Using the Transformer

BERT – Pre-training

Learning goals

- Understand the two pre-training tasks
- Learn how samples are constructed
- Understand the pre-training process

MASKED LANGUAGE MODELING (MLM)

First remark:

- It has nothing to do with Masked Self-Attenion
  $\rightarrow$ Masked Self-Attention is an architectural detail in the decoder of the Transformer, i.e. used by e.g. GPT
- Masked Self-Attention as a way to prevent causality issues in a Transformer decoder
- MLM is a self-supervised *modeling objective* introduced to couple Self-Attention and (deep) bidirectionality without violating causality

MASKED LANGUAGE MODELING (MLM) CTD.

- **Training objective:**

  Given a sentence, predict `[MASK]`ed tokens

- **Generation of samples:**

  Randomly replace* a fraction of the words by `[MASK]`

  *Sample 15% of the tokens; replace 80% of them by `[MASK]`, 10% by a random token & leave 10% unchanged

- **Input:**

The	quick	brown	`[MASK]`	jumps	over	the	`[MASK]`	dog	.

- **Targets:**

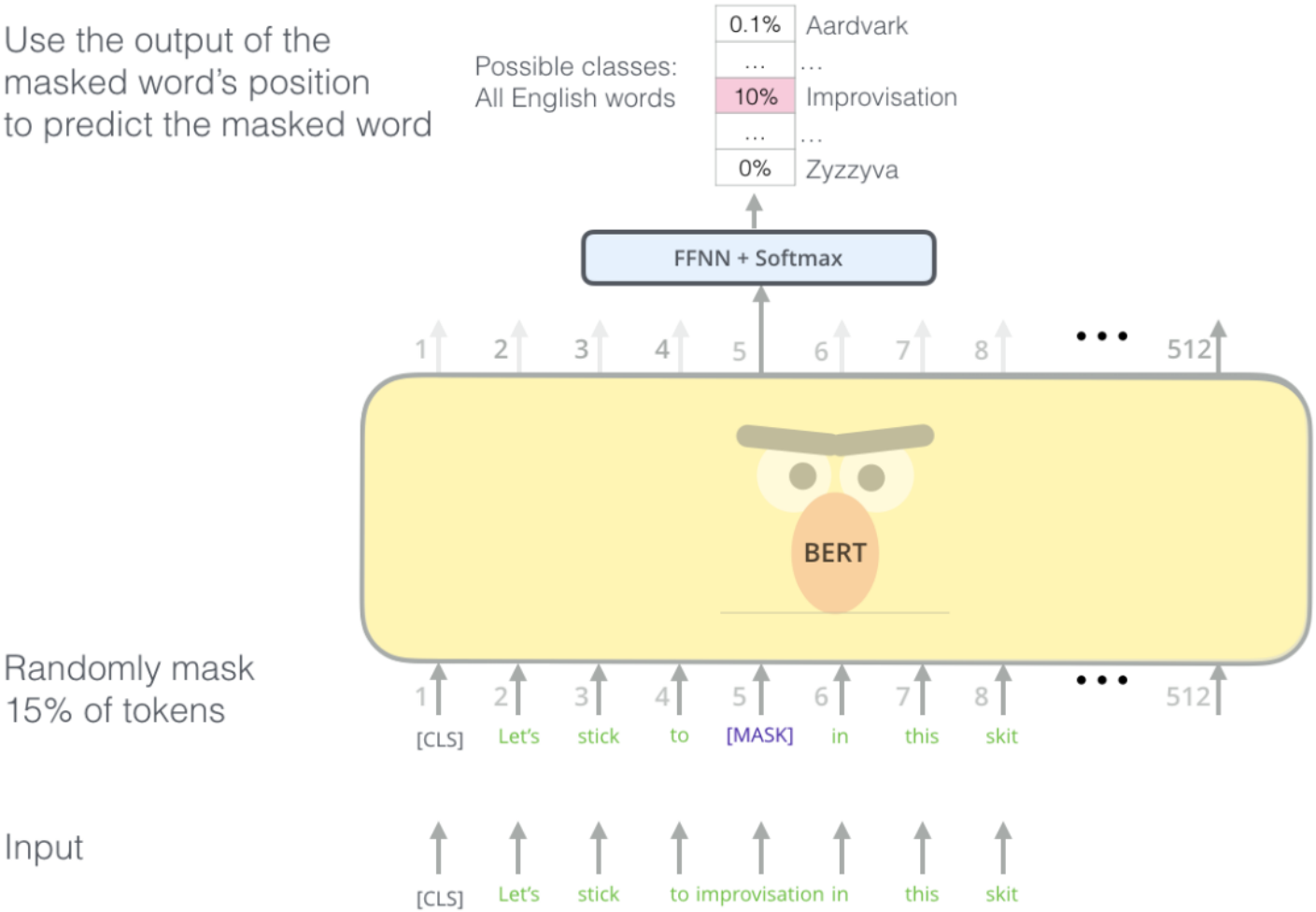
  (*fox*, *lazy*)

