PhD Thesis Defense

Multimodality in Neural Machine Translation

Jindřich Libovický

June 13, 2019

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

Outline of the Talk

Multimodal Machine Translation

Multi-Source RNN Sequence-to-Sequence Learning

Multi-Source Transformer Model

Further Improving Translation Quality

Multimodal Machine Translation

Multimodal Translation

Translation of image description from one language into another.

Source:

en: A boy in a red suitsuit plays in the water.

Targets:

de: Ein Junge in einem roten BadeanzugBadeanzug spielt im Wasser.

fr: Un garçon en maillot de bainmaillot de bain rouge joue dans l'eau.

cs: Chlapec v červených plavkáchplavkách si hraje ve vodě.

Multimodal Translation: Lexical Ambiguity Again

Sentence context seems to be enough: skála vs. kámen

Source:

en: A woman sitting on a very large rock smiling at the camera with trees in the background.

Targets:

de: Eine Frau sitzt vor Bäumen im Hintergrund auf einem sehr großen Stein und lächelt in die Kamera.

fr: Une femme assise sur un très gros rocher, souriant pour la photo avec des arbres en arrière-plan.

cs: Žena sedící na velmi velkém kameni, usmívající se na kameru, se stromy v pozadí.

Multimodal Translation: Gender

Some languages have gendered nouns, some not.

Source:

en: A baseball player in a black shirt just tagged a player in a white shirt.

Targets:

de: Eine Baseballspielerin in einem schwarzen Shirt fängt eine Spielerin in einem weißen Shirt.

fr: Une joueuse de baseball en maillot noir vient de toucher une joueuse en maillot blanc.

cs: Basebalistka v černém triku právě vyoutovala hráčku v bílém triku.

Multi30k Dataset

- Based on Flickr30k dataset: 30k images with crowsourced descriptions
- Language: simple sentences, present tense, no named entities
- One description selected for translation: crowd-sourced translation into German, French

For competition at WMT18 we created a Czech version

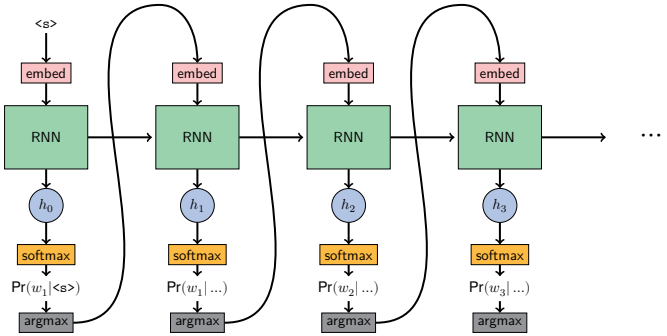
Available at:

https://github.com/multi30k/data

Multi-Source RNN Sequence-to-Sequence Learning

RNN Sequence-to-Sequence Model

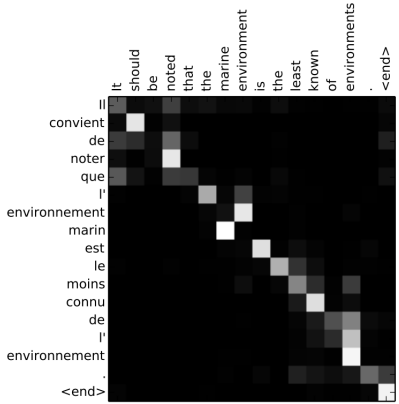
- two parts: encoder + decoder
- encoder bidirectional RNN
- decoder: autoregressively fed with its own outputs



Attentive Sequence-to-Sequence Learning

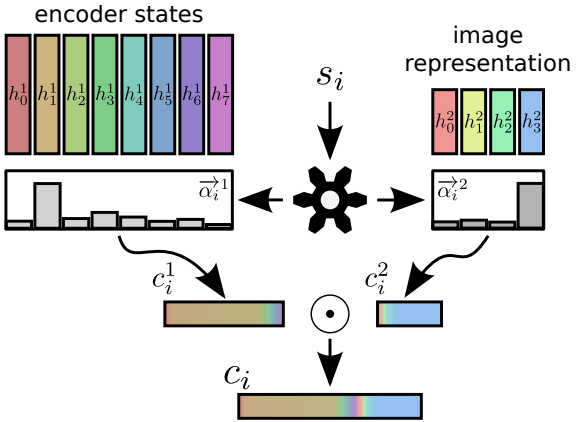
In each decoder step:

