Multimodality in Neural Machine Translation

Jindřich Libovický

February 12, 2018

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

Outline of the Talk

Multimodal Machine Translation

Multimodal Translation

- Translation of image description from language into another
  = 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 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: Stein vs. Felsen

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.

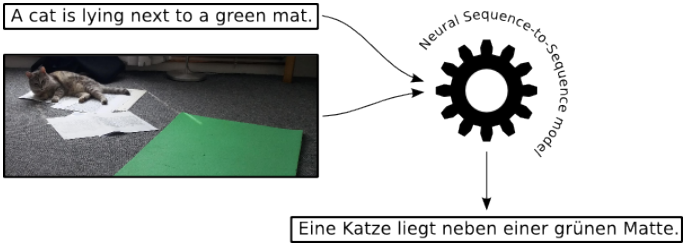
Multi30k Dataset

- 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 and Czech

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%

Multi-source Sequence-to-Sequence Learning

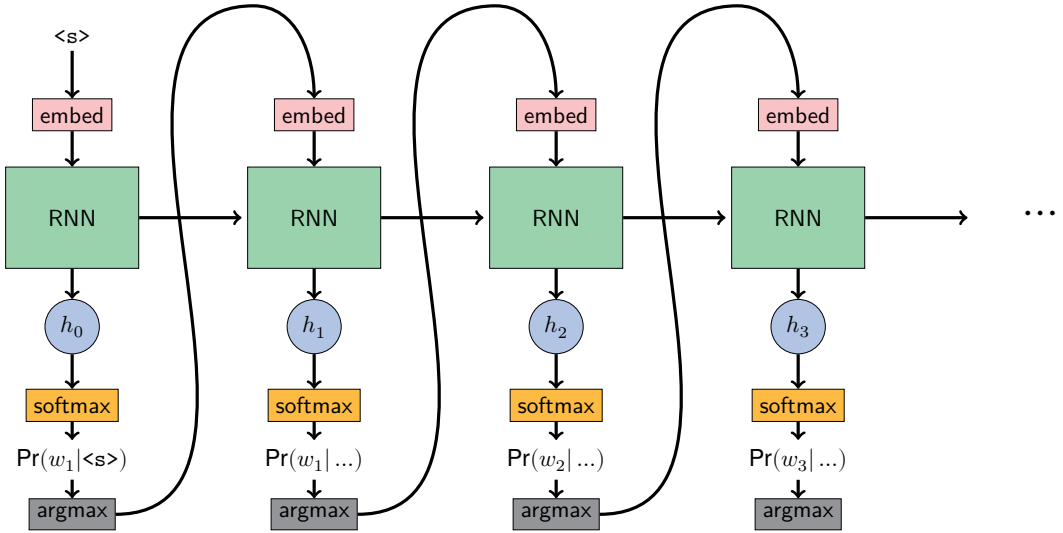


Other use cases

Automatic MT post-editing, Multi-source machine translation, ...

Multi-Source RNN Sequence-to-Sequence Learning

Autoregressive RNN decoding



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

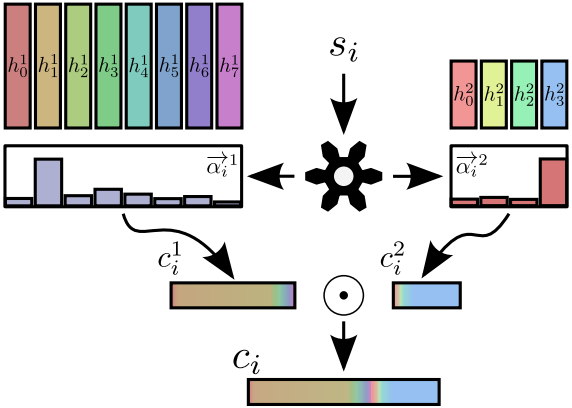
$$e_{ij} = v_a^\top \tanh(W_a s_i + U_a h_j) \quad (1)$$

