How Deep Learning Quietly Revolutionized NLP

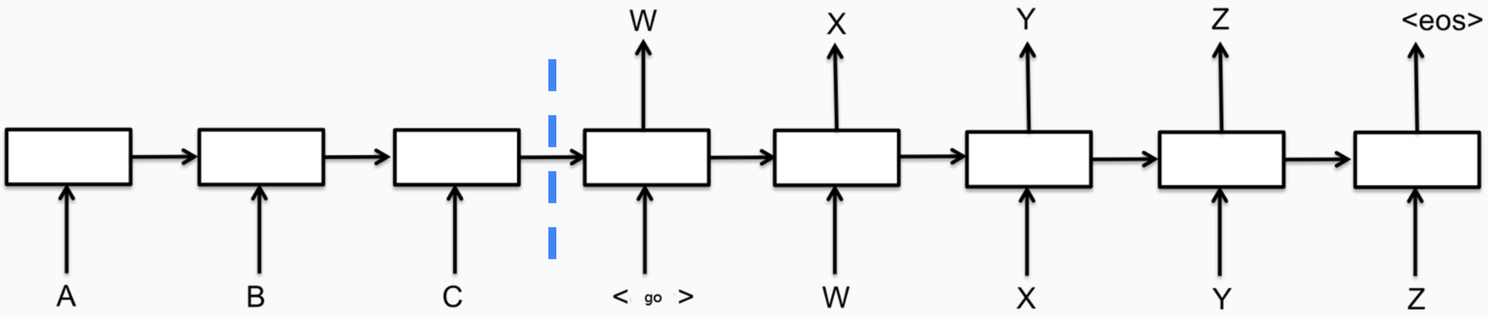
Łukasz Kaiser, Google Brain

What NLP tasks are we talking about?

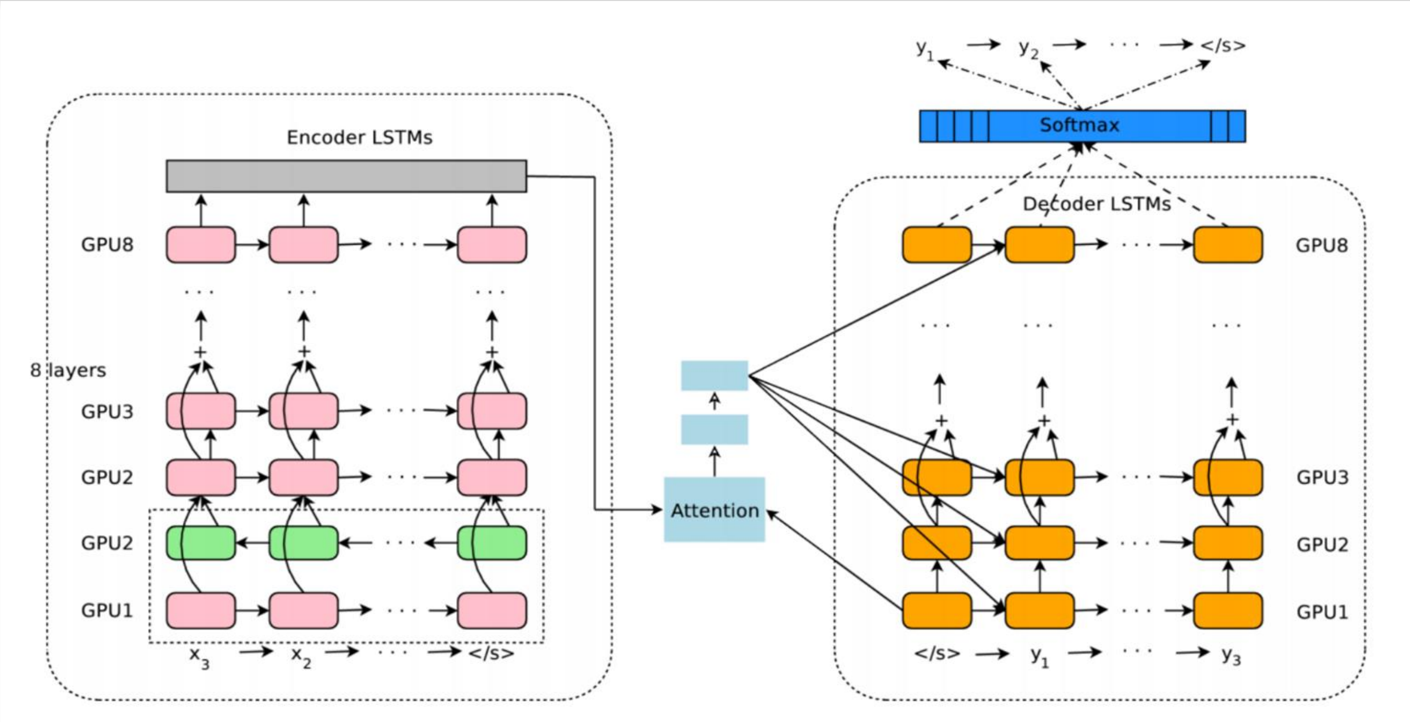
- Part Of Speech Tagging — Assign part-of-speech to each word.
- Parsing — Create a grammar tree given a sentence.
- Named Entity Recognition — Recognize people, places, etc. in a sentence.
- Language Modeling — Generate natural sentences.
- Translation — Translate a sentence into another language.
- Sentence Compression — Remove words to summarize a sentence.
- Abstractive Summarization — Summarize a paragraph in new words.

