Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Author: Melvin Johnson , Mike Schuster , Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean

Presented by: Kejia Jiang

Introduction

- A **single** Neural Machine Translation (NMT) model to translate between multiple languages.
- **Simplicity**

  Requires no change to the traditional NMT model architecture.
- **Low-resource language improvements**

  Language pairs with little available data and language pairs with abundant data are mixed together.
- **Zero-shot translation**

  Translates between arbitrary languages, including unseen language pairs during the training process.

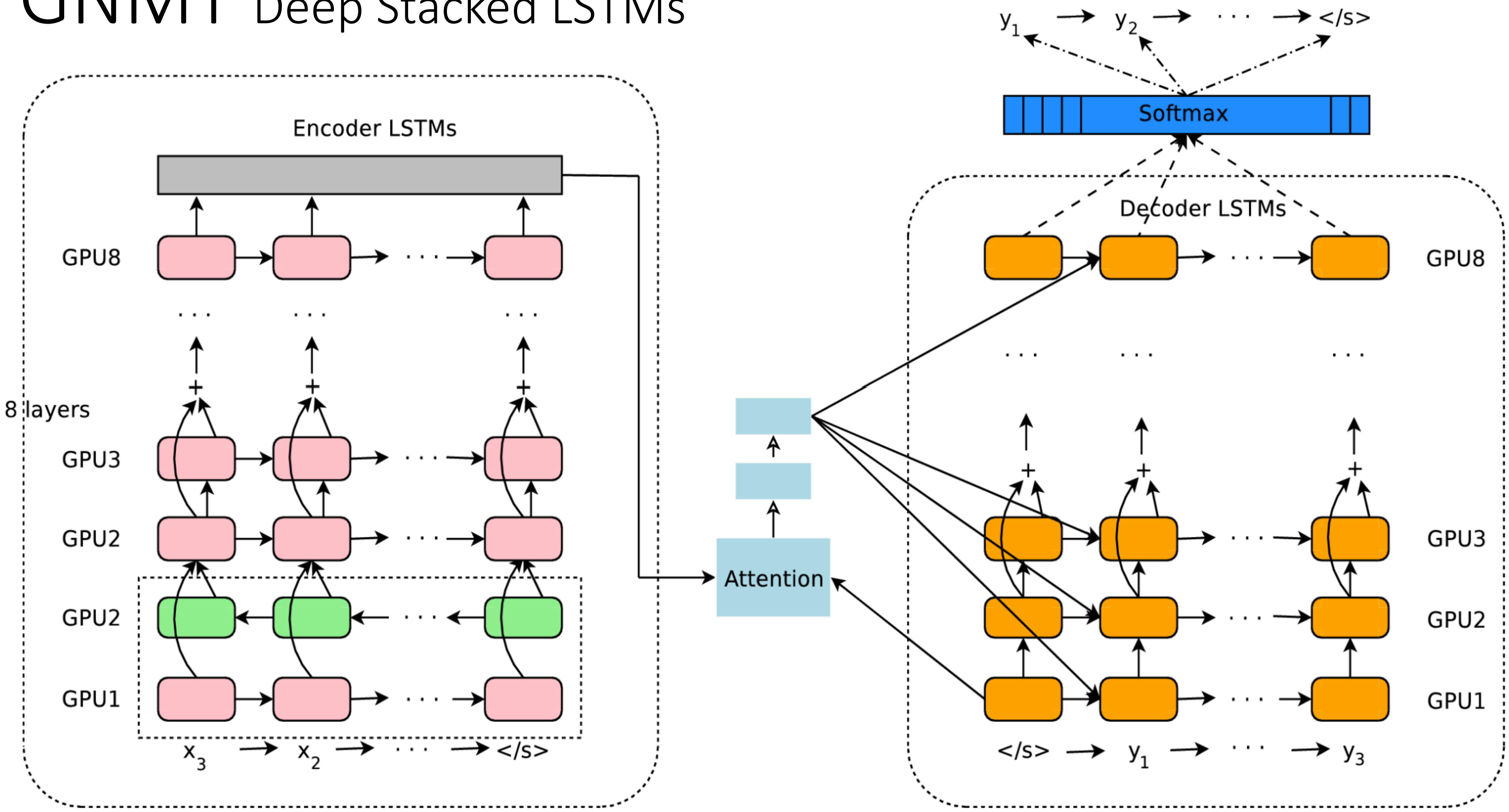
Related work

- The multilingual model architecture is identical to Google’s Neural Machine Translation (GNMT) system (Wu et al., 2016)

  Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (Wu et al., 2016)

- GNMT model consists of a deep LSTM network with 8 encoder and 8 decoder layers using residual connections and attention connections.
  - Accurate
  - Fast
  - Robustness to rare words

GNMT Deep Stacked LSTMs



GNMT attention module

- Context a_i for the current time step is computed according to the following formulas:

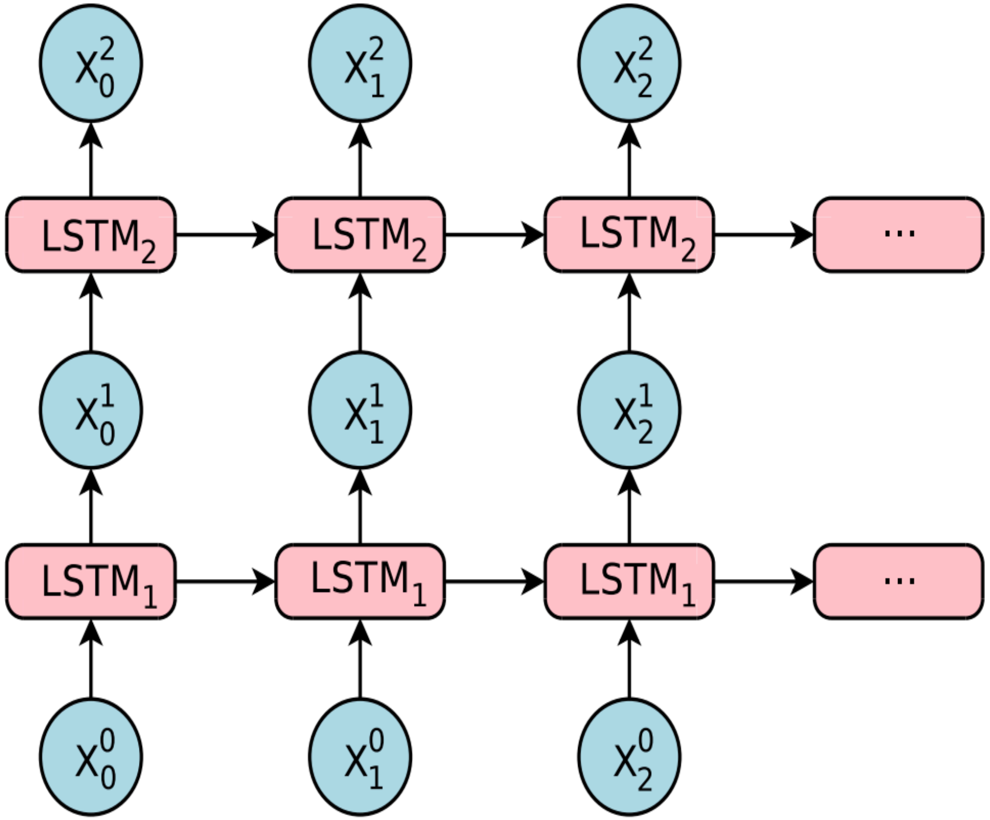
$$s_t = AttentionFunction(\mathbf{y}_{i-1}, \mathbf{x}_t) \quad \forall t, \quad 1 \le t \le M$$

