T5 - UNIFIED TEXT-TO-TEXT TRANSFER TRANSFORMER

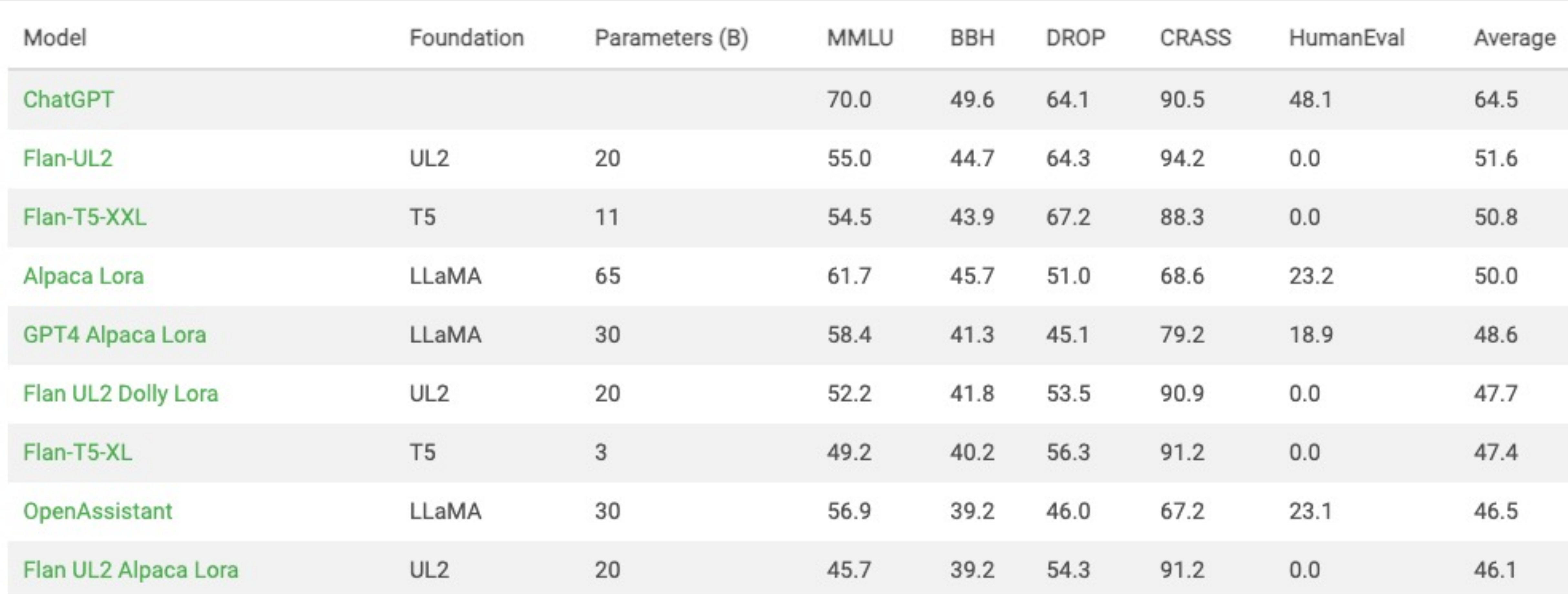
Houston Machine Learning-LLM Reading Group
Yan Xu

What is T5

- **T5: T**ext-**t**o-**T**ext **T**ransfer **T**ransformer
  - Use the complete encoder-decoder
  - Pretrained with Cleaned dataset: Colossal Clean Crawled Corpus (C4)