Can deep learning solve these tasks?

- Inputs and outputs have variable size, how can neural networks handle it?
- Recurrent Neural Networks can do it, but how do we train them?
- Long Short-Term Memory [Hochreiter et al., 1997], but how to compose it?
- Encoder-Decoder (sequence-to-sequence) architectures [Sutskever et al., 2014; Bahdanau et al., 2014; Cho et al., 2014]

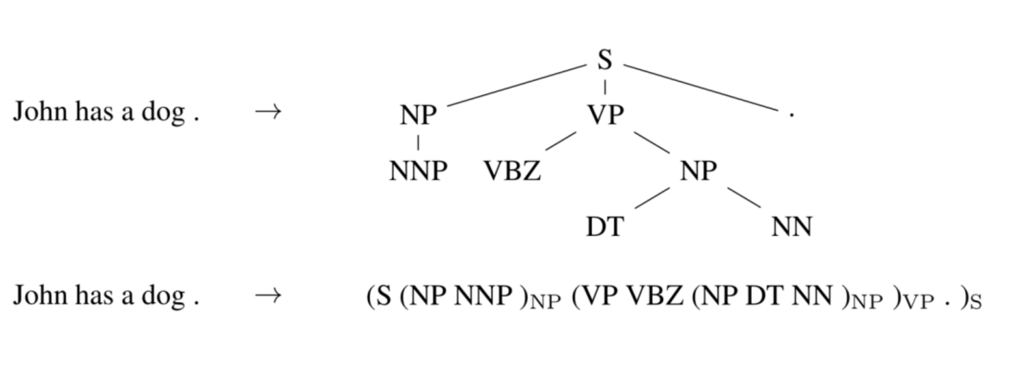


Advanced sequence-to-sequence LSTM



Parsing with sequence-to-sequence LSTMs

(1) Represent the tree as a sequence.



(2) Generate data and train a sequence-to-sequence LSTM model.

(3) Results: 92.8 F1 score vs 92.4 previous best [Vinyals & Kaiser et al., 2014]

Language modeling with LSTMs

Language model performance is measured in perplexity (lower is better).

- Kneser-Ney 5-gram: 67.6 [Chelba et al., 2013]
- RNN-1024 + 9-gram: 51.3 [Chelba et al., 2013]
- LSTM-512-512: 54.1 [Józefowicz et al., 2016]
- 2-layer LSTM-8192-1024: 30.6 [Józefowicz et al., 2016]
- 2-l LSTM-4096-1024+MoE: 28.0 [Shazeer & Mirhoseini et

Language modeling with LSTMs: Examples

Raw (not hand-selected) sampled sentences: [Józefowicz et al., 2016]

About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortege .

It is now known that coffee and cacao products can do no harm on the body .

Yuri Zhirkov was in attendance at the Stamford Bridge at the start of the second half but neither Drogba nor Malouda was able to push on through the Barcelona

Sentence compression with LSTMs

Example:

Input: *State Sen. Stewart Greenleaf discusses his proposed human trafficking bill at Calvery Baptist Church in Willow Grove Thursday night.*

Output: Stewart Greenleaf discusses his human trafficking bill.

Results: readability

informativeness

MIRA (previous best): 4.31 3.55

LSTM [Filippova et al., 2015]: 4.51 3.78

Translation with LSTMs

Translation performance is measured in BLEU scores (higher is better, EnDe):

- Phrase-Based MT: 20.7 [Durrani et al., 2014]
- Early LSTM model: 19.4 [Sébastien et al., 2015]
- DeepAtt (large LSTM): 20.6 [Zhou et al., 2016]
- GNMT (large LSTM): 24.9 [Wu et al., 2016]
- GNMT+MoE: 26.0 [Shazeer & Mirhoseini et al., 2016]

Again, model size and tuning seem to be the decisive factor.

Translation with LSTMs: Examples

German:

Probleme kann man niemals mit derselben Denkweise lösen, durch die sie entstanden sind.

PBMT Translate:

No problem can be solved from solved
the same consciousness that thinking

GNMT Translate:

Problems can never be
with the same way of

Translation with LSTMs: How good is it?

Google Translate production data, median score by human evaluation on the scale 0-6. [Wu et al., ‘16]

	PBMT	GNMT	Human	Relative improvement
English → Spanish	4.885	5.428	5.504	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

What are the problems with LSTMs?

Speed is a problem for sequence-to-sequence LSTM models.

- These models are large, we need more computing power.

  Solution: new hardware trading less precision for more computing.

- These model are sequential, can be slow even at small sizes.

  Solution: new, more parallel models (Neural GPU, ByteNet).

Sequence-to-sequence LSTMs require a lot of data.

- Attention and other new architectures increase data efficiency.
- Use regularizers like dropout, confidence penalty and layer normalization.

Thank You

- Deep learning has profoundly changed the field of NLP.
- Sequence-to-sequence LSTMs yield state-of-the-art results on many NLP tasks.
- Google Translate uses LSTMs in production. Big improvements in translation quality.
- New models are addressing problems with LSTMs.