Bert For Sequence Classification

Size([2, 2])

Prediction

Size([2, 768])

Dropout Layer p=0.1

Size([2, 768])

Linear Layer

Bert Model

Size([2, 768])

Bert Pooler

Tanh Activation Function

Size([2, 768])

Linear Layer

Size([2, 768])

Keep First Token [CLS]

Bert Encoder

Size([2, 9, 768])

X 12 Bert Layers

Bert Layer

Bert Output

Layer Normalization

Size([2, 9, 768])

＋

Size([2, 9, 768])

Size([2, 9, 768])

Dropout Layer p=0.1

Size([2, 9, 768])

Linear Layer

Size([2, 9, 3072])

Bert Intermediate

Gelu Activation Function

Size([2, 9, 3072])

Linear Layer

Size([2, 9, 768])

Size([2, 9, 768])

Bert Attention

Size([2, 9, 768])

Bert Self Output

Layer Normalization

Size([2, 9, 768])

＋

Size([2, 9, 768])

Size([2, 9, 768])

Dropout Layer p=0.1

Size([2, 9, 768])

Linear Layer

Size([2, 9, 768])

Bert Self Attention

R

Size([2, 9, 12, 64])

P

Size([2, 12, 9, 64])

•

Size([2, 12, 9, 9])

Dropout Layer p=0.1

Attention Scores Size([2, 12, 9, 9])

Softmax Layer

Size([2, 12, 9, 9])

sqrt(64 attention head size) —8.0→ /

Size([2, 12, 9, 9])

•   Size([2, 12, 64, 9])   T

Size([2, 12, 9, 64])   Size([2, 12, 9, 64])   Size([2, 12, 9, 64])

Reshape Tensors
Query
768 size = 12 attention heads
x 64 attention heads size

Reshape Tensors
Key
768 size = 12 attention heads
x 64 attention heads size

Reshape Tensors
Value
768 size = 12 attention heads
x 64 attention heads size

Size([2, 9, 768])   Size([2, 9, 768])   Size([2, 9, 768])

Linear Layer   Linear Layer   Linear Layer

Size([2, 9, 768])   Size([2, 9, 768])   Size([2, 9, 768])

Size([2, 9, 768])

Size([2, 9, 768])

Bert Embeddings

Dropout Layer p=0.1

Size([2, 9, 768])

Layer Normalization

Size([2, 9, 768])

＋

Size([2, 9, 768])   Size([1, 9, 768])   Size([2, 9, 768])

Word Embeddings Layer   Position Embeddings Layer   Tokens Types Embeddings Layer

Size([1, 9])

Created Tokens
Positions IDs

Size([2, 9])   Size([2, 9])

Inputs

Bert Tokenizer Output

Tokens
IDs

Tokens
Types IDs

token_type_ids:
[[0, 0, 0, 0, 0, 0, 0, 0, 0],
[0, 0, 0, 0, 0, 0, 0, 0, 0]]
Size:
[2,9]

Bert Tokenizer

attention_mask:
[[1, 1, 1, 1, 1, 1, 0, 0, 0],
[1, 1, 1, 1, 1, 1, 1, 1, 1]]
Size:
[2,9]

I love cats!
He hates pineapple pizza.

input_ids:
[[101, 146, 1567, 11771, 106, 102, 0, 0, 0],
[101, 1124, 18457, 10194, 11478, 7136, 13473, 119, 102]]
Size:
[2,9]

Processed Text:
[CLS] I love cats! [SEP] [PAD] [PAD] [PAD]
[CLS] He hates pineapple pizza. [SEP]

Notations:

＋   Adding Tensors.

T   Transpose Tensor.

•   Matrix product of two tensors.

/   Divide tensor by scalar value.

P   Permute Tensor.

R   Reshape Tensor.