- compute distribution over encoder states given the decoder state
- the decoder gets a context vector to decide about its output



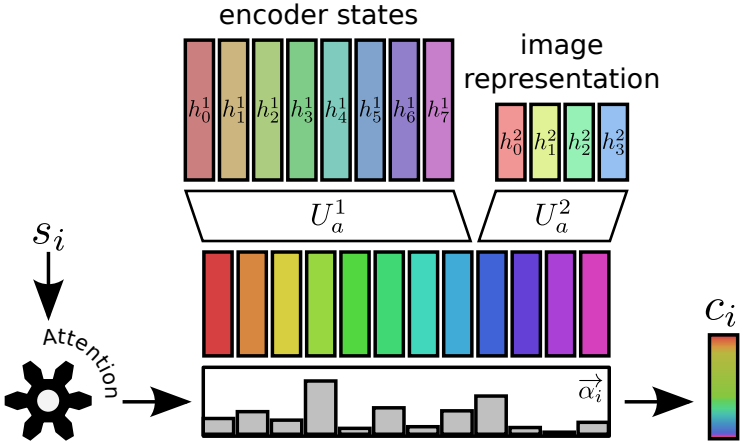
What about multiple inputs?

Context Vector Concatenation



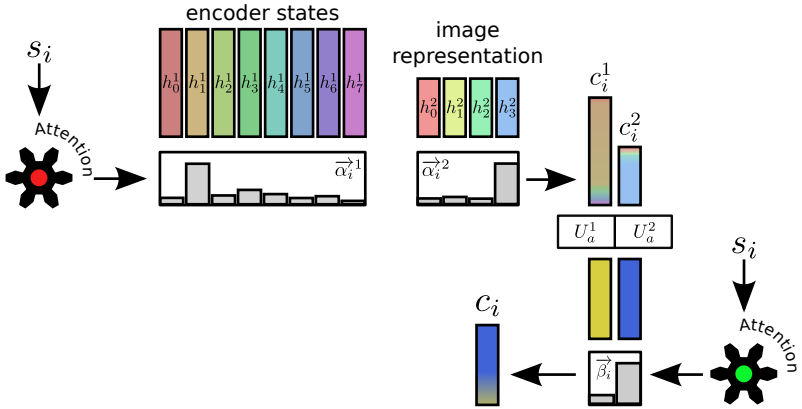
- Attention over input sequences computed independently.
- Combination resolved later on in the network

Flat Attention Combination



Importance of different inputs reflected in the **joint** attention distribution.

Hierarchical Attention Combination



Attention distribution is **factored** by input.

Translation Quality on Multi30k

	en→de		en→fr		en→cs	
	BLEU	adv.BLEU	BLEU	adv.BLEU	BLEU	adv.BLEU
text-only	36.7 ± .8	—	48.3 ± .8	—	30.0 ± .8	—
concatenation	35.7 ± .8	30.9 ± .8	47.7 ± .8	41.3 ± .8	29.3 ± .8	24.2 ± .8
flat	34.6 ± .8	33.8 ± .8	46.0 ± .9	43.5 ± .8	29.1 ± .8	26.5 ± .8
hierarchical	37.6 ± .8	34.2 ± .8	48.2 ± .9	44.9 ± .8	29.5 ± .8	28.1 ± .8

Adversarial evaluation: model provided with incorrect image.

In WMT17 **manual evaluation**, hierarchical strategy outperformed text-only model.

When multimodality helps (1)

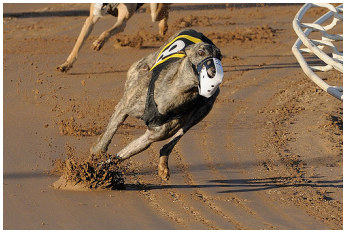
SRC a dark-haired beardedbearded man in glasses and a hawaiian shirt is sitting on the grass .

REF ein dunkelhaariger mann mit bartmit bart, brille und hawaiihemd sitzt auf dem gras .

TXT ein dunkelhaariger mann mit brille und einem hawaii-hemd sitzt auf dem gras .

MMT ein dunkelhaariger bärtigerbärtiger mann in brille und hawaiihemd sitzt auf dem gras .

