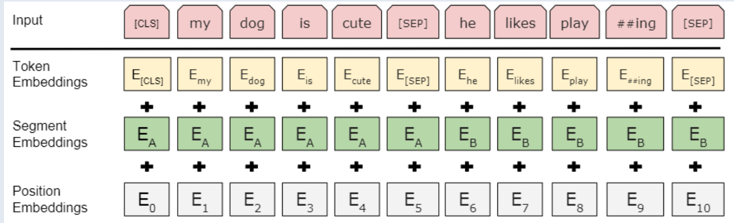
BERT

- **Model Architecture**
  - multi-layer bidirectional Transformer en-coder
  - pre-training and fine-tuning
- **Input/Output Representations**
  - a single sentence or a pair of sentences(e.g., 〈 Question, Answer 〉) in one token
    - Sentence pairs are packed together into a single sequence, separate them with a special token ([SEP])
  - The first token of every sequence: [CLS]
    - add a learned embedding to every token indicating whether it belongs to sentence A or sentence B



Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

- **pre-train BERT using two unsupervised tasks**
  - chooses 15% Masked: we replace the i-th token with (1) the [MASK] token 80% of the time (2) a random token 10% of the time (3) the unchanged i-th token 10% of the time.
    - Then, Ti will be used to predict the original token with cross entropy loss.
      - 这是因为预训练的时候用到了很多MASK，但是微调（网络参数）的时候不会出现MASK。
  - <A,B> : 50% B is the actual next sentence that follows A (labeled as IsNext), and 50% it is a random sentence from the corpus (labeled as NotNext)
- **Fine-tuning BERT**