Neural Machine Translation
by Jointly Learning
to Align and Translate

Dzmitry Bahdanau*

KyungHyun Cho, Yoshua Bengio

*work was done during an internship at Université de Montréal
This Talk Is About...

... a neural network that translates...

Like in this simple example!

Comme dans cet exemple simple!
Neural Machine Translation

Different from

(Schwenk et al. 2006)  (Devlin et al. 2014)
Encoder-Decoder Approach

(Ñeco&Forcada, 1997)
(Kalchbrenner et al., 2013)
(Cho et al., 2014)
(Sutskever et al., 2014)

RNN Encoder-Decoder (Cho et al. 2014):

Encoder

Decoder

Representation

$h_1$ $h_2$ ... $h_L$

$x_1$ $x_2$ ... $x_L$

$y_1$ $y_2$ ... $y_T$
RNN Encoder-Decoder: Issues

- has to remember the whole sentence
- fixed size representation can be the bottleneck
- humans do it differently

performance drops on long sentences:
Deviations in the end of long sentences:

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

Un privilège d’admission est le droit d’un médecin de reconnaître un patient à l’hôpital ou un centre médical d’un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

[based on his state of health???]
Key Idea

Tell Decoder what is now translated:

The agreement on European Economic Area was signed in August 1992.

L'accord sur l'Espace économique européen a été signé en ???

Have such hints computed by the net itself!
New Encoder

Bidirectional RNN: $h_j$ contains $x_j$ together with its context ($..., x_{j-1}, x_{j+1}, ...$).

$(h_1, ..., h_L)$ is the new *variable-length* representation instead of *fixed-length* $c$. 
Step i:

- compute alignment
- compute context
- generate new output
- compute new decoder state
Alignment Model

\[ e_{ij} = v^T \tanh(Ws_{i-1} + Vh_j) \]  
\[ \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{L} \exp(e_{ik})} \]  

- nonlinearity (tanh) is crucial!
- simplest model possible
- \( Vh_j \) is precomputed => quadratic complexity with low constant

![Diagram of Alignment Model](Image)
Experiment: English to French

Model:
- RNN Search, 1000 units

Baseline:
- RNN Encoder-Decoder, 1000 units
- Moses, a SMT system (Koehn et al. 2007)

Data:
- English to French translation, 348 million words,
- 30000 words + UNK token for the networks, all words for Moses

Training:
- Minimize mean $\log P(y|x,\theta)$ w.r. $\theta$
- $\log P(y|x,\theta)$ is differentiable w.r. $\theta$ => usual methods
Quantitative Results

- No performance drop on long sentences
- Much better than RNN Encoder-Decoder

<table>
<thead>
<tr>
<th>Model</th>
<th>All</th>
<th>No UNK°</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNNencdec-30</td>
<td>13.93</td>
<td>24.19</td>
</tr>
<tr>
<td>RNNsearch-30</td>
<td>21.50</td>
<td>31.44</td>
</tr>
<tr>
<td>RNNencdec-50</td>
<td>17.82</td>
<td>26.71</td>
</tr>
<tr>
<td>RNNsearch-50</td>
<td>26.75</td>
<td>34.16</td>
</tr>
<tr>
<td>RNNsearch-50*</td>
<td>28.45</td>
<td>36.15</td>
</tr>
<tr>
<td>Moses</td>
<td>33.30</td>
<td>35.63</td>
</tr>
</tbody>
</table>

- Without unknown words
- Comparable with the SMT system
An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

Un privilège d’admission est le droit d’un médecin d’admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l’hôpital.  

Encoder-Decoder  

.... d’un diagnostic ou de prendre un diagnostic en fonction de son état de santé.  

[based on his state of health???]
Qualitative Results: Alignment

The agreement on the European Economic Area was signed in August 1992.


It is known, that the verb often occupies the last position in German sentences.

Es ist bekannt, dass das Verb oft die letzte Position in deutschen Strafen einnimmt.
Related Work: Neural MT

- **Sutskever et al. (2014)**
  - 30.6 BLEU with 4-layer LSTM Encoder-Decoder, 90k words
- **Jean et al. (2015)**
  - 32.8 BLEU, RNNSearch, 500k words by importance sampling
- **Better results by using dictionaries and ensembles**
  - Jean et al. (2015), Luong et al. (2015), both achieve state-of-the-art
Related Work: Attention Mechanisms

Our alignment model is an *attention mechanism*.

- First differentiable attention model for handwriting synthesis: (Graves et al. 2013)
  - monotonic alignment only
  - predicts shifts instead of selecting location

- Non-differentiable attention mechanism for image classification: (Mnih et al. 2014)
Summary

● Novel approach to neural machine translation
  ○ No fixed size representation
  ○ Plausible alignments

● Applicable to many other structured input/output problems
  ○ response generation (not exactly, but Shang et. al 2015)
  ○ speech recognition (Chorowski et. al 2014)
  ○ caption generation (Xu et. al, 2015)
  ○ video description generation (Yao et. al, 2015)

Thanks!