When multimodality helps (2)



SRC	a muzzled greyhound dog wearing yellow and black is running on the track .
REF	ein windhund mit maulkorb in gelb und schwarz läuft auf der strecke .
TXT	ein windhund mit maulkorb in der hand und schwartzer kleidung läuft auf dem weg .
MMT	ein windhund mit maulkorb in gelber und schwarzer rennt auf der rennstrecke .

Hierarchical Attention Example

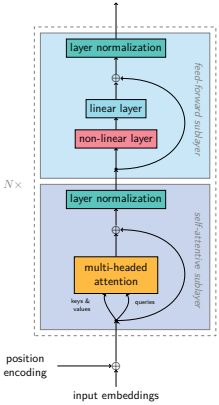
Source: a brown dog is running after the black dog .

Reference de: ein brauner hund rennt dem schwarzen hund hinterher .
Reference fr: un chien brun court après le chien noir .
Reference cs: hnědý pes běží za černým psem .

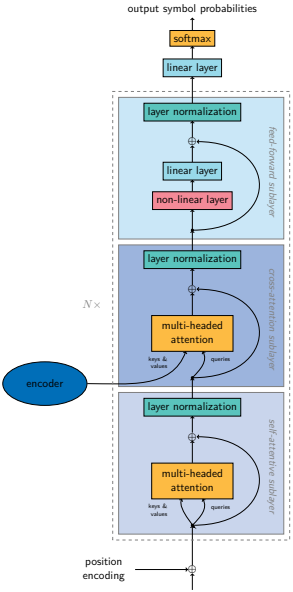
Output with attention:

Multi-Source Transformer Model

Transformer

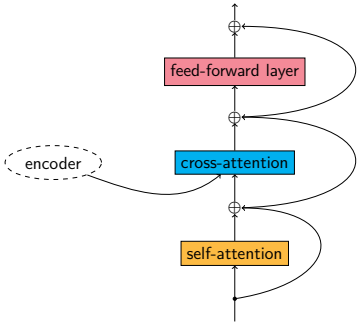


- Architecture for sequence-to-sequence learning
- Encoder and decoder part
- Consists of attention and feed-forward layers only

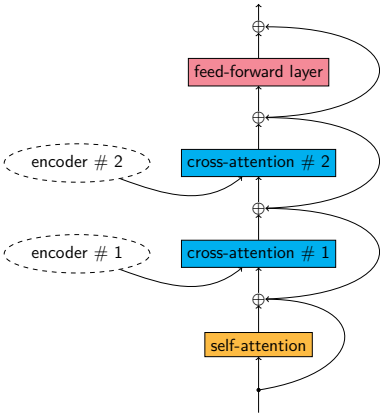


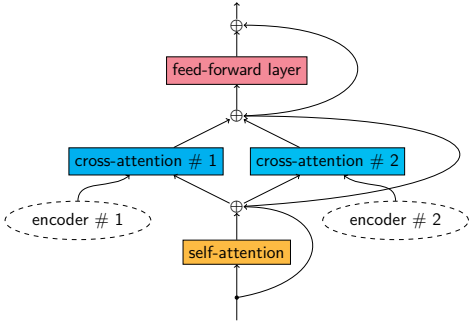
Encoder-Decoder Attention

- Encoder and decoder have several layers
- Each layer = three sub-layers:
  - Attending to previously generated words
  - Attending from encoder to decoder
  - Non-linear transformation of states

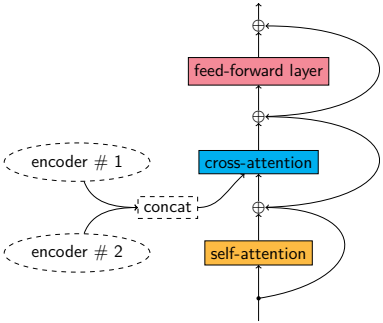


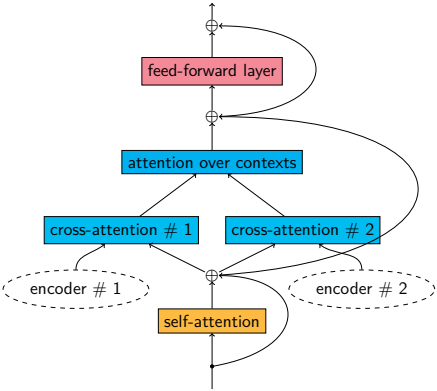
Serial & Parallel





Flat & Hierarchical





Multimodal Translation: Results

	en→de		en→fr		en→cs	
	BLEU	adv.BLEU	BLEU	adv.BLEU	BLEU	adv.BLEU
RNN text-only	36.7 ± .8	—	48.3 ± .8	—	30.0 ± .8	—
Transformer text-only	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.

