Using the Transformer

RoBERTa (Liu et al., 2019)

Learning goals

- Understand the improvements over BERT
- Dynamic Masking

ROBERTA

Improvements in Pre-Training:

- Authors claim that BERT is seriously undertrained
- Change of the `MASK`ing strategy
  $\rightarrow$ BERT masks the sequences once before pre-training
  $\rightarrow$ RoBERTa uses dynamic `MASK`ing
  $\Rightarrow$ RoBERTa sees the same sequence `MASK`ed differently
- RoBERTa does not use the additional NSP objective during pre-training
- 160 GB of pre-training resources instead of 13 GB
- Pre-training is performed with larger batch sizes (8k)

DYNAMIC VS. STATIC MASKING ▸ LIU ET AL., 2019

Static Masking (BERT):

- Apply `MASK`ing procedure to pre-training corpus once
- (additional for BERT: Modify the corpus for NSP)
- Train for approximately 40 epochs

Dynamic Masking (RoBERTa):

- Duplicate the training corpus *ten* times
- Apply `MASK`ing procedure to each duplicate of the pre-training corpus
- Train for 40 epochs
- Model sees each training instance in ten different "versions" (each version four times) during pre-training

ROBERTA ‣ LIU ET AL., 2019

Architectural differences:

- Architecture (layers, heads, embedding size) identical to BERT
- 50k token BPE vocabulary instead of 30k
- Model size differs (due to the larger embedding matrix) $\Rightarrow \sim$ 125M (360M) for the BASE (LARGE) variant

Performance differences:

	MNLI	QNLI	QQP	RTE	SST	MRPC	CoLA	STS	WNLI	Avg
Single-task single models on dev										
$\text{BERT}_{\text{LARGE}}$	86.6/-	92.3	91.3	70.4	93.2	88.0	60.6	90.0	-	-
$\text{XLNet}_{\text{LARGE}}$	89.8/-	93.9	91.8	83.8	95.6	89.2	63.6	91.8	-	-
RoBERTa	**90.2/90.2**	**94.7**	**92.2**	**86.6**	**96.4**	**90.9**	**68.0**	**92.4**	**91.3**	-

Source: Liu et al. (2019)

Note: Liu et al. (2019) report the accuracy for QQP while Devlin et al. (2018) report the F1 score (cf. results displayed in chapter 6.2.3); XLNet: see next Chapter.