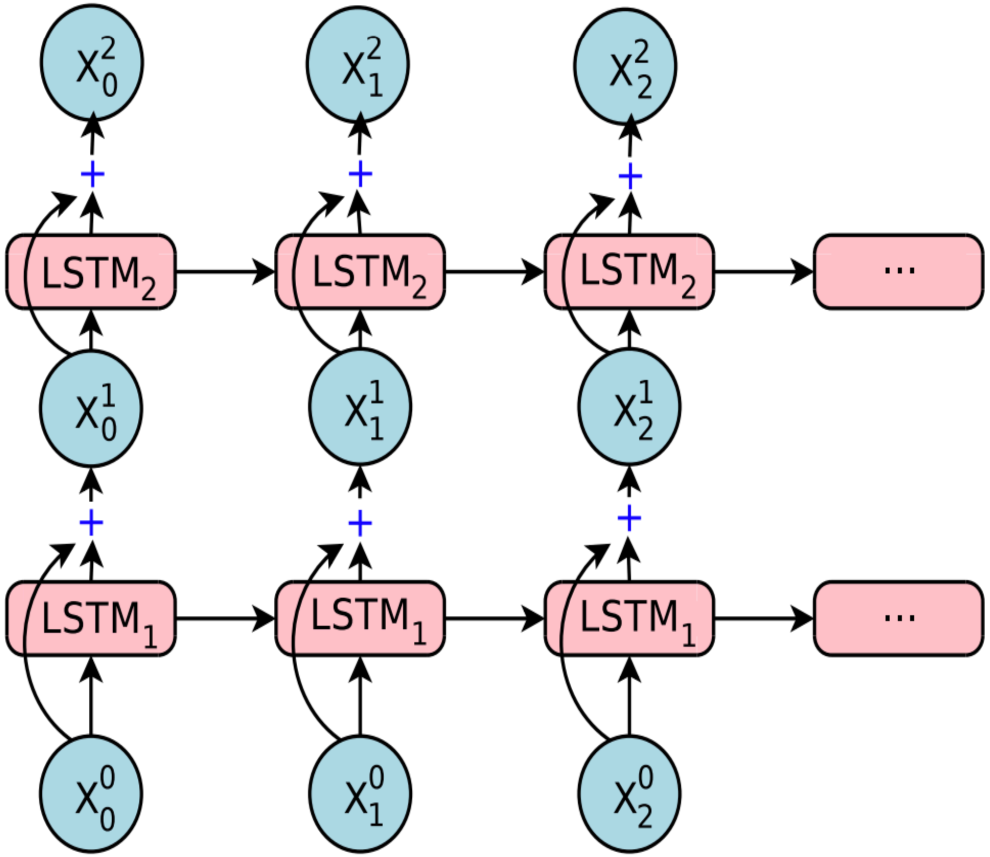
$$p_t = \exp(s_t) / \sum_{t=1}^{M} \exp(s_t) \qquad \forall t, \quad 1 \le t \le M$$

$$\mathbf{a}_i = \sum_{t=1}^{M} p_t . \mathbf{x}_t$$

- Here the *AttentionFunction* is a feed forward network with one hidden layer.

GNMT Residual Connections





GNMT Residual Connections

$$\mathbf{c}_t^i, \mathbf{m}_t^i = \mathrm{LSTM}_i(\mathbf{c}_{t-1}^i, \mathbf{m}_{t-1}^i, \mathbf{x}_t^{i-1}; \mathbf{W}^i)$$
$$\boxed{\mathbf{x}_t^i = \mathbf{m}_t^i}$$
$$\mathbf{c}_t^{i+1}, \mathbf{m}_t^{i+1} = \mathrm{LSTM}_{i+1}(\mathbf{c}_{t-1}^{i+1}, \mathbf{m}_{t-1}^{i+1}, \mathbf{x}_t^i; \mathbf{W}^{i+1})$$

- With residual connections between LSTM_i and LSTM_{i+1}, the above equations become:

$$\mathbf{c}_t^i, \mathbf{m}_t^i = \mathrm{LSTM}_i(\mathbf{c}_{t-1}^i, \mathbf{m}_{t-1}^i, \mathbf{x}_t^{i-1}; \mathbf{W}^i)$$
$$\boxed{\mathbf{x}_t^i = \mathbf{m}_t^i + \mathbf{x}_t^{i-1}}$$
$$\mathbf{c}_t^{i+1}, \mathbf{m}_t^{i+1} = \mathrm{LSTM}_{i+1}(\mathbf{c}_{t-1}^{i+1}, \mathbf{m}_{t-1}^{i+1}, \mathbf{x}_t^i; \mathbf{W}^{i+1})$$

GNMT Wordpiece Model

- To address the translation of out-of-vocabulary (OOV) words, GNMT applies sub-word units to do segmentation.
- Example:

  *Word: **Jet** makers **feud** over seat width with big orders at stake.*

  *Wordpieces: **_J et** _makers **_fe ud** _over _seat _width _with _big _orders _at _stake.*
- This method provides a good balance between the flexibility of “character”-delimited models and the efficiency of “word”-delimited models.