When multimodality helps in Transformers

SRC a man in an orange jumpsuit and matching hard hat is helping with a blue hose

REF ein mann in einem orangefarbenen overall und passendem schutzhelm hilft mit einem blauen schlauch

TXT ein mann in orangefarbenem overall und mit goldener kopfbedeckung hilft einem blauen schlauch .

MMT ein mann in einem orangefarbenen overall und passenden schutzhelm hilft mit einem blauen schlauch .

Multi-Source Translation – Task Overview

es También la región del Mar Negro tiena una gran importancia.
fr La région de la mer Noire revêt, elle aussi, une grande importance.
de Auch die Schwarzmeerregion hat eine große Bedeutung.
en The Black Sea region, too, is of great importance.

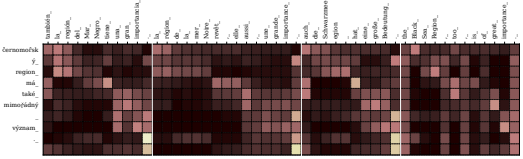
Černomořský region má také mimořádný význam.

Multi-Source Translation – Results

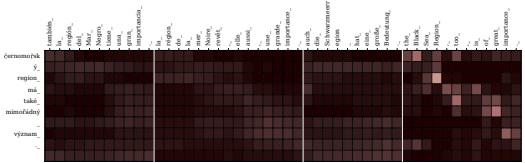
	BLEU	Adversarial evaluation (BLEU)			
		en	de	fr	es
baseline	18.5 ± .5	—	—	—	—
serial	20.5 ± .6	8.1 ± .4	19.7 ± .5	19.5 ± .6	18.4 ± .5
parallel	20.5 ± .6	1.4 ± .2	18.7 ± .5	17.9 ± .5	20.3 ± .5
flat	20.4 ± .6	0.2 ± .1	19.9 ± .6	20.0 ± .6	19.6 ± .5
hierarchical	19.4 ± .5	4.2 ± .3	18.3 ± .5	18.3 ± .5	15.3 ± .5

Multi-Source Translation – Analysis

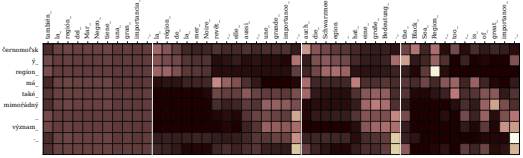
Serial



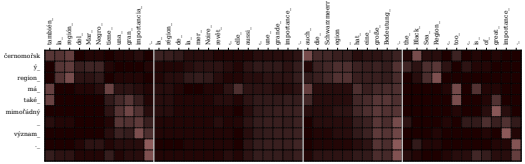
Flat



Parallel



Hierarchical



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

Further Improving Translation Quality

Get More Data

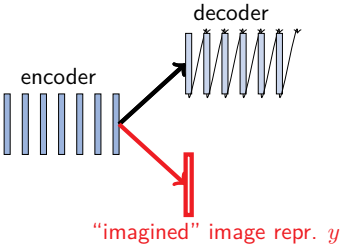
1. More textual data

- Treat the task as domain adaptation
- Trained char.-level LM, used to score sentences by perplexity
- Filter available parallel corpora
- Filter monolingual corpora in target language, do backtranslation

→ orders of magnitude bigger textual training data

2. More image data

- Imagination model (Elliott and Kádár, 2017): multi-task learning – translation + predict image representation
- Images used at train time only, at inference time



Results

		en→de	en→fr	en→cs
Multi30k	Textual	38.3 ± .8	59.6 ± .9	30.9 ± .8
	Multimodal (serial)	38.7 ± .8	60.8 ± .9	31.0 ± .8
More data	+ parallel data	40.4 ± .9	62.5 ± .8	32.3 ± .9
	+ Imagination	42.6 ± .8	62.8 ± .9	36.3 ± .9

