Input Combination Strategies for Multi-Source Transformer Decoder

Jindřich Libovický, Jindřich Helcl, David Mareček

📅 November 1, 2018

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Overview

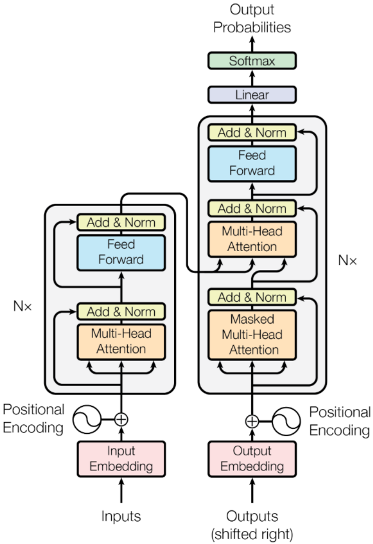
1. Transformer decoder overiew
2. Input combination strategies
3. Experiments
    - Multimodal translation
    - Multi-source translation

Overview

1. **Transformer decoder overiew**
2. Input combination strategies
3. Experiments
   - Multimodal translation
   - Multi-source translation

Transformer

- Architecture for sequence-to-sequence learning
- Encoder and decoder part
- Consists of attention and feed-forward layers only



Encoder-Decoder Attention

Scaled dot-product attention:

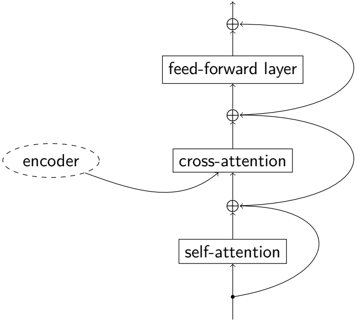
$$\mathcal{A}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) V.$$

Multi-headed setup:

$$\mathcal{A}^h(Q, K, V) = \sum_{i=1}^{h} C_i W_i^O$$

$$C_i = \mathcal{A}(QW_i^Q, KW_i^K, VW_i^V)$$

W^Q, W^K, $W^V \in \mathbb{R}^{d \times d_h}$ trainable

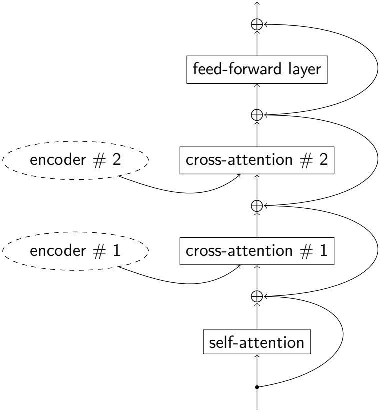


Overview

1. Transformer decoder overiew
2. **Input combination strategies**
3. Experiments
   - Multimodal translation
   - Multi-source translation

Serial

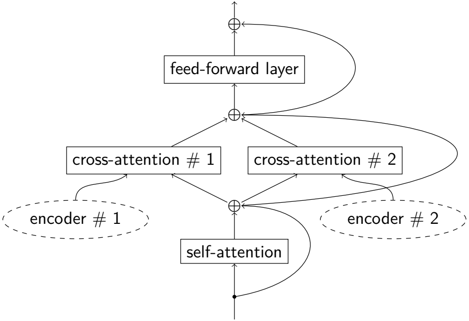
Stack the layers after each other.



Parallel

Run attentions independently, sum up the outputs.

$$\mathcal{A}^{h}_{\textit{para}}(Q, K_{1:n}, V_{1:n}) = \sum_{i=1}^{n} \mathcal{A}^{h}(Q, K_i, V_i)$$

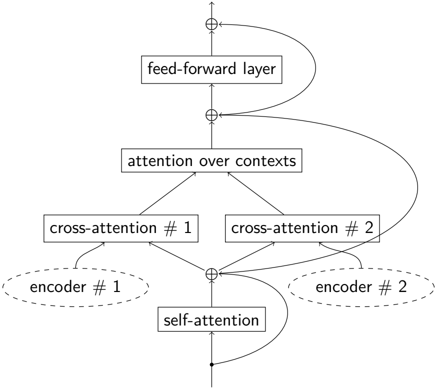


Hierarchical

Run the attentions independently, put another attention layer on top.

