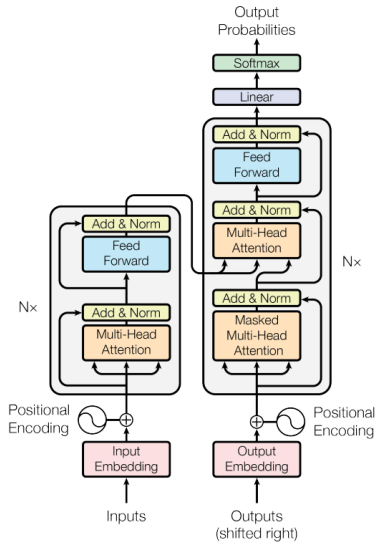
Transformer

Transformer for MT



Learning goals

- Understand the use of the Transformer

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

describe BLEU for context

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**	$2.3 \cdot 10^{19}$	

TRANSFORMER FOR MT

TRANSFORMER FOR MT