Main Contributions

- Czech version of the Multi30k dataset
- Introduced novel deep learning models that utilize both the textual and visual information
- Competitive models using domain adaptation and multi-task learning

https://ufal.mff.cuni.cz/jindrich-libovicky

Supplemental Slides

Supplemental Slides

Multi30k Dataset

Multi30k Dataset: Statistics

split	sentences	English			German		
		tok./sent.	ch./tok.	OOV	tok./sent.	ch./tok.	OOV
train	29,000	13.0	4.8	—	12.4	5.9	—
validation	1,014	13.1	4.8	1.2%	12.7	6.0	3.0%
test 2016	1,000	13.0	4.9	1.1%	12.1	5.9	2.6%
		French			Czech		
		tok./sent.	ch./tok.	OOV	tok./sent.	ch./tok.	OOV
train	29,000	14.1	5.4	—	10.2	6.0	—
validation	1,014	14.2	5.4	1.2%	10.2	5.9	3.9%
test 2016	1,000	14.0	5.5	1.1%	10.5	6.0	4.0%

Czech Multi30k: Annotation Quality

	Proportion of data	Annotator agreement
No spelling errors	94%	92%
Stylistically appropriate	75%	73%
Adequate in meaning	96%	93%
No inappropriate lexical anglicism	94%	90%
No inappropriate syntactic anglicism	93%	91%

Supplemental Slides

RNN Sequence-to-Sequence Models

Attentive Sequence Learning

In each decoder step i

- compute distribution over encoder states given the decoder state
- the decoder gets a context vector to decide about its output

$$e_{ij} = v_a^\top \tanh(W_a s_i + U_a h_j) \quad (1)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (2)$$

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j \quad (3)$$

Flat Attention Combination

$$
\begin{aligned}
\textbf{one source} &\rightarrow N \textbf{ sources} \\
e_{ij} = v_a^\top \tanh(W_a s_i + U_a h_j) &\rightarrow e_{ij}^{(k)} = v_a^\top \tanh(W_a s_i + U_a{}^{(k)} h_j) \\
\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} &\rightarrow \alpha_{ij}^{(k)} = \frac{\exp(e_{ij}^{(k)})}{\sum_{n=1}^{N} \sum_{m=1}^{T_x^{(n)}} \exp\left(e_{im}^{(n)}\right)} \\
c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j &\rightarrow c_i = \sum_{k=1}^{N} \sum_{j=1}^{T_x^{(k)}} \alpha_{ij}^{(k)} U_c{}^{(k)} h_j^{(k)}
\end{aligned}
$$

- $U_a^{(k)}$, $U_c^{(k)}$ project states to a common space
- Question: Should $U_a^{(k)} = U_c^{(k)}$? (i.e. should the projection parameters be shared?)

Hierarchical Attention Combination

1.

$k = 1 \dots N$

inputs

Compute the context vector:

$$c_i^{(k)} = \sum_{j=1}^{T_x^{(k)}} \alpha_{ij}^{(k)} h_j^{(k)}, \text{ where } \alpha_{ij}^{(k)} = \dots$$

2.

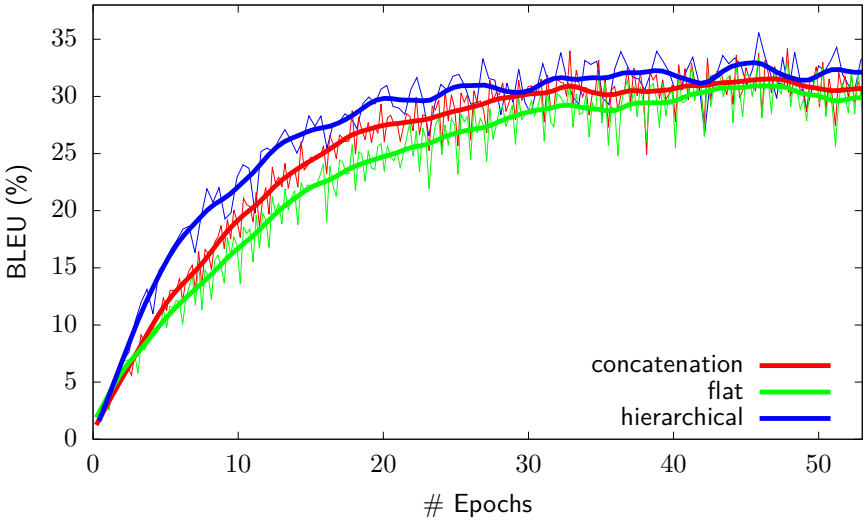
Compute another attention distribution over the intermediate context vectors $c_i^{(k)}$ and get the resulting context vector c_i.

