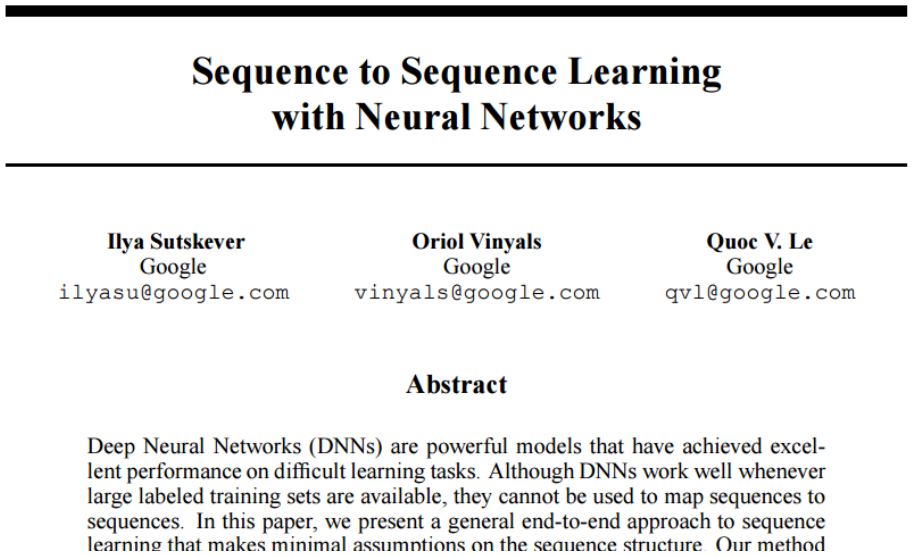
Machine Translation with RNN



L] 14 Dec 2014

Sequence to Sequence Learning with Neural Networks

Ilya Sutskever
Google
ilyasu@google.com

Oriol Vinyals
Google
vinyals@google.com

Quoc V. Le
Google
qvl@google.com

Abstract

Deep Neural Networks (DNNs) are powerful models that have achieved excellent performance on difficult learning tasks. Although DNNs work well whenever large labeled training sets are available, they cannot be used to map sequences to sequences. In this paper, we present a general end-to-end approach to sequence learning that makes minimal assumptions on the sequence structure. Our method