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

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

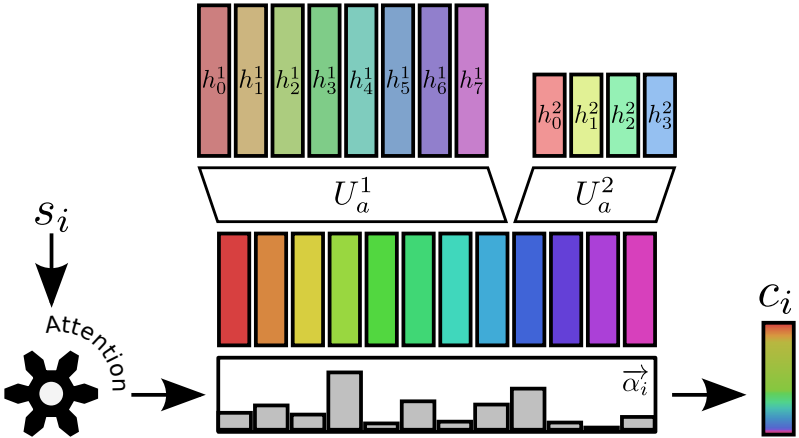
What about multiple inputs?

Context Vector Concatenation



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

Flat Attention Combination



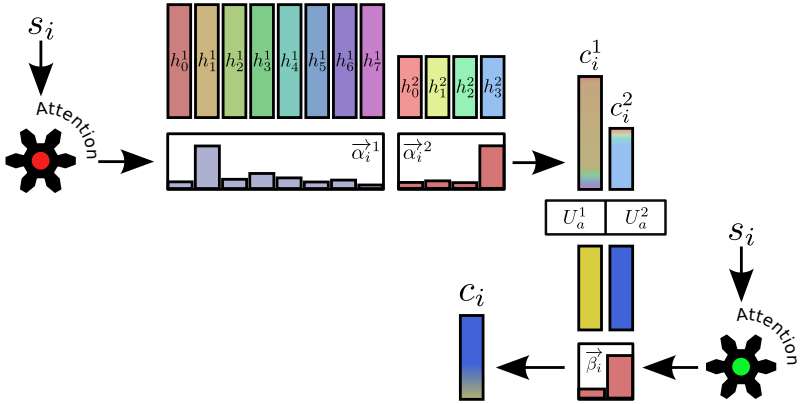
Importance of different inputs reflected in the **joint** attention distribution.

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

Hierarchical Attention Combination



Attention distribution is **factored** by input.

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.

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

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

Results

	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	—
decoder initialization	36.9 ± .8	35.8 ± .8	48.1 ± .8	48.1 ± .9	29.6 ± .8	29.3 ± .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

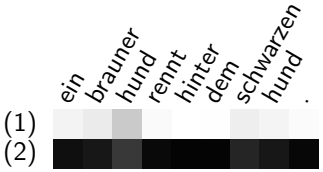
In WMT17 manual evaluation, hierarchical strategy outperformed text-only model.

Hierarchical Attention Example (1)

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:

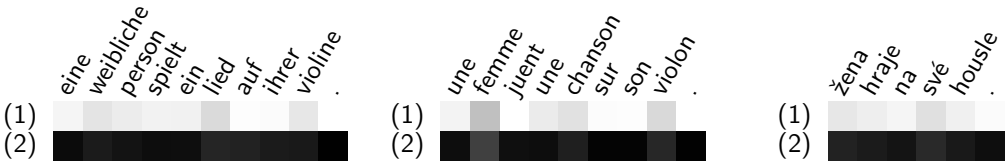


Hierarchical Attention Example (2)

Source: a female playing a song on her violin .

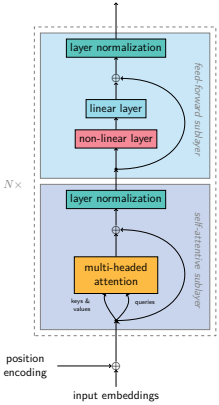
Reference de: eine frau spielt ein lied auf ihrem cello .
Reference fr: une femme jouant un morceau sur son violon .
Reference cs: žena hraje píseň na housle .

