Attention and sequential structure

Position Embeddings

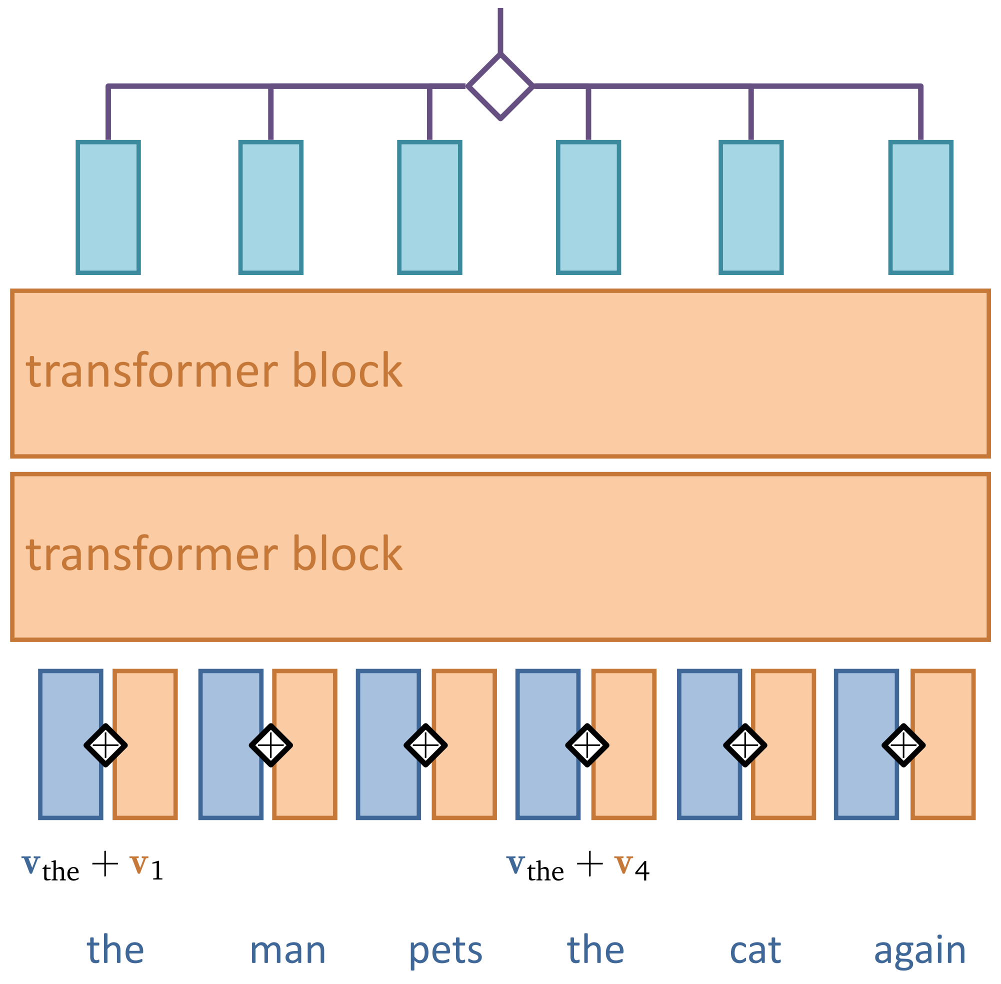
Conceptually simple

Easy to implement

Adds set of learnable parameters

Maximum context length at test time limited to max sequence length in training set