GNMT with zero-shot translation

- Based on the GNMT, the system adds an artificial token at the beginning of the input sentence to indicate the target language the model should translate to.
- Exmaple: En→Es

Instead of :

How are you? -> ¿Cómo estás?

put <2es> at the beginning:

<2es> How are you? -> ¿Cómo estás?

Zero-shot translation

- The system use implicit bridging to deal with the problem. No explicit parallel training data has been seen.
  - Although the source and target languages should be seen individually during the training at some point.

Table 5: Portuguese→Spanish BLEU scores using various models.

	Model	Zero-shot	BLEU
(a)	PBMT bridged	no	28.99
(b)	NMT bridged	no	30.91
(c)	NMT Pt→Es	no	31.50
(d)	Model 1 (Pt→En, En→Es)	yes	21.62
(e)	Model 2 (En↔{Es, Pt})	yes	24.75
(f)	Model 2 + incremental training	no	31.77

To improve zero-shot translation quality

- Incrementally training the multilingual model on the additional parallel data for the zero-shot directions.
- Zero-shot:
  En↔{Be,Ru,Uk}
- From-scratch:
  En↔{Be,Ru,Uk}
  + Ru↔{Be, Uk}
- Incremental:
  Zero-shot
  + From-scratch

	Zero-Shot	From-Scratch	Incremental
En→Be	16.85	17.03	16.99
En→Ru	22.21	22.03	21.92
En→Uk	18.16	17.75	18.27
Be→En	25.44	24.72	25.54
Ru→En	28.36	27.90	28.46
Uk→En	28.60	28.51	28.58
Be→Ru	56.53	82.50	78.63
Ru→Be	58.75	72.06	70.01
Ru→Uk	21.92	25.75	25.34
Uk→Ru	16.73	30.53	29.92

Mixed language

- Can a multilingual model successfully handle multi-language input (code-switching) in the middle of a sentence?
- Yes! Because the individual characters/wordpieces are present in the shared vocabulary.

- **Japanese:** 私は東京大学の学生です。 → I am a student at Tokyo University.
- **Korean:** 나는 도쿄 대학의 학생입니다. → I am a student at Tokyo University.
- **Japanese/Korean:** 私は東京大学학생입니다. → I am a student of Tokyo University.

Mixed language (2)

- What happens when a multilingual model is triggered with a linear mix of two target language tokens?

- Example:

  Using a multilingual En→{Ja, Ko} model, feed a linear combination (1−w)<2ja>+w<2ko> of the embedding vectors for “<2ja>” and “<2ko>”, 0 <= w <= 1.

  Result : with w = 0.5, the model switches languages mid-sentence.

w_{ko}	I must be getting somewhere near the centre of the earth.
0.00	私は地球の中心の近くにどこかに行っているに違いない。
0.40	私は地球の中心近くのどこかに着いているに違いない。
0.56	私は地球の中心の近くのどこかになっているに違いない。
0.58	私は지구の中心의가까이에어딘가에도착하고있어야한다。
0.60	나는지구의센터의가까이에어딘가에도착하고있어야한다。
0.70	나는지구의중심근처어딘가에도착해야합니다。
0.90	나는어딘가지구의중심근처에도착해야합니다。
1.00	나는어딘가지구의중심근처에도착해야합니다。

Conclusion

- Use a single model where all parameters are shared, which improves the translation quality of low resource languages in the mix.
- Zero-shot translation without explicit bridging is possible.
- To improve the zero-shot translation quality:

  Incrementally training the multilingual model on the additional parallel data for the zero-shot directions.
- Mix languages on the source or target side can yield interesting but reliable translation results.

Thank you!