample Outputs: NMT or Humans?

SRC	The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.
REF	Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.
MT	Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

Sample Outputs: NMT or Humans?

SRC	The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.
REF	Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.
MT	Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

SRC	There were creative differences on the set and a disagreement.
	Došlo ke vzniku kreativních rozdílů na scéně a k neshodám.
	Na place byly tvůrčí rozdíly a neshody.

Sample Outputs: NMT or Humans?

SRC	The police arrested two men, who on Tuesday attacked a thirty-five-year-old man with a knife and a machete.
REF	Policie **obvinila** dva útočníky, kteří v úterý v centru Olomouce napadli nožem a mačetou pětatřicetiletého muže.
MT	Policie **zatkla** dva muže, kteří v úterý napadli pětatřicetiletého muže nožem a mačetou.

SRC	There were creative differences on the set and a disagreement.
REF	Došlo ke vzniku kreativních rozdílů na scéně a k neshodám.
MT	Na place byly tvůrčí rozdíly a neshody.

Sample Outputs: NMT or Humans?

SRC **Economy Secretary** Keith Brown visited the site today and was among the first to walk from the land on to the bridge.

Ekonomický tajemník Keith Brown stavbu dnes navštívil a byl mezi prvními, kteří přišli z pevniny na most.

Ministr hospodářství Keith Brown dnes místo navštívil a byl mezi prvními, kteří vyšli ze země na most.

Sample Outputs: NMT or Humans?

SRC	**Economy Secretary** Keith Brown visited the site today and was among the first to walk from the land on to the bridge.
REF	**Ekonomický tajemník** Keith Brown stavbu dnes navštívil a byl mezi prvními, kteří přišli z pevniny na most.
MT	**Ministr hospodářství** Keith Brown dnes místo navštívil a byl mezi prvními, kteří vyšli ze země na most.

Luckily ;-), Catastropic Errors Happen

Same system as above (UEDIN, winner of WMT17):

SRC	... said Frank initially stayed in **hostels**...
MT	... řekl, že Frank původně zůstal v **Budějovicích**...
↳*Gloss*	... said that Frank initically stayed in **Budweis**...
SRC	Most of the **Clintons'** income...
MT	Většinu příjmů **Kliniky**...
↳*Gloss*	Most of the income of the **Clinic**...
SRC	The 63-year-old has now been made a special repres
MT	63letý **mladík** se nyní stal zvláštním zástupcem...
↳*Gloss*	The 63-year-old **youngster** has now become a spec

Catastropic Errors Happen (2/2)

SRC	Criminal Minds star Thomas Gibson sacked after hitting producer
REF	Thomas Gibson, hvězda seriálu Myšlenky zločince, byl propuštěn po té, co uhodil režiséra
MT	**Kriminalisté Minsku** hvězdu Thomase Gibsona **vyhostili** po **zásahu** producenta
↳*Gloss*	**Minsk criminal investigators** have **expelled** the star Thomas Gibson after **striking** the producer

SRC	...add to that its long-standing grudge...
REF	...přidejte k tomu svou dlouholetou nenávist...
MT	...přidejte k tomu svou dlouholetou **záštitu**...
↳*Gloss*	...add to that its long-standing **auspices**... (grudge = zášť $\rightarrow$ záštita = auspices)

Summary

- Machine translation as a sequence-to-sequence ML task.
- Non-neural statististical MT relied on *decomposition* into smaller bits.
- Neural MT is "just" a complex classifier / language model:
  - Given the whole input sentence
    and words produced so far
    predict the next word (or sentence end).
  - Repeat.
- Deep NNs "learn representation".
  - Down-to-earth: new coordinates, new features.
  - Blue-skies: the meanings of words or sentences.

Would you like to learn more?
NPFL087 Statistical Machine Translation (summer term)

References

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *CoRR*, abs/1412.3555.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. pages 3104–3112.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *CoRR*, abs/1611.03530.