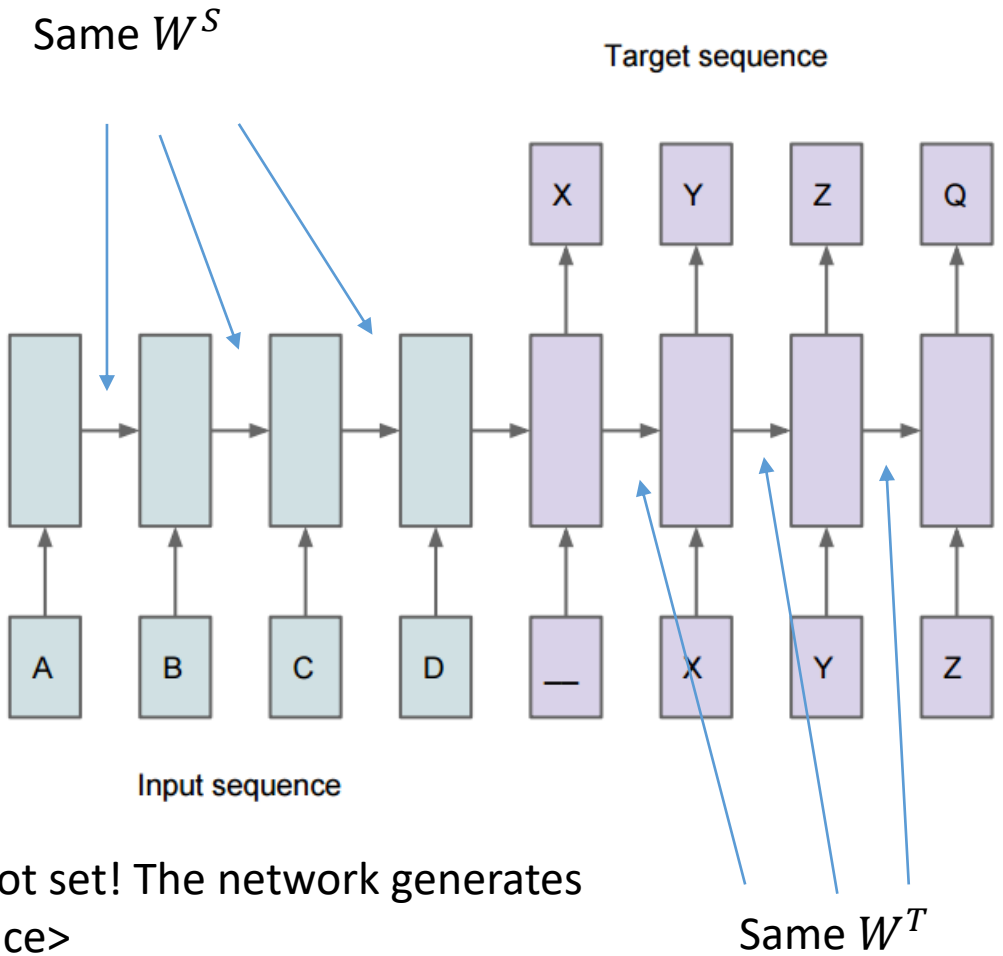
Proc. NIPS 2014

Slides from Ilya Sutskever

Main Idea

Want to translate “ABCD<end sentence>” to “XYZQ<end sentence>”

- One encoder RNN
  - All the information from the sequence ABCD is stored in the hidden vector at the end of the input sequence
- One decoder RNN
  - Generates the output



Note: the sequence lengths are not set! The network generates words until it gets to <end sentence>

LSTM

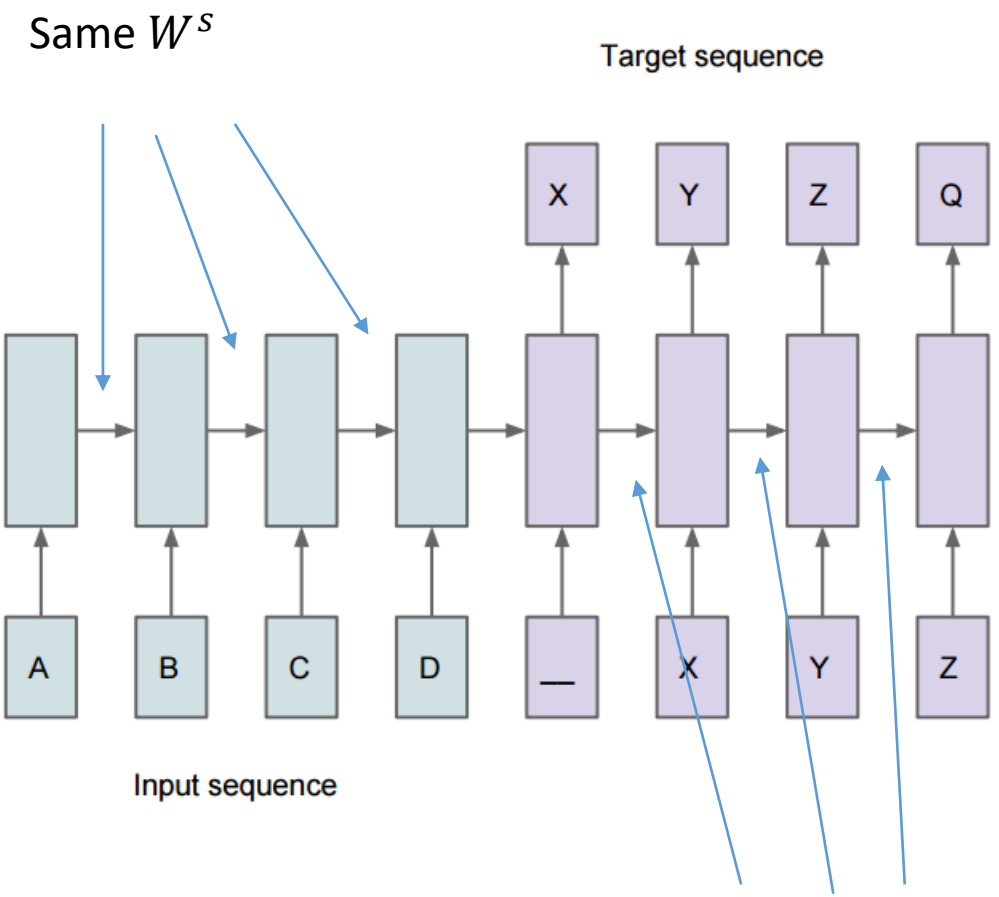
- A modification of RNN hidden units that helps with the vanishing/exploding gradient problem
- Just think of LSTM units as ordinary RNN units
  - Or take a course on Neural Networks!