$$\begin{aligned} e_i^{(k)} &= v_b^\top \tanh(W_b s_i + U_b^{(k)} c_i^{(k)}) \\ \beta_i^{(k)} &= \frac{\exp(e_i^{(k)})}{\sum_{n=1}^{N} \exp(e_i^{(n)})} \\ c_i &= \sum_{k=1}^{N} \beta_i^{(k)} U_c^{(k)} c_i^{(k)} \end{aligned}$$

RNN Multimodal Translation: Experiment Setup

- Bidirectional GRU encoder, 300 dimensions
- Conditional GRU decdoer, 500 dimenions
- Vocabulary of approx. 20k BPE subwords
- Image representation: convolutional maps from ResNet

Transformer Multimodal Translation – Learning Curves



ACL 2017 Results

	shared proj.	sentinel	MMT		APE	
			BLEU	METEOR	BLEU	HTER
baseline			32.4	49.3	62.3	24.8
concatenation			31.4 $\pm$ .8	48.0 $\pm$ .7	62.3 $\pm$ .5	24.4 $\pm$ .4
flat	×	×	30.2 $\pm$ .8	46.5 $\pm$ .7	62.6 $\pm$ .5	24.2 $\pm$ .4
	×		29.3 $\pm$ .8	45.4 $\pm$ .7	62.3 $\pm$ .5	24.3 $\pm$ .4
		×	30.9 $\pm$ .8	47.1 $\pm$ .7	62.4 $\pm$ .6	24.4 $\pm$ .4
			29.4 $\pm$ .8	46.9 $\pm$ .7	62.5 $\pm$ .6	24.2 $\pm$ .4
hierarchical	×	×	**32.1 $\pm$ .8**	**49.1 $\pm$ .7**	62.3 $\pm$ .5	24.1 $\pm$ .4
	×		28.1 $\pm$ .8	45.5 $\pm$ .7	62.6 $\pm$ .6	24.1 $\pm$ .4
		×	26.1 $\pm$ .7	42.4 $\pm$ .7	62.4 $\pm$ .5	24.3 $\pm$ .4
			22.0 $\pm$ .7	38.5 $\pm$ .6	62.5 $\pm$ .5	24.1 $\pm$ .4

Results on the Multi30k dataset and the APE dataset. The column 'share' denotes whether the projection matrix is shared for energies and context vector computation, 'sent.' indicates whether the sentinel vector has been used or not.

Supplemental Slides

Transformer Models

Encoder-Decoder Attention in Transformer

Scaled dot-product attention:

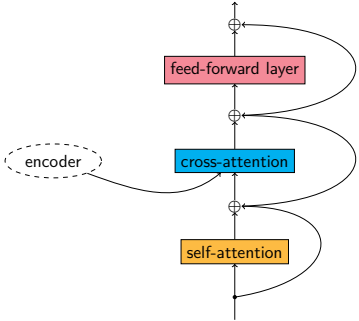
$$\mathcal{A}(Q,K,V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

Multi-headed setup:

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

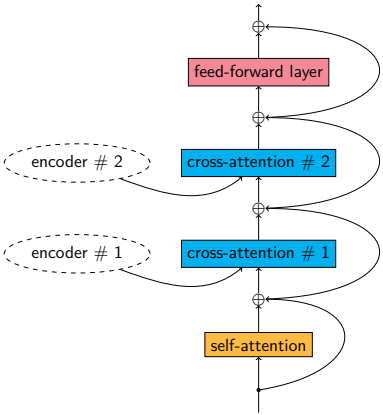
$$C_i = \mathcal{A}(QW_i^Q, KW_i^K, VW_i^V)$$

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



Serial

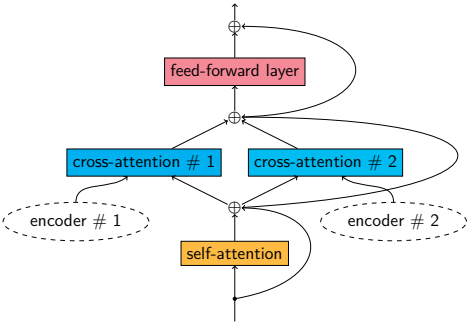
Stack the layers after each other.



Parallel

Run attentions independently, sum up the outputs.