Embedding quality diminishes (in theory) with *t*

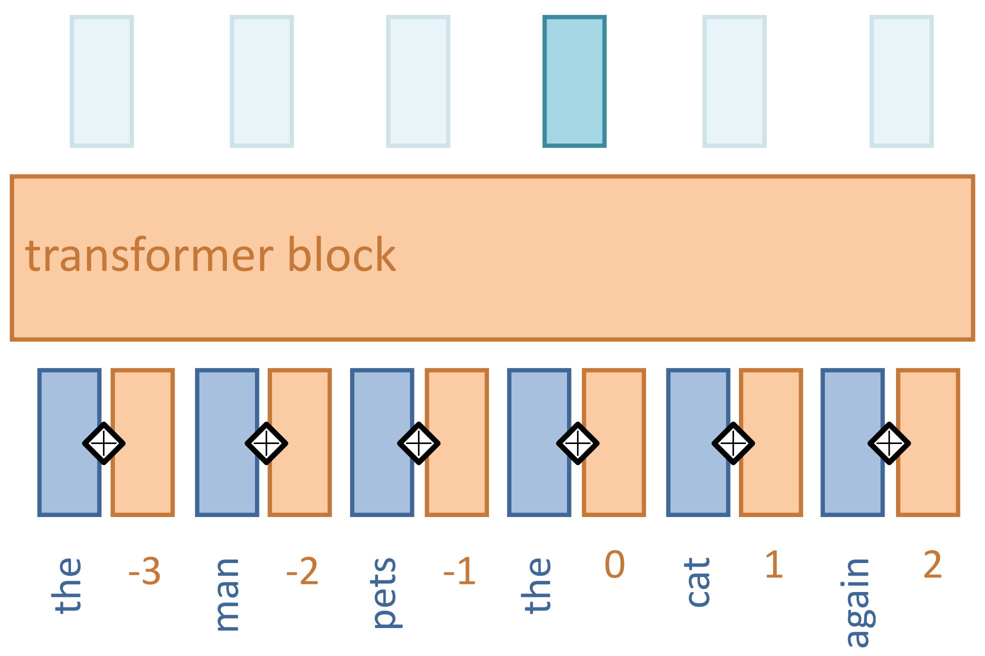


Relative position embeddings

Captures the property of absolute position invariance which is desirable

Like absolute position embeddings, limited in context length at inference

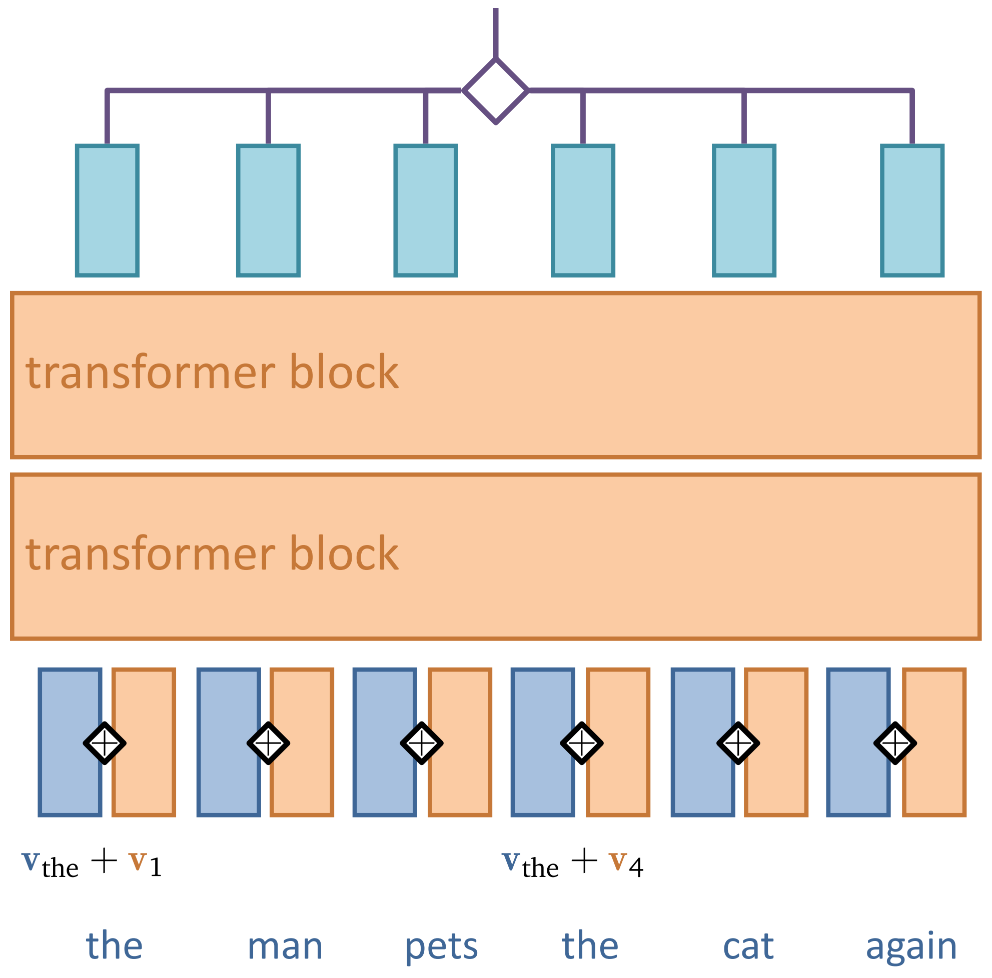
Conceptually and practically more complex; on the surface this results in (T-1)^2 different inputs because each input will have T-1 separate representations (there is a hack that turns this into 2T-1)