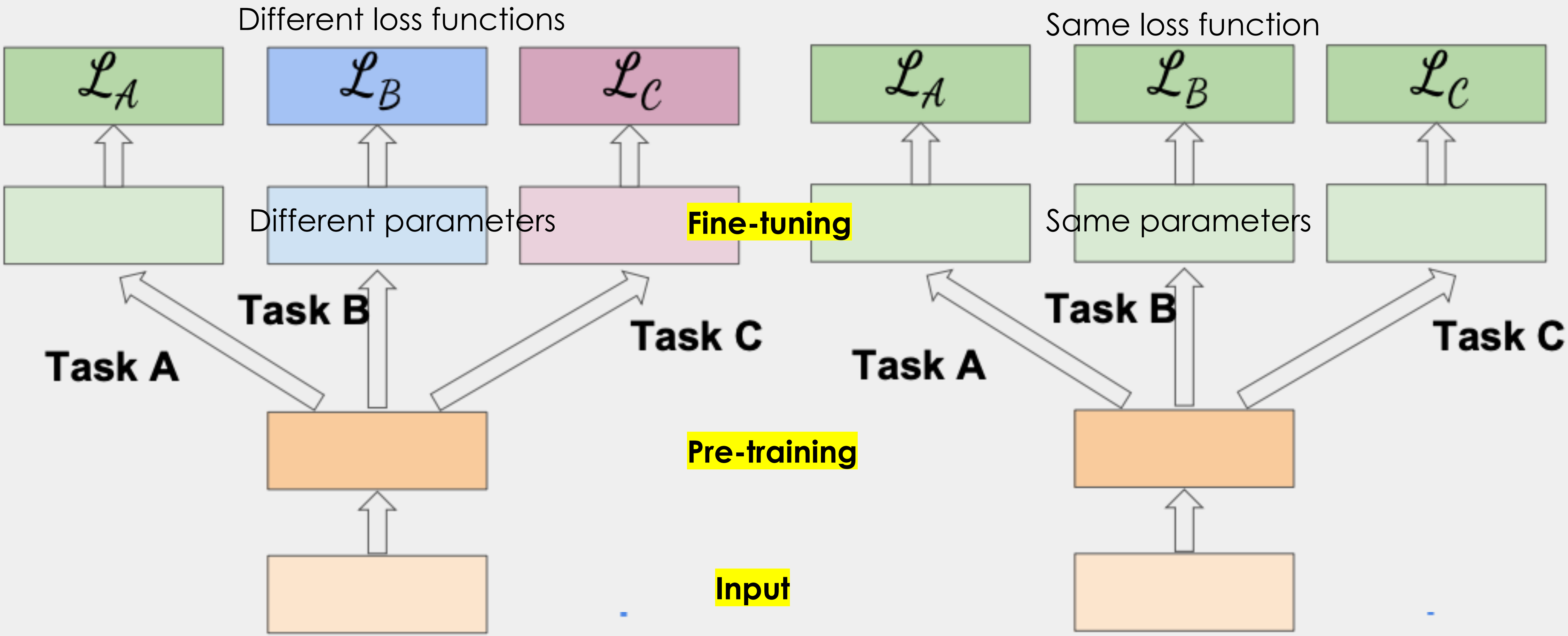
LLM based on T5

https://declare-lab.net/instruct-eval/



Model	Foundation	Parameters (B)	MMLU	BBH	DROP	CRASS	HumanEval	Average
ChatGPT			70.0	49.6	64.1	90.5	48.1	64.5
Flan-UL2	UL2	20	55.0	44.7	64.3	94.2	0.0	51.6
Flan-T5-XXL	T5	11	54.5	43.9	67.2	88.3	0.0	50.8
Alpaca Lora	LLaMA	65	61.7	45.7	51.0	68.6	23.2	50.0
GPT4 Alpaca Lora	LLaMA	30	58.4	41.3	45.1	79.2	18.9	48.6
Flan UL2 Dolly Lora	UL2	20	52.2	41.8	53.5	90.9	0.0	47.7
Flan-T5-XL	T5	3	49.2	40.2	56.3	91.2	0.0	47.4
OpenAssistant	LLaMA	30	56.9	39.2	46.0	67.2	23.1	46.5
Flan UL2 Alpaca Lora	UL2	20	45.7	39.2	54.3	91.2	0.0	46.1

Multi-task transfer learning



Classical Paradigm **T5 Paradigm**

Prerequisites

Step by Step into Transformer
Step by Step into BERT

By medium.com/@YanAIx

Setup: Pre-training Dataset

- Goal: analyze the effect of the quality, characteristics and size of unlabeled data
- Source: https://commoncrawl.org/ (20 TB/month, noisy data)
- Data cleaning using heuristics
  - Only retain lines ending in a terminal punctuation mark (“.”, “!”, “?” etc.)
  - Remove obscene words
  - Removing pages containing Javascript code
  - Remove duplicate sentences
  - Retain only English webpages
  - 750 GB

Setup: Fine-tuning tasks

- Text classification: GLUE and SuperGLUE
- Abstractive summarization: CNN/Daily Mail
- Question Answering: SQuAD
- Translation: WMT English to German, French, and Romanian

