Neural Machine Translation: Approaches, Challenges and By-Products

Kyunghyun Cho¹  Caglar Gulcehre¹  Bart van Merriënboer¹  
Jean Pouget-Abadie²  Dzmitry Bahdanau³  Coline Devin⁴  
Yoshua Bengio¹

¹Université de Montréal, Canada  
²Ecole Polytechnique, France  
³Jacobs University, Germany  
⁴Harvey Mudd College, USA
Neural Machine Translation

La croissance économique a ralenti ces dernières années.

\[ [z_1, z_2, \ldots, z_d] \]

Économic growth has slowed down in recent years.

### Properties
- Variable-length input
- Variable-length output
- End-to-end trainable
- Many-to-many mapping

### Implications
- Order-sensitive
- Probabilistic

---

Kalchbrenner & Blunsom (2013), Kalchbrenner et al. (2014), Sutskever et al. (2014), Cho et al. (2014), Grefenstette et al. (2014)
(1) Recurrent Neural Network

Architecture
- Encoder: Recurrent Network
- Decoder: Recurrent Network

Important note:
- Gating mechanism!

---

Sutskever et al. (2014), Cho et al. (2014)
(2) Convolutional neural networks

- **Encoder**: Convolutional Network
- **Decoder**: (deconv +) Recurrent Neural Network

---

Kalchbrenner & Blunsom (2013), Kalchbrenner et al. (2014)
(3) Gated Recursive Convolutional Neural Network

Architecture

- **Encoder**: *Gated* Recursive Convolutional Network
- **Decoder**: Recurrent Neural Network

(Soft) Gating learns syntax

---

Cho et al. (2014a)
Challenges – (1) Curse of Sentence Length (a)

Issues:

1. Only few long sentences in the training corpus
2. Difficult to encode a long sentence into a fixed-length vector
3. Difficult to generate a coherent long sentence

Cho et al. (2014a,b)
Challenges – (1) Curse of Sentence Length (b)

Segmentation-based Approach

1. Find a segmentation \((s_1, s_2, \ldots)\) that maximizes

\[
\sum_n \max_i (\log p(f_i | s_n) + \log p(s_n | f_i)) / C_{s_n}
\]

2. Translate each \(s_n\) separately into \(f_n\)

3. Concatenate \((f_1, f_2, \ldots)\)

---

**Without Segmentation**

**With Segmentation**

---

Cho et al. (2014b)
Challenges – (2) Unknown words

Issues:

- Word-level modeling: The vocabulary needs to be fixed
- Character-level modeling: A sentence becomes a super long sequence
- Something in-between?

---

Sutskever et al. (2014), Cho et al. (2014a)
Cho et al. (2014a), Edges are thresholded at 0.1.
By-Products – (2) Better Embedding?

Ongoing work (+ Felix Hill* and Tae-ho Kim°)
By-Products – (3) Translation across Modalities

Encoder–Decoder Architecture

- Not restricted to text
- Image → Description
- Speech → Transcription
- Description → Image?
- Speech → Translated Sentence?
- Speech → Foreign Speech?