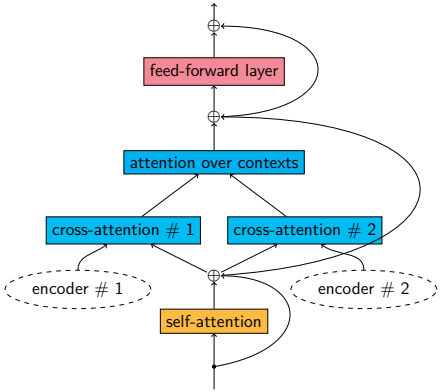
$$\mathcal{A}^h_{para}(Q, K_{1:n}, V_{1:n}) = \sum_{i=1}^{n} \mathcal{A}^h(Q, K_i, V_i)$$



Hierarchical

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

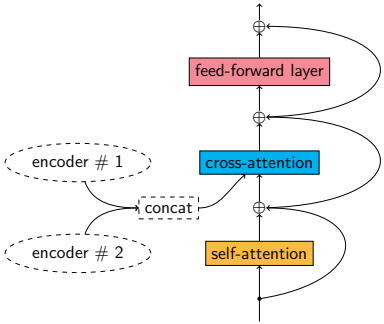
$$K_{hier} = V_{hier} = \text{concat}_i(\mathcal{A}^h(Q, K_i, V_i))$$
$$\mathcal{A}^h_{hier}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{hier}, V_{hier})$$



Flat

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

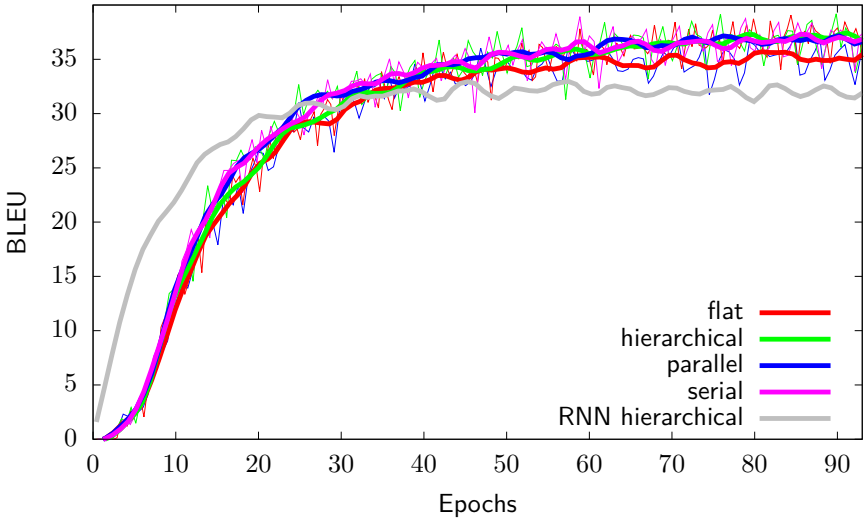
$$K_{flat} = V_{flat} = \text{concat}_i(K_i)$$
$$\mathcal{A}^h_{flat}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{flat}, V_{flat})$$



Transformer 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

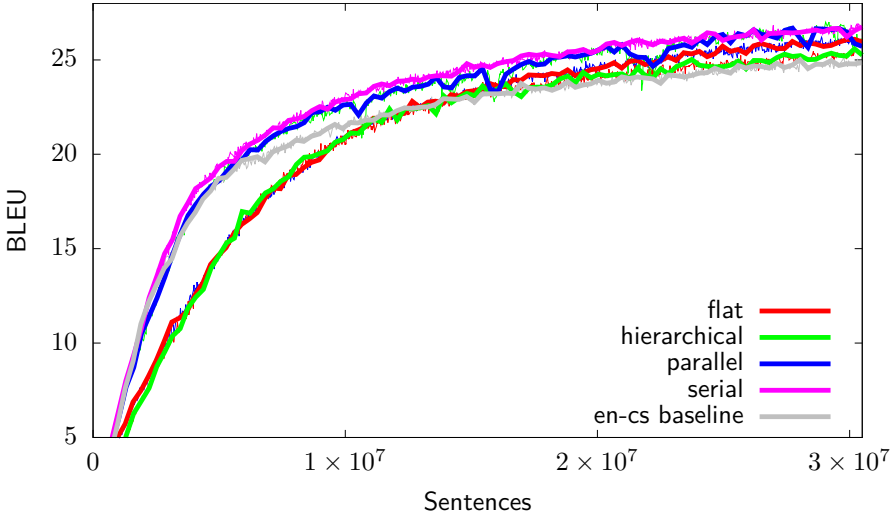
Transformer Multimodal Translation – Learning Curves