Cost Function

- Source sentence $S^{(i)}$, target translation $T^{(i)}$
- What to maximize $\sum_i \log P(T^{(i)}|S^{(i)})$

- Compute the hidden state h_t for every time-step in the input sequence using the RNN
- Use it as the input hidden state to the target RNN
- Compute the $\hat{y}_t$ for every time-step in the target sequence
- Multiply the probs of the actual outputs in the target sequence

$$P(T^{(i)}|S^{(i)})$$



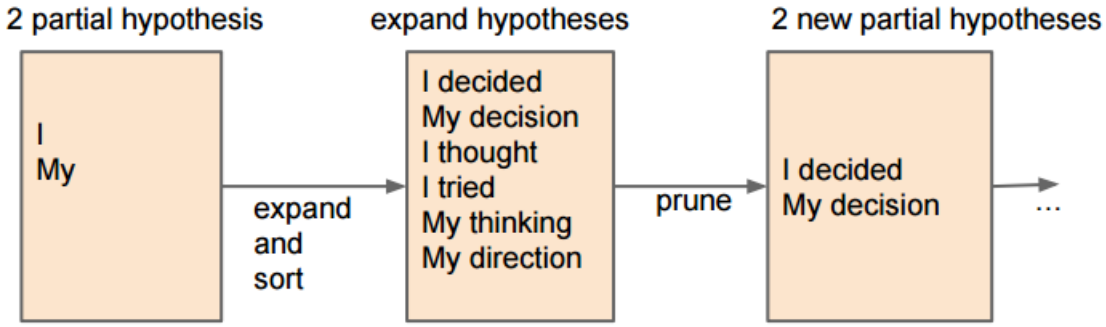
$$P(T|S) = P(\tau_1, \dots, \tau_l|s_1, \dots s_m) = \Pi_t p(\tau_t|h_t, s_1, \dots, s_m, \tau_1, \dots \tau_{t-1})$$