$$K_{\textit{hier}} = V_{\textit{hier}} = \mathsf{concat}_i(\mathcal{A}^h(Q, K_i, V_i))$$

$$\mathcal{A}^h_{\textit{hier}}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{\textit{hier}}, V_{\textit{hier}})$$

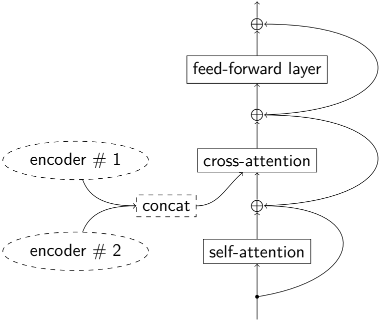


Flat

Concatenate the input states, then run a single attention layer.

$$K_{\textit{flat}} = V_{\textit{flat}} = \mathsf{concat}_i(K_i)$$

$$\mathcal{A}^h_{\textit{flat}}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{\textit{flat}}, V_{\textit{flat}})$$



Overview

1. Transformer decoder overiew
2. Input combination strategies
3. **Experiments**
    - **Multimodal translation**
    - **Multi-source translation**

Multimodal Translation – Task Overview

- Translation of image captions from Flickr30k dataset
- Multi30k dataset: images with English captions, German, French and Czech translations

Source:
en: A boy in a red suit plays in the water.

Targets:
de: Ein Junge in einem roten Badeanzug spielt im Wasser.
fr: Un garçon en maillot de bain rouge joue dans l'eau.
cs: Chlapec v červených plavkách si hraje ve vodě.

Multimodal Translation – Experiment Setup

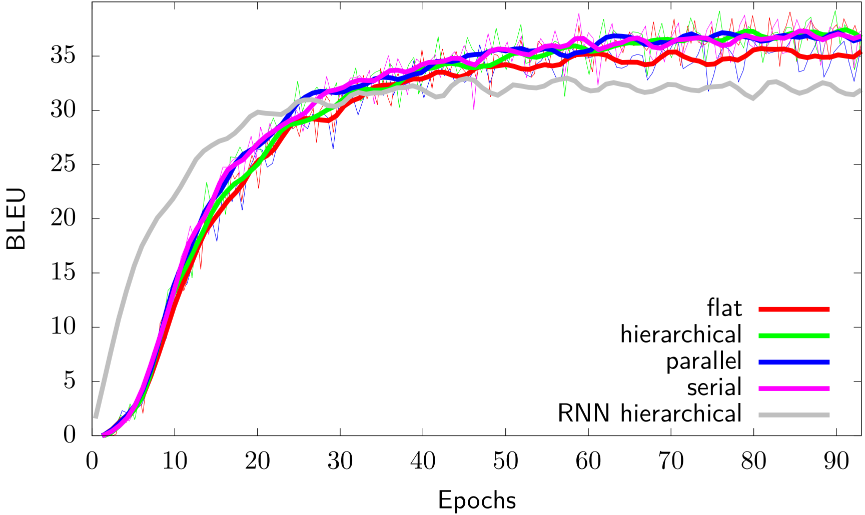
- Model dimension 512
- 6 layers in both encoder and decoder
- Vocabulary of approx. 20k wordpieces
- Image representation: convolutional maps from ResNet

Multimodal Translation – Results

	en→de		en→fr		en→cs	
	BLEU	adv.BLEU	BLEU	adv.BLEU	BLEU	adv.BLEU
baseline	38.3 ± .8	—	59.6 ± .9	—	30.9 ± .8	—
serial	38.7 ± .9	37.3 ± .6	60.8 ± .9	58.9 ± .9	31.0 ± .8	29.7 ± .8
parallel	38.6 ± .9	38.2 ± .8	60.2 ± .9	58.9 ± .9	31.1 ± .9	30.4 ± .8
flat	37.1 ± .8	35.7 ± .8	58.0 ± .9	57.0 ± .9	29.9 ± .8	28.2 ± .8
hierarchical	38.5 ± .8	38.1 ± .8	60.8 ± .9	60.2 ± .9	31.3 ± .9	31.0 ± .8

Quantitative results of the MMT experiments on the 2016 test set. Column ‘adv. BLEU’ is an adversarial evaluation with randomized image input.

