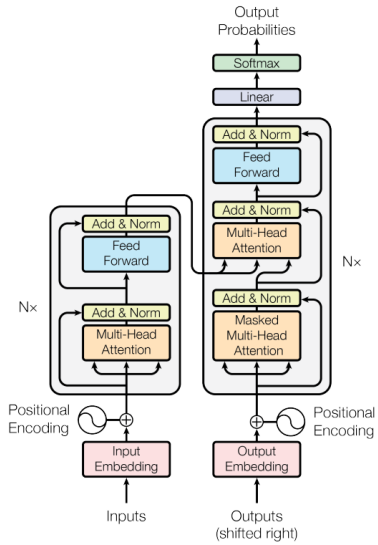
Transformer

Transformer for MT



Learning goals

- Understand the use of the Transformer

MACHINE TRANSLATION

- Sequence-to-sequence task
- Already served as a motivation for introducing the "ordinary" Attention-mechanism by Bahdanau et al. (2014)
- Crucial, that the decoder has access to the whole input sequence
  $\rightarrow$ This is very well solved by cross-attention
- Good contextualization in the encoder improves translation quality
  - (Bidirectional) RNNs/LSTMs are only (concatenated) unidirectional architectures
  - Transformer-Encoder layers are bidirectional by construction
  - Stacking them on top of each other makes this bidirectional contextualization even "deeper"

WMT 2014 EN-TO-DE AND EN-TO-FR

Parallel training data:

Parallel data:

File	Size	CS-EN	DE-EN	HI-EN	FR-EN	RU-EN	Notes
Europarl v7	628MB	✓	✓		✓		same as previous year, corpus home page
Common Crawl corpus	876MB	✓	✓		✓	✓	same as previous year
UN corpus	2.3GB				✓		same as previous year, corpus home page
News Commentary	77MB	✓	✓		✓	✓	updated, data with document boundaries
10^9French-English corpus	2.3 GB				✓		same as previous year [md5|sha1]
CzEng 1.0	115MB	✓					same as previous year, corpus home page (avoid sections 98 and 99)
Yandex 1M corpus	121MB					✓	corpus home page; v1.3 now in original case
Wiki Headlines	7.8MB			✓		✓	Provided by CMU. The ru-en is unchanged from last year.
HindEnCorp	25MB			✓			Collected by Charles University
~~The JHU Corpus~~				~~✓~~			This is fully contained in HindEnCorp, so not made available here.

THE BLEU SCORE

- Based on n-gram overlap from candidate and reference sentence
- Precision (for each n-gram) calculated as follows:

$$P_n = \frac{\sum_{C \in \{\text{ Candidates }\}} \sum_{n-gram \in C} \text{ Count }_{\text{clip}} (n- \text{ gram })}{\sum_{C' \in \{\text{ Candidates }\}} \sum_{n-gram' \in C'} \text{ Count } (n- \text{ gram }')}$$

- Finally, the BLEU score can be computed as

$$\texttt{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \cdot \log(P_n)\right),$$

- where BP is a "brevity penalty" to penalize short generations, N is the number of n-grams & w_n the weight for each P_n (usually $\frac{1}{N}$)

TRANSFORMER FOR MT

The Transformer ..

- .. outperforms the previous SOTA models
- .. at a lower number of required training FLOPs

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

VARYING THE TRANSFORMER ARCHITECTURE

Table 3: Variations on the Transformer architecture. Unlisted values are identical to those of the base model. All metrics are on the English-to-German translation development set, newstest2013. Listed perplexities are per-wordpiece, according to our byte-pair encoding, and should not be compared to per-word perplexities.

	N	d_{model}	d_{ff}	h	d_k	d_v	P_{drop}	ϵ_{ls}	train steps	PPL (dev)	BLEU (dev)	params $\times 10^6$
base	6	512	2048	8	64	64	0.1	0.1	100K	4.92	25.8	65
(A)				1	512	512				5.29	24.9	
				4	128	128				5.00	25.5	
				16	32	32				4.91	25.8	
				32	16	16				5.01	25.4	
(B)					16					5.16	25.1	58
					32					5.01	25.4	60
(C)	2									6.11	23.7	36
	4									5.19	25.3	50
	8									4.88	25.5	80
		256			32	32				5.75	24.5	28
		1024			128	128				4.66	26.0	168
			1024							5.12	25.4	53
			4096							4.75	26.2	90
(D)							0.0			5.77	24.6	
							0.2			4.95	25.5	
								0.0		4.67	25.3	
								0.2		5.47	25.7	
(E)	positional embedding instead of sinusoids									4.92	25.7	
big	6	1024	4096	16			0.3		300K	**4.33**	**26.4**	213