BERT-base analysis: -

	Operation	Input_1	Size_1	Input_2	Size_2	Output	Size_O
1. Multi-Head Attention							
	a) Linear layer for Q,v,k	Encoder Input	S x H (512 x 768)	WQ	H x H (768 x 768)	Query (Q)	S x H (512 x 768)
		Encoder Input	S x H (512 x 768)	WK	H x H (768 x 768)	Key (K)	S x H (512 x 768)
		Encoder Input	S x H (512 x 768)	WV	H x H (768 x 768)	Value (V)	S x H (512 x 768)
	b) Scaled Dot-Product	Query (Q)	h x S x D_h (12 x 512 x 64)	K^T	h x D_h x S (12 x 64 x 512)	Attention Scores	h x S x S (12 x 512 x 512)
	c) SoftMax	Attention Scores	h x S x S (12 x 512 x 512)			Attention Weights	h x S x S (12 x 512 x 512)
	d) Weighted Sum	Attention Weights	h x S x S (12 x 512 x 512)	Value (V)	h x S x Dh (12 x 512 x 64)	Heads Output	h x S x Dh (12 x 512 x 64)
	e) Concat & Project	Heads Output	S x (h * Dh) (512 x 768)	Weight Matrix (WO)	H x H (768 x 768)	Attention Output	S x H (512 x 768)
2. Add & Norm	a) Residual Connection	Encoder Input	S x H (512 x 768)	Attention Output	S x H (512 x 768)	Sub-layer Sum	S x H (512 x 768)
	b) Layer Normalization	Sub-layer Sum	S x H (512 x 768)			Norm 1 Output	S x H (512 x 768)
3. Feed-Forward Network	a) Linear 1 + GELU	Norm 1 Output	S x H (512 x 768)	Weight Matrix (W1)	H x H_ff (768 x 3072)	Intermediate	S x H_ff (512 x 3072)
	b) Linear 2	Intermediate	S x H_ff (512 x 3072)	Weight Matrix (W2)	H_ff x H (3072 x 768)	FFN Output	S x H (512 x 768)
4. Add & Norm	a) Residual Connection	Norm 1 Output	S x H (512 x 768)	FFN Output	S x H (512 x 768)	Sub-layer Sum 2	S x H (512 x 768)
	b) Layer Normalization	Sub-layer Sum 2	S x H (512 x 768)			Encoder Layer Output	S x H (512 x 768)

Max Sequence Length (S): 512, **Embedding Dimension (H)**: 768.

Number of Attention Heads (h): 12, **Attention Head Size (D_h)**: 64, **Feed-Forward Intermediate Size (H_ff)**: 3072 (4∗H)

For BERT large: - H → 1024, h→ 16, D_h→ 64, h_ff → 4096.