Multimodal Translation – Learning Curves



Multi-Source Translation – Task Overview

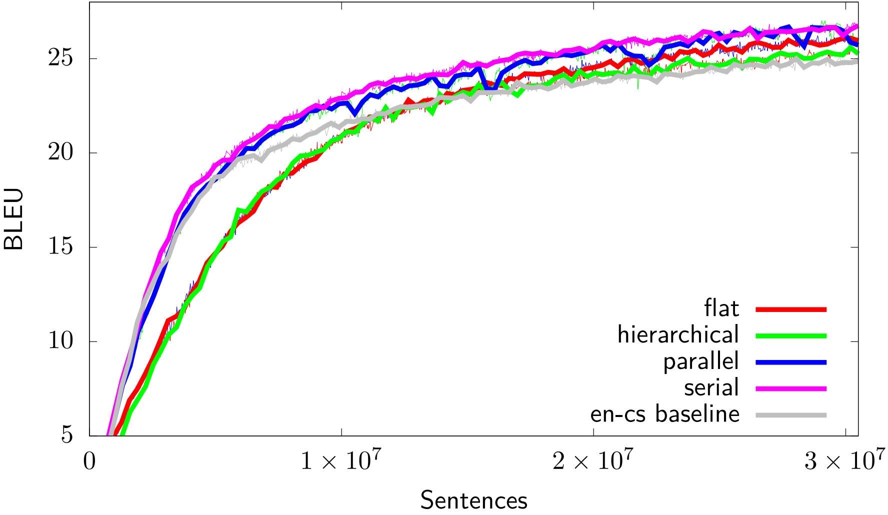
- Source languages: English, German, French, Spanish
- Target language: Czech
- Data: intersection of Europarl, 511k five-way parallel sentences
- Shared vocabulary of 42k wordpieces
- Model dimension 256, 6 layers in both encoder and decoder

Multi-Source Translation – Results

	BLEU	Adversarial evaluation (BLEU)			
		en	de	fr	es
baseline	16.5 ± .5	—	—	—	—
serial	20.5 ± .6	8.1 ± .4	19.7 ± .5	19.5 ± .6	18.4 ± .5
parallel	20.5 ± .6	1.4 ± .2	18.7 ± .5	17.9 ± .5	20.3 ± .5
flat	20.4 ± .6	0.2 ± .1	19.9 ± .6	20.0 ± .6	19.6 ± .5
hierarchical	19.4 ± .5	4.2 ± .3	18.3 ± .5	18.3 ± .5	15.3 ± .5

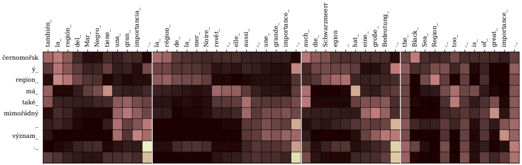
Quantitative results of the MMT experiment. The adversarial evaluation shows the BLEU score when one input language was changed randomly.

Multi-Source Translation – Learning Curves

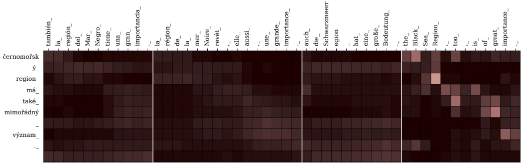


Multi-Source Translation – Analysis

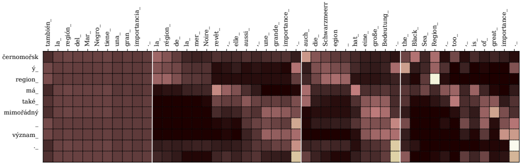
Serial



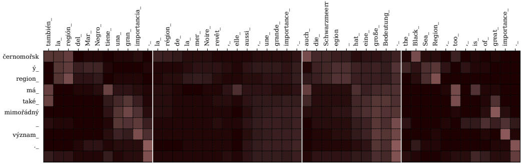
Flat



Parallel



Hierarchical



Visualization of attention for sentence *The Black Sea region, too, is of great importance.*
Language order in figures: *es, fr, de, en*

Conclusions

- Introduced 4 strategies: serial, parallel, hierarchical, flat
- All strategies perform approximately the same
- Slightly better than text-only baseline for multimodal MT
- Multi-source MT better than single-source

https://ufal.mff.cuni.cz