Decoding/Translating

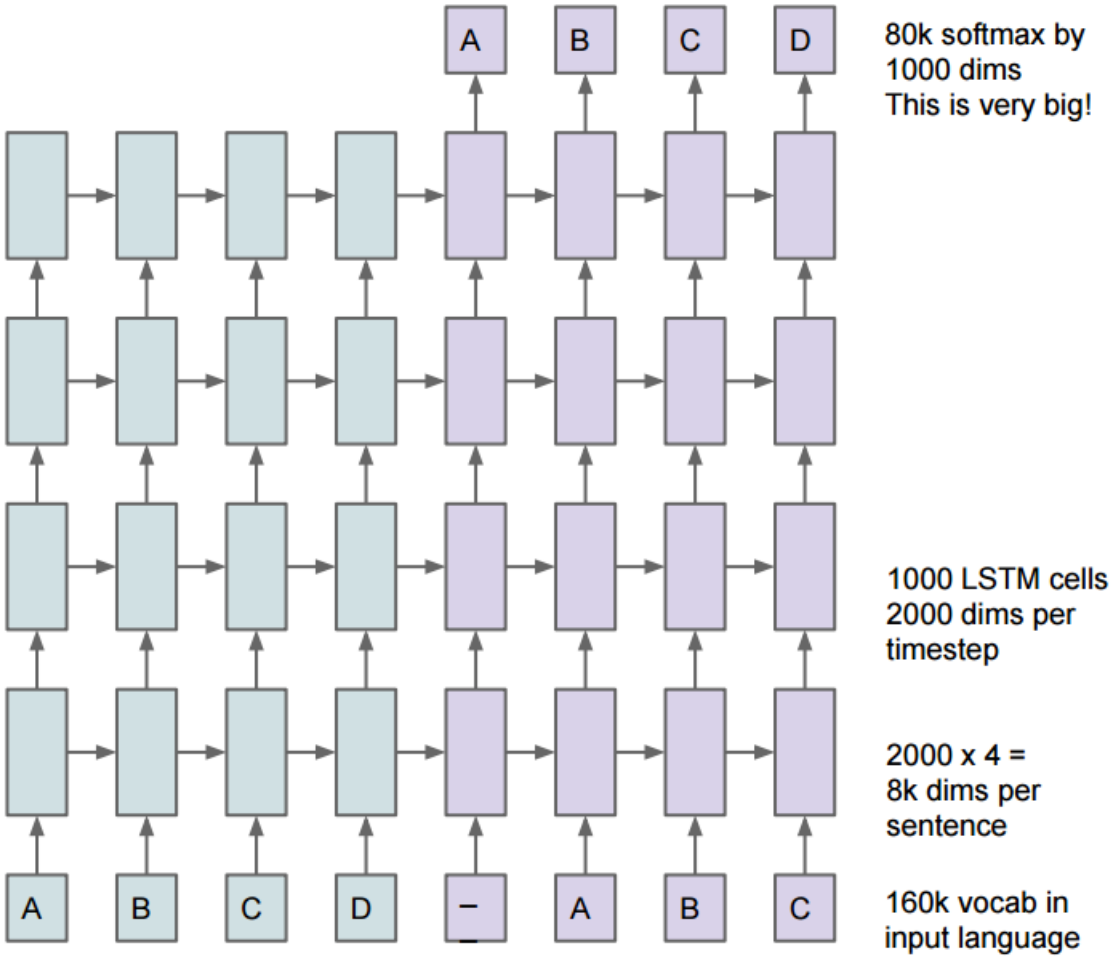
- Technically, we want $argmax_T P(T|S)$
  - Would need to try all possible sentences T!
- Instead, use greedy beam search

Beam-Search Decoding

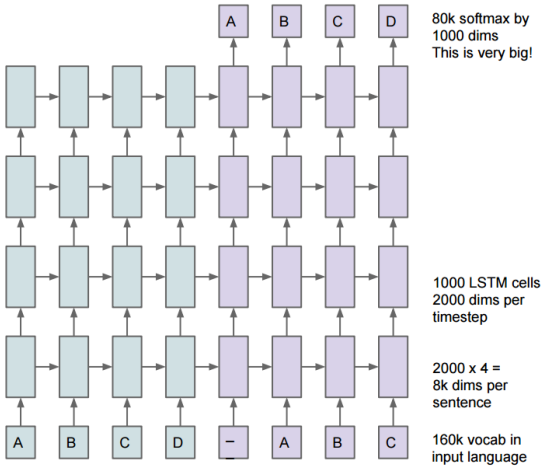
- Generate the translation hypotheses left-to-right
- Maintain N partial translation hypotheses
- Expand each translation with all the possible next words
- Discard all but the top N (in terms of probability) new partial translations



The Model


A
B
C
D
80k softmax by
1000 dims
This is very big!
1000 LSTM cells
2000 dims per
timestep
2000 x 4 =
8k dims per
sentence
A
B
C
D
–
A
B
C
160k vocab in
input language

The Model



- Every rectangle represents 1000 LSTM cells
  - Same colour => same incoming weights
- Every input word is encoded into a 1000-dimensional vector first
  - Like word2vec (alternative: a 160k-dimensional one-hot encoded vector)
  - That way, we can have a deep RNN (since input is smaller, need fewer weights)
- Each sentence is encoded using the memory/hidden state of the right-most green LSTM stack
  - NOTE! A SENTENCE IS ENCODED INTO A FIXED-DIMENSIONAL VECTOR!