Setup: Input & Output

- “text-to-text” format
  - Preprocessed Examples in Appendix D in T5 paper
- consistent training objective: maximum likelihood
- task-specific (text) prefix
- Mismatch label Issue
  - e.g. given a premise and hypothesis, classify into one of 3 categories - ‘entailment’, ‘contradiction’ and ‘neutral’
  - Potentially possible for decoder to output ‘hamburger’
  - This issue never observed with their trained models

T5: Exploring the Limits of Transfer Learning with a unified Text-to-Text Transformer

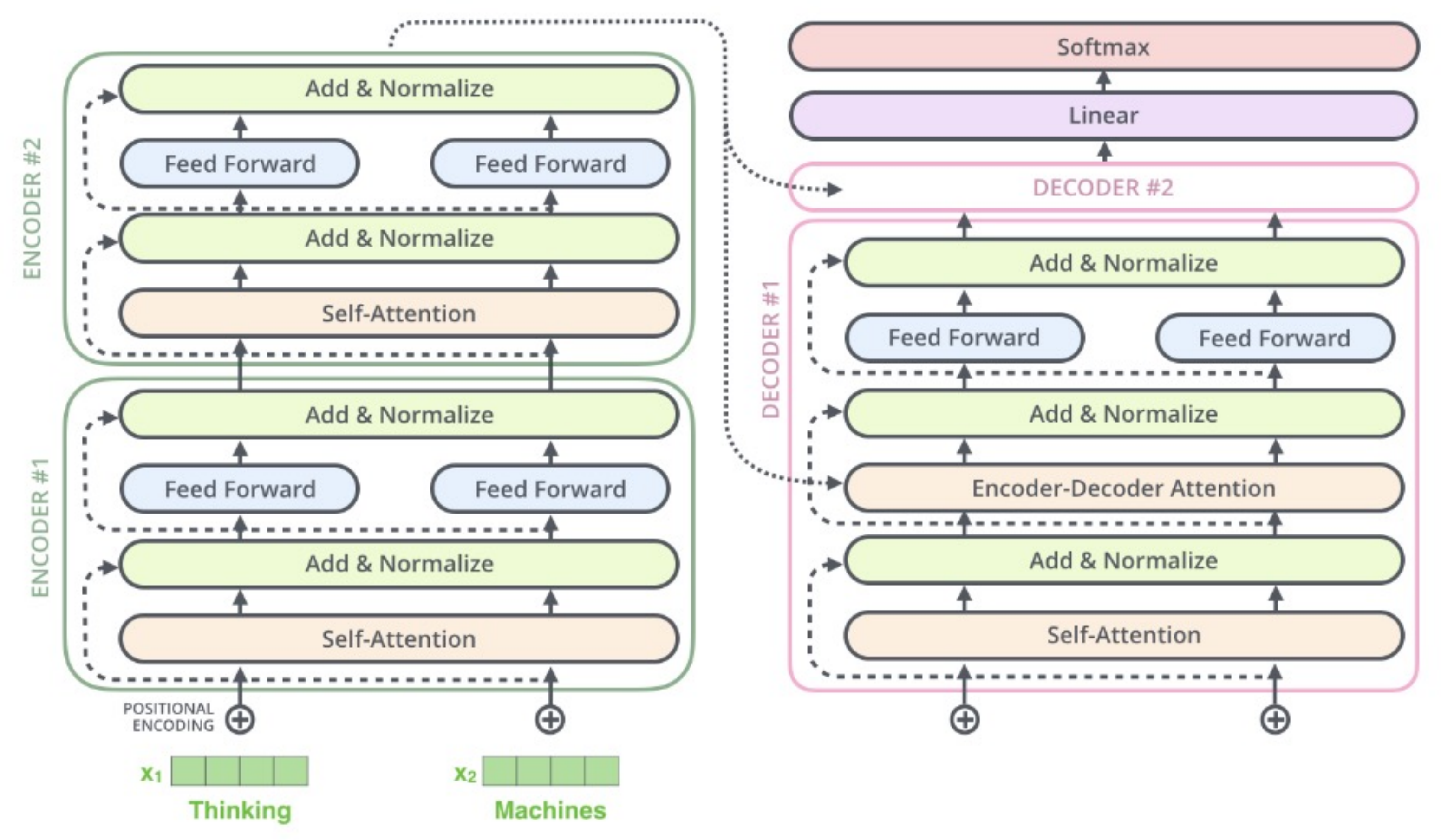
Empirical Survey

Baseline → Architecture → Objective →

→ Dataset → Transfer Approach → Scaling

Baseline

- Encoder-Decoder architecture as in original Transformer paper (Vaswani et al., 2017)