Output with attention:

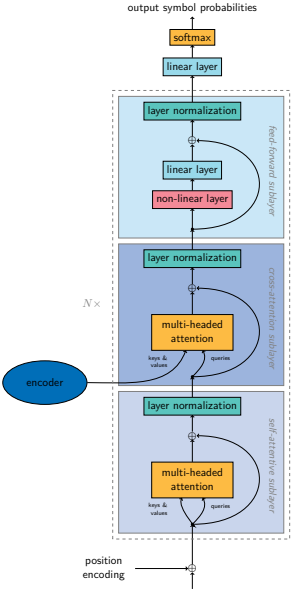


Multi-Source Transformer Model

Transformer



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



Encoder-Decoder Attention

Scaled dot-product attention:

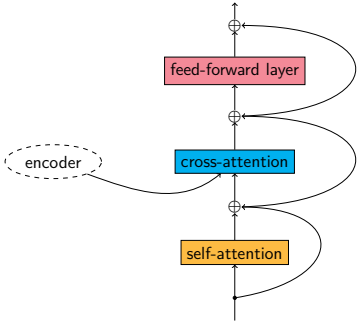
$$\mathcal{A}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V.$$

Multi-headed setup:

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

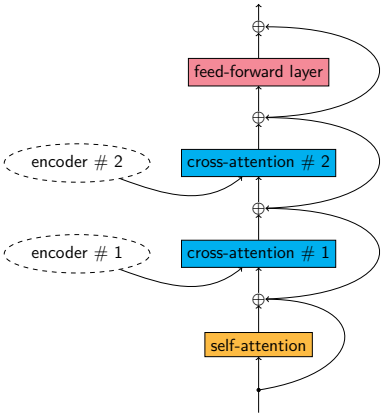
$$C_i = \mathcal{A}(QW_i^Q, KW_i^K, VW_i^V)$$

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



Serial

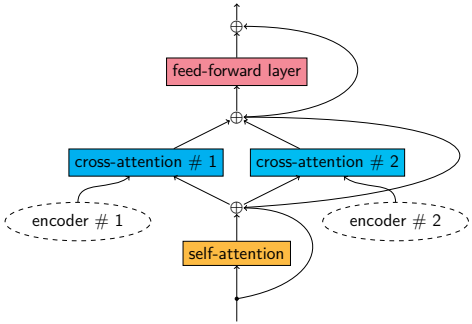
Stack the layers after each other.



Parallel

Run attentions independently, sum up the outputs.

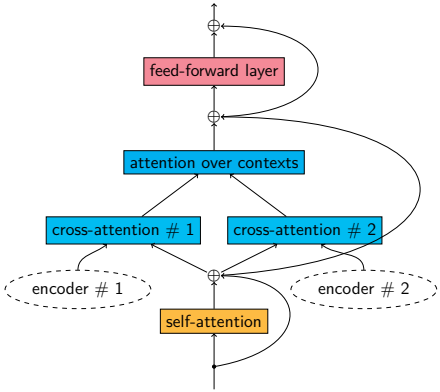
$$\mathcal{A}^h_{\textit{para}}(Q, K_{1:n}, V_{1:n}) = \sum_{i=1}^{n} \mathcal{A}^h(Q, K_i, V_i)$$



Hierarchical

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

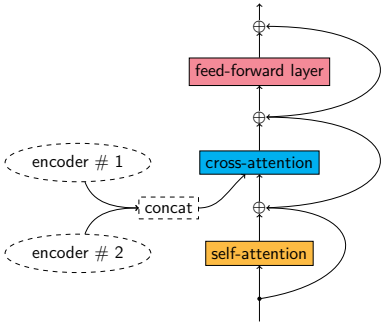
$$K_{hier} = V_{hier} = \text{concat}_i(\mathcal{A}^h(Q, K_i, V_i))$$
$$\mathcal{A}^h_{hier}(Q, K_{1:n}, V_{1:n}) = \mathcal{A}^h(Q, K_{hier}, V_{hier})$$



Flat

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

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



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
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. Column 'adv. BLEU' is an adversarial evaluation with randomized image input.

