Table 1: Replacing the position-wise feed-forward networks with multihead-attention-over-parameters produces similar results to the base model. All metrics are on the English-to-German translation development set, newstest2013.

			h_p	d_{pk}	d_{pv}	n_p	PPL (dev)	BLEU (dev)	params $\times 10^6$	training time
base	512	2048					4.92	25.8	65	12 hours
AOP_1	512		8	64	64	1536	4.92	25.5	65	16 hours
AOP_2	512		16	64	64	512	**4.86**	**25.9**	65	16 hours