- Relative Positional self-attention (Shaw et al., 2018)
- removing the Layer Norm bias

Baseline

- Relative Positional self-attention with edge representations

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V + \boxed{a_{ij}^V})$$

output

Weighted average of values

$$a_{ij}^K = w_{\mathrm{clip}(j-i,k)}^K$$

$$a_{ij}^V = w_{\mathrm{clip}(j-i,k)}^V$$

$$\mathrm{clip}(x,k) = \max(-k, \min(k,x))$$

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{n} \exp e_{ik}}$$

Attention weights

$$w^K = (w_{-k}^K, \ldots, w_k^K) \text{ and } w^V = (w_{-k}^V, \ldots, w_k^V)$$

Learnable edge keys and edge values

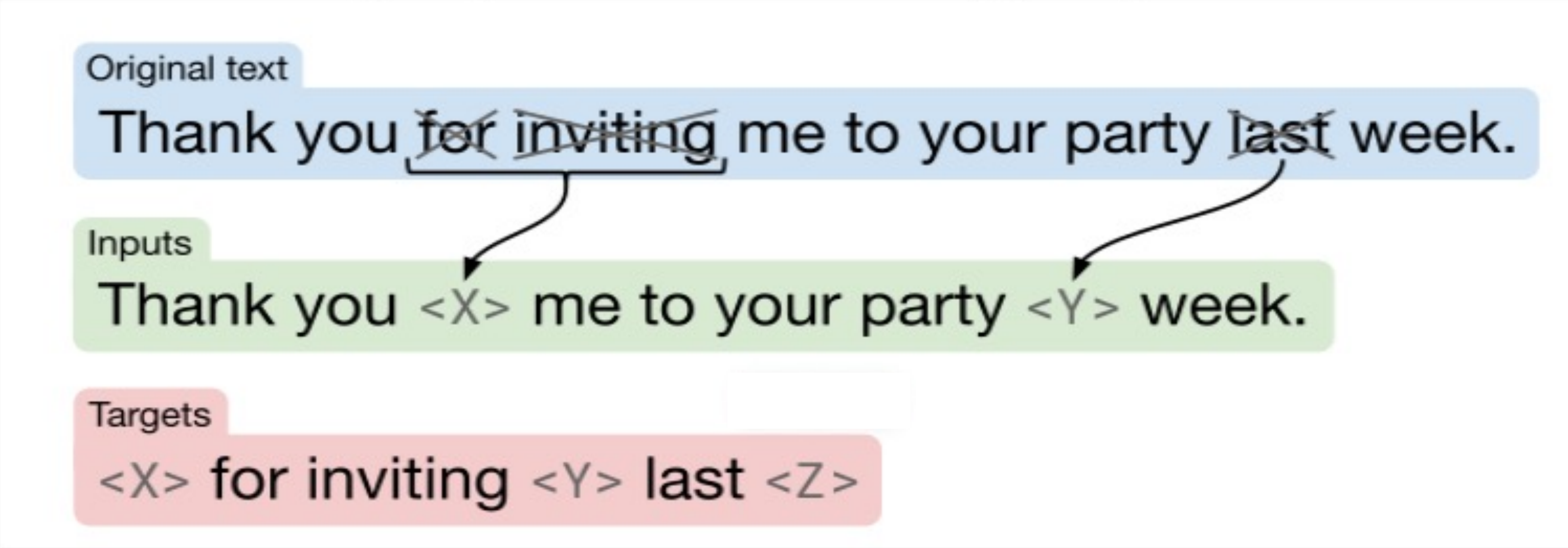
Queries

Keys

$$e_{ij} = \frac{x_i W^Q (x_j W^K + \boxed{a_{ij}^K})^T}{\sqrt{d_z}}$$

Baseline

- Pre-training objective: Denoising(drop 15 % tokens randomly)



Original text

Thank you ~~for inviting~~ me to your party ~~last~~ week.

Inputs

Thank you <X> me to your party <Y> week.

