

Table 1: 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		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