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

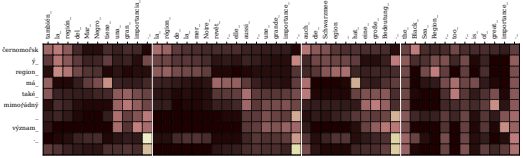
Multi-Source Translation – Results

	BLEU	Adversarial evaluation (BLEU)			
		en	de	fr	es
baseline	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

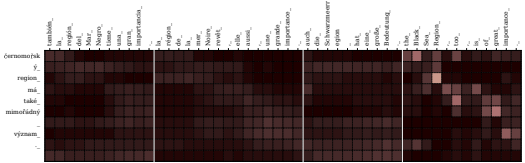
Quantitative results of the MMT experiment. The adversarial evaluation shows the BLEU score when one input language was changed randomly.

Multi-Source Translation – Analysis

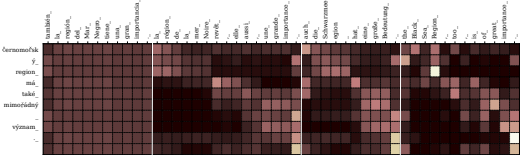
Serial



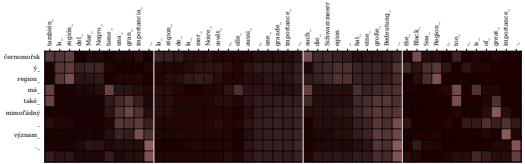
Flat



Parallel



Hierarchical



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

How to Score Well in the WMT Multimodal Task

Get More Data

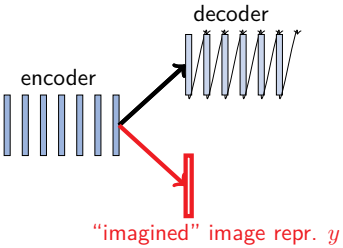
1. More textual data

- 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

$\rightarrow$ 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
RNN RNN	Cons.	Textual	36.7 ± .8	48.3 ± .8	30.0 ± .8
		Imagination	36.8 ± .8	47.6 ± .8	29.8 ± .8
		Multimodal (hierarchical)	37.6 ± .8	48.2 ± .9	29.5 ± .8
	Unc.	Textual	38.7 ± .8	42.8 ± .8	27.7 ± .8
		Imagination	38.2 ± .8	43.4 ± .8	31.0 ± .8
Trans. Transformer	Cons.	Textual	38.3 ± .8	59.6 ± .9	30.9 ± .8
		Imagination	39.2 ± .8	59.7 ± .9	30.5 ± .9
		Multimodal (serial)	38.7 ± .8	60.8 ± .9	31.0 ± .8
	Unc.	Textual	40.4 ± .9	62.5 ± .8	32.3 ± .9
		Imagination	42.6 ± .8	62.8 ± .9	36.3 ± .9

State-of-the-Art Results

method	BLEU
Moses	32.5
RNN, text only	36.7
Transformer, text only	38.3
Original Imagination	40.2
Our unconstrained Imagination	42.6
WMT18 winner	45.1

Assessing the Representations

CCA & Image Retrieval

- mean-pool encoder states
- Canonical Correlation Analysis: find projection that is highly correlated with image representation
- representation fixed (do not propagate there)
- evaluate on image retrieval by captions (Recall @ 10)

$$W_t, W_v = \underset{W'_t, W'_v}{\text{argmax}}\, \text{corr}\left(W_t^\top \mathbf{T}, W_v^\top \mathbf{V}\right)$$

- **T** and **V** are sets of aligned representations
- W_t, W_v the learned projections

Semantic Textual Similarity

- SemEval sentence similarity task: manually annotated semantic similarity of sentences
- General domain, not related to image captions
- Sentence representation = mean-pooled encoder states
- Spearman correlation of representation of cosine distance of representations and semantic similarity