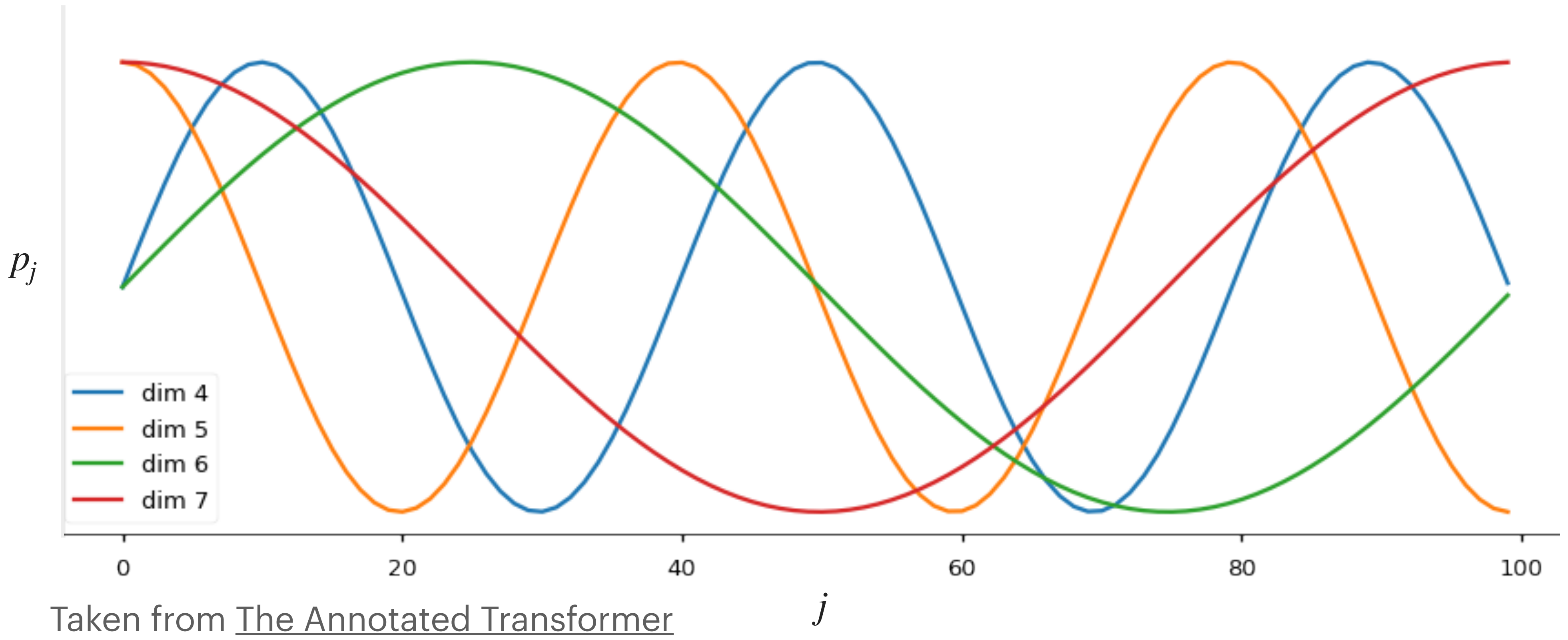


Position encodings

Captures the property of absolute position invariance which is desirable

Conceptually and practically more complex; on the surface this results in (T-1)^2 different inputs because each input will have T-1 separate representations (there is a hack that turns this into 2T-1)





Taken from The Annotated Transformer

Embedding illustrations taken from Bloem (2020)

BERT (2018)

BERT

The most popular/used/studied Transformer model to date

30K citations [1]

Trained on Wikipedia and Book Corpus (~10K books)

Wordpiece embedding [2]

Non-causal, uses masked LM procedure for pretraining

340 M parameters in total

LM prediction head: 4 ReLU (D=4096)

