…

…

**768** input neurons

**50** hidden neurons

**3** output neurons

**Feed-Forward Neural Network**

**BERT fine-tuning**

**BERT base uncased**

• **12** layers of encoders

• **12** attention heads

• **768** neurons in the hidden layers

**110M** parameters

Output of **MLM**

(not used in the classification task)

Output of **NSP**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ |  | ↑ |
| h[CLS] | hMy | hdog | his | h[MASK] | hand | hlikes | hplaying | h[SEP] | h[PAD] | h[PAD] | … | h[PAD] |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| E[CLS] | EMy | Edog | Eis | E[MASK] | Eand | Elikes | Eplaying | E[SEP] | E[PAD] | E[PAD] | … | E[PAD] |
| ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ |  | ↑ |
| 101 | 2026 | 3899 | 2003 | 103 | 1998 | 7777 | 2652 | 102 | 0 | 0 | … | 0 |
| ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ |  | ↑ |
| [CLS] | My | dog | is | [MASK] | and | likes | playing | [SEP] | [PAD] | [PAD] | … | [PAD] |

Input Embeddings

**Ei** of dimension **768**

Tokens

Tokens IDs

Hidden states

**hi** of dimension **768**

Fixed length of maximum **512** tokens

**BERT pre-trained model**