MASKED LANGUAGE MODELING (MLM) CTD.

Discrepancy between pre-training & fine-tuning:

- `[MASK]`-token as central part of pre-training procedure
- `[MASK]`-token does not occur during fine-tuning
- **Modified pre-training task:**
  Predict 15% of the tokens of which only 80% have been replaced by `[MASK]`
  - 80% of the selected tokens:
    `The quick brown fox` $\rightarrow$ `The quick brown [MASK]`
  - 10% of the selected tokens:
    `The quick brown fox` $\rightarrow$ `The quick brown went`
  - 10% of the selected tokens:
    `The quick brown fox` $\rightarrow$ `The quick brown fox`

MASKED LANGUAGE MODELING (MLM) CTD.



Source: *Jay Alammar*

NEXT SENTENCE PREDICTION (NSP)

- **Training objective:**

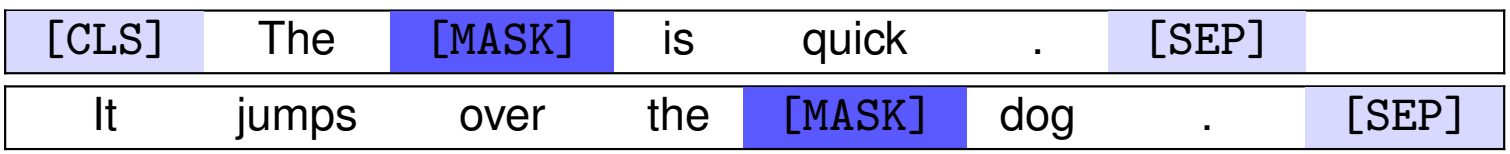
  Given two sentences, predict whether s_2 follows s_1

- **Generation of samples:**

  Randomly sample* negative examples (cf. word2vec)