LM training took 4 days on 64 TPU cores

$$D = 1024$$

$$N_H = 16$$

$$N_B = 24$$

$$T_{max} = 1024$$

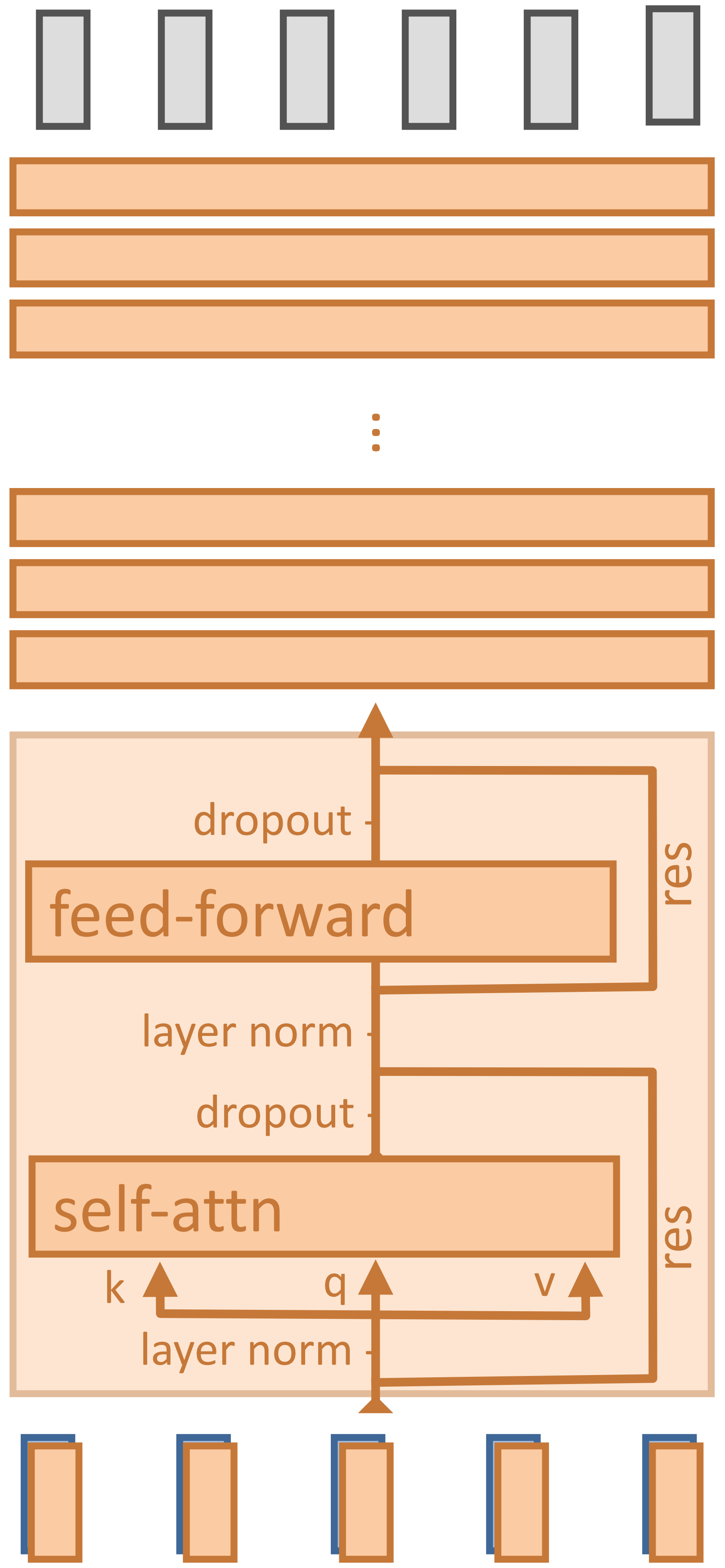
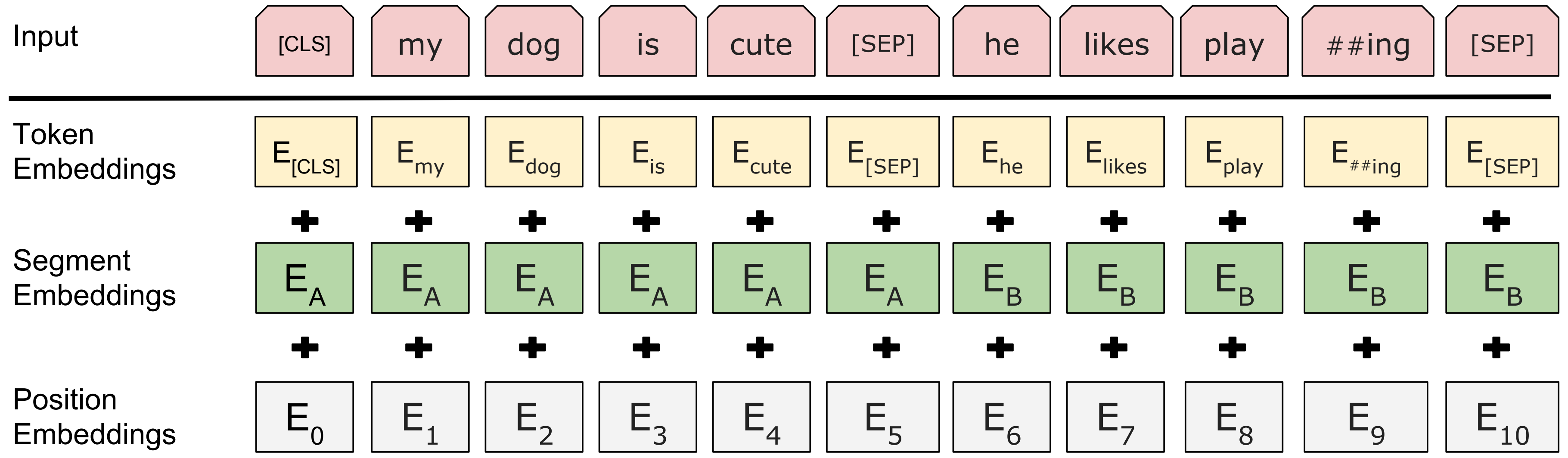


Figure taken from Bloem, 2020



BERT Input Features [1]

Hyperparams			Dev Set Accuracy			
#L	#H	#A	LM (ppl)	MNLI-m	MRPC	SST-2
3	768	12	5.84	77.9	79.8	88.4
6	768	3	5.24	80.6	82.2	90.7
6	768	12	4.68	81.9	84.8	91.3
12	768	12	3.99	84.4	86.7	92.9
12	1024	16	3.54	85.7	86.9	93.3
24	1024	16	3.23	86.6	87.8	93.7

Ablation Study [1]

[1] Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)*, [2] Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016)*