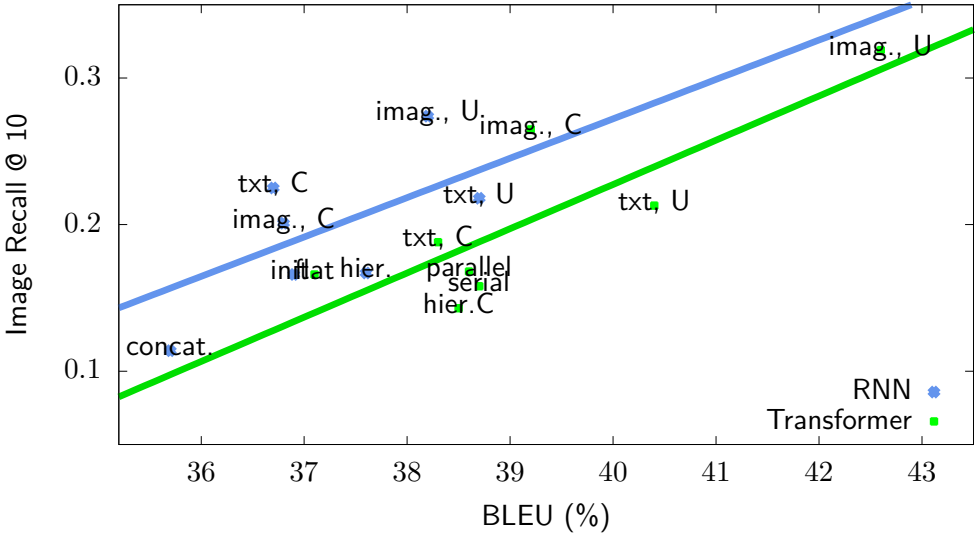
Results

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

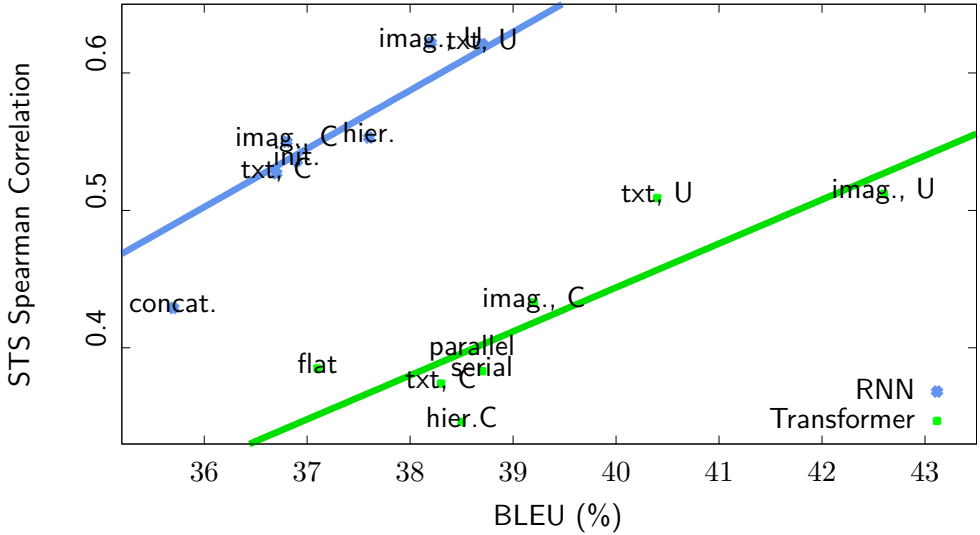
Textual MT		BLEU↑	Img.↑	STS↑
RNN	Textual	36.7	22.5	.527
	Textual U	38.7	21.8	.621
	Imagination	36.8	20.1	.550
	Imagination U	38.2	27.4	.622
Trans.	textual	38.3	18.8	.374
	textual U	40.4	21.3	.509
	Imagination	39.2	26.5	.433
	Imagination U	42.6	31.9	.512

Multimodal MT		BLEU↑	Img.↑	STS↑
RNN	Decoder init.	36.9	16.6	.536
	Att. concatenation	35.7	11.4	.429
	Flat att. comb.	34.6	14.6	.487
	Hierar. att. comb.	37.6	16.7	.553
Trans.	Serial att. comb.	38.7	15.8	.383
	Parallel att. comb.	38.6	16.8	.398
	Flat att. comb.	37.1	16.6	.385
	Hierar. att. comb.	38.5	14.3	.346

BLEU vs. Image Retrieval



BLEU vs. Semantic Similarity Performance



Conclusions

- Multimodal Translation Tasks was in fact domain adaptaion
- Multimodal information can help the translation
  - ... in a different way than we expected
- Multi-source models work
  - ... but do not improve MMT quality
- Target language and image provide a stronger training signal than LM