Examples

- **FR:** Les avionneurs se querellent au sujet de la largeur des sièges alors que de grosses commandes sont en jeu
- **LSTM:** Aircraft manufacturers are concerned about the width of seats while large orders are at stake
- **Translation:** Jet makers feud over seat width with big orders at stake

Out of vocabulary word (777x)

- **FR**: Aujourd' hui , Airbus en appelle directement au public avant le salon aéronautique de Dubaï , où le 777X devrait prendre le pas sur ses concurrents avec plus de 100 commandes.
- **LSTM**: Today , Airbus is calling directly to the public before the Dubai Airshow , where it is expected to take over its competitors with more than 100 orders
- **Translation:** Now , Airbus is appealing directly to the public ahead of the Dubai Airshow , where the 777X is expected to dominate with more than 100 orders .

Special token for “out of vocabulary word” is one of the inputs/outputs

Example (nonsense)

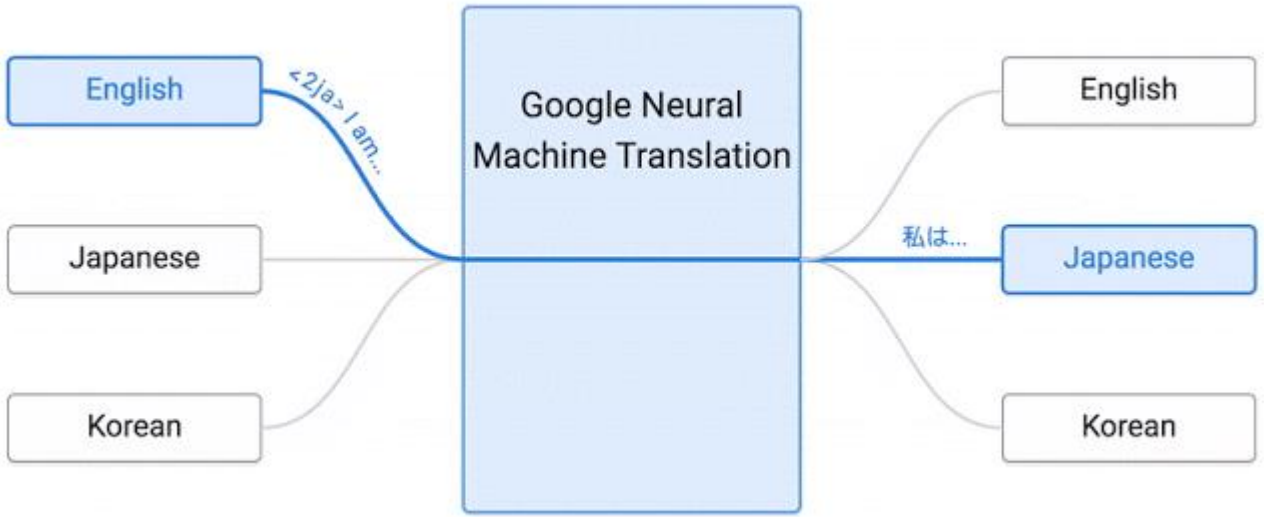
- **FR:** Toutefois , le comité a recommandé que la FAA autorise les pilotes à demander aux passagers d' éteindre leurs appareils pendant les atterrissages aux instruments dans des conditions de faible visibilité .
- **LSTM:** However , the Committee recommended that the FAA allow pilots to ask passengers to stop their aircraft from landing in instruments under low visibility conditions
- **Translation**: However , the committee recommended the FAA allow pilots to order passengers to shut off devices during instrument landings in low visibility .

Example

- **FR:** Les compagnies aériennes permettent à leurs passagers d' utiliser le Wi-Fi aux altitudes de croisière depuis plusieurs années .
- **LSTM:** The airlines allow their passengers to use the Wi-Fi at cruising altitudes for several years .
- **Translation:** Airlines have been offering Wi-Fi use at cruising altitudes to passengers for several years .

Language-Independent Vectors that Represent Sentences?

- https://research.googleblog.com/2016/11/zero-shot-translation-with-googles.html



But is it really learning an "interlingua"?

- "Interlingua": a language comprehensible to speakers of multiple languages
- Here: a vector that represents language-independent meaning

Visualize the activations when plugging the same sentence in different languages