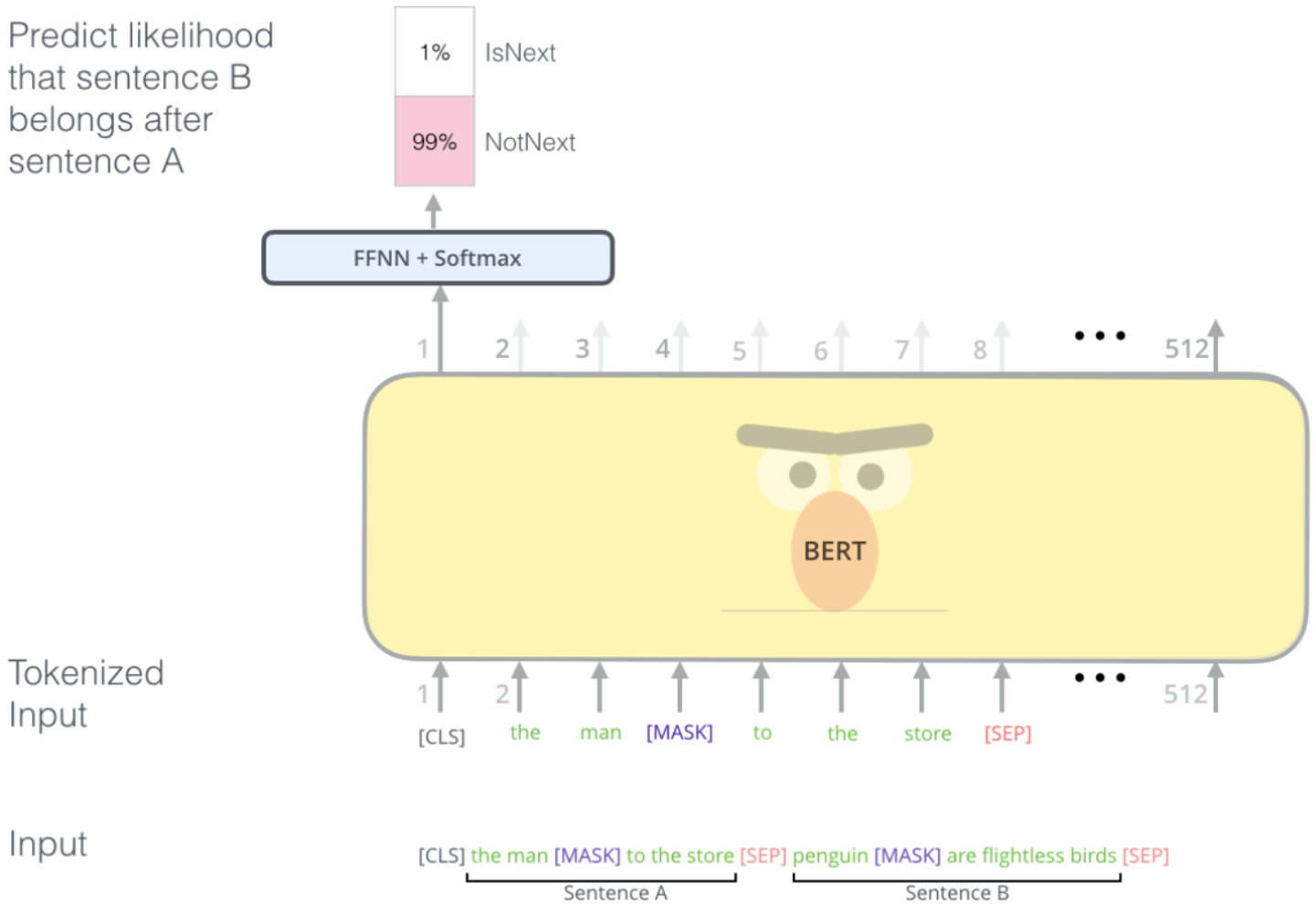
  *50% of the time the second sentence is the actual next sentence, 50% of the time it is a randomly sampled sentence

- **Full Input:**



- [CLS] token as sequence representation for classification
- [SEP] token for separation of the two input sequences

NEXT SENTENCE PREDICTION (NSP) CTD.



Source: *Jay Alammar*

PRE-TRAINING BERT

Ingredients:

- Massive lexical resources (BooksCorpus + Eng. Wikipedia)
  $\rightarrow$ 13 GB in total
- Train for approximately* 40 epochs
- 4 (16) Cloud TPUs for 4 days for the BASE (LARGE) variant
- Loss function:

$$Loss_{BERT} = Loss_{MLM} + Loss_{NSP}$$

*1.000.000 steps on batches of 256 sequences with a sequence length of 512 tokens

- For their experiments:
  - Pre-train w/ sequence length 128 for 90% of the steps
  - Pre-train w/ sequence length 512 for 10% of the steps
    (Reason: Learn positional embeddings for all positions)