Multi-Source Translation – Task Steup

- 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 – Learning Curves



Supplemental Slides

Additional Data

Data for the Unconstrained Setup

	de	fr	cs
Multi30k		29k	
oversampling factor	273×	366×	9×
Task 2 BT	145k	—	—
in-domain parallel	3k	—	15k
in-domain back-translation	30k	—	29k
oversampling factor	39×	—	7×
EU Bookshop	9.3M	10.6M	445k
MS COCO (English only)		414k	

Equations for the Imagination Model

$$\hat{y}_{\text{img}} = W_2^R \,\text{ReLu}\left(W_1^R \sum_j h_j\right)$$

$$L_{\text{imag}} = \max\left(0, \alpha + \text{dist}(\hat{y}, y) - \text{dist}(\hat{y}, y_c)\right)$$

- where h_i are the encoder states, W_1^R and W_2^R are trainable parameters
- contrastive examples: randomly shuffled batch

Supplemental Slides

Assessing Sentence Representations

Image Retrieval & Semantic Textual Similarity

Sentence represenation = mean-pool of encoder states

Image Retrieval

- Canonical Correlation Analysis: find projection that is highly correlated with image representation
- Representation fixed
- Evaluate on image retrieval by captions (Recall @ 10)

Semantic Textual Similarity

- SemEval sentence similarity task: manually annotated semantic similarity of sentences
- General domain, not related to image captions
- Spearman correlation of cosine distance of representations and semantic similarity

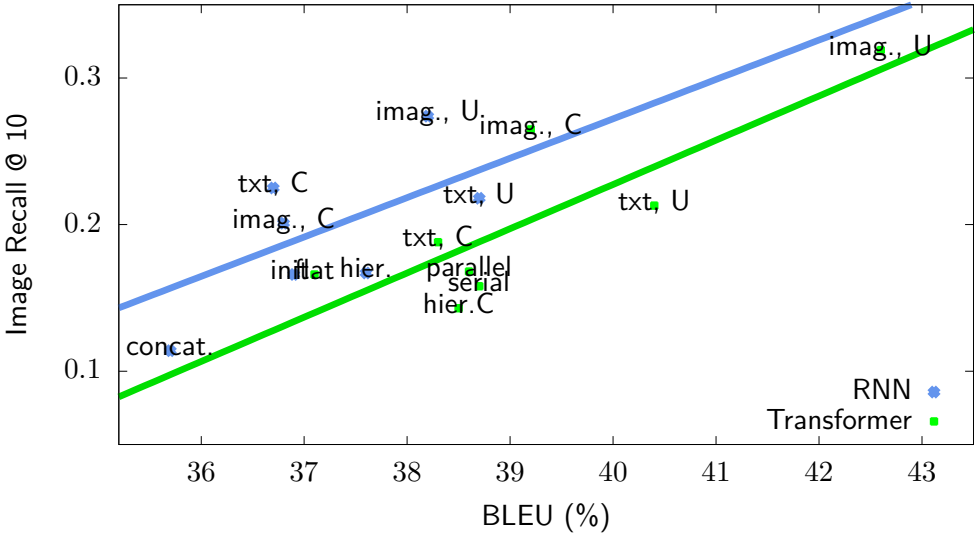
Results

Language Model	Img.	STS
RNN on Flickr30k	22.4	.340
ELMo	28.4	.631
BERT	22.4	.624

Textual MT		Img.	STS
RNN	Textual	22.5	.527
	+ parallel data	21.8	.621
	+ Imagination	27.4	.622
Trans.	Textual	18.8	.374
	+ parallel data	21.3	.509
	+ Imagination	31.9	.512

Multimodal MT		Img.	STS
RNN	Decoder init.	16.6	.536
	Att. concatenation	11.4	.429
	Flat att. comb.	14.6	.487
	Hierar. att. comb.	16.7	.553
Trans.	Serial att. comb.	15.8	.383
	Parallel att. comb.	16.8	.398
	Flat att. comb.	16.6	.385
	Hierar. att. comb.	14.3	.346

BLEU vs. Image Retrieval



BLEU vs. Semantic Similarity Performance