Targets

<X> for inviting <Y> last <Z>

- BERT-base Size Encoder and Decoder (L=12, H=768, A=12)
- Multilingual Vocabulary: SentencePiece (32k word pieces)

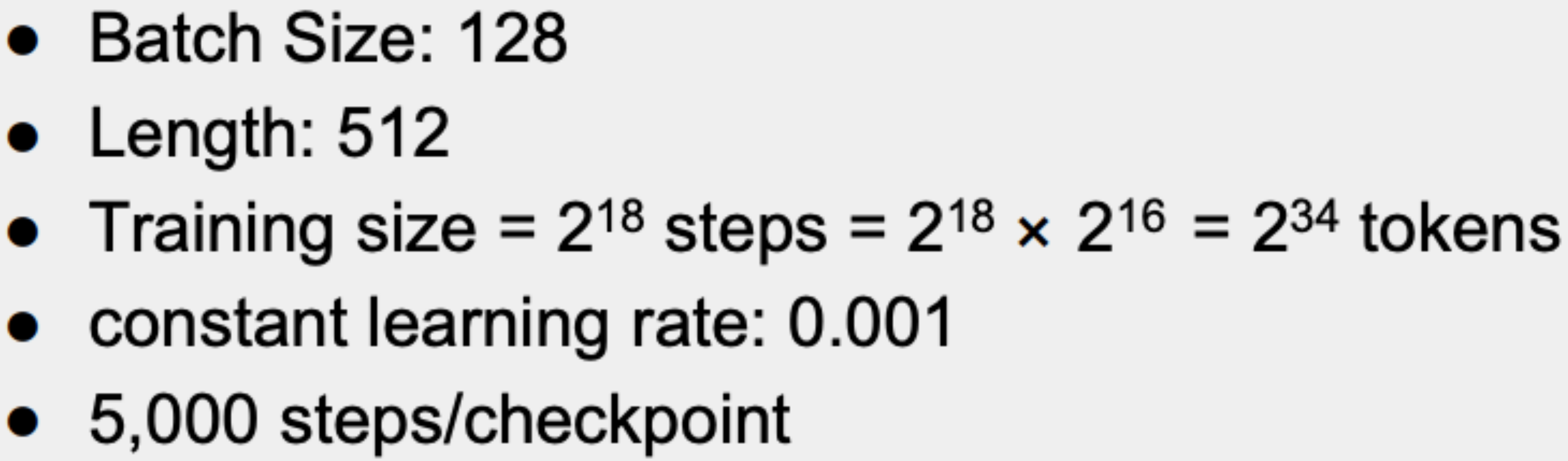
L: number of transformer blocks; H: hidden size; A: the number of self-attention heads

Baseline(Pre-training details)

- Max Sequence length: 512 tokens
- Batch size: 128 sequences = 128 × 512 = 2^{16} tokens
- Training size = 2^{19} steps = $2^{19} \times 2^{16} = 2^{35}$ tokens ≈ 34 B tokens << BERT (137B) << RoBERTa (2.2T)
- inverse square root learning rate schedule, where k = 10^4 (warm-up steps)

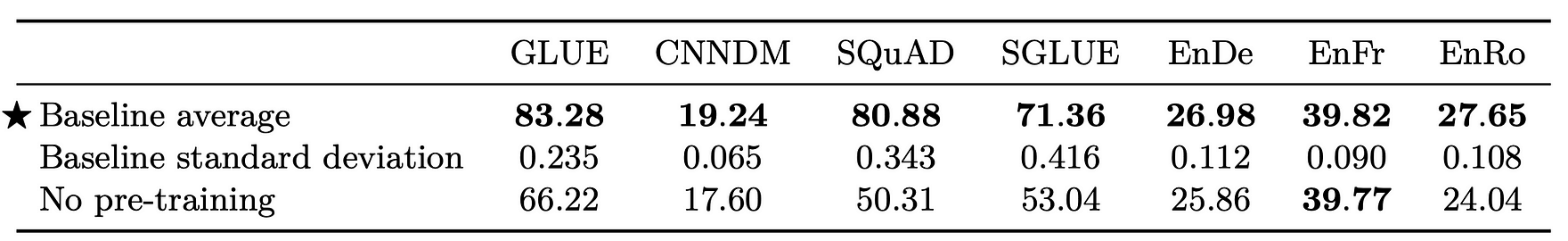
  $1/\sqrt{\max(n, k)}$, constant learning rate for the first k step then decays the learning rate until pre-training is over.
- AdaFactor
- Dropout: 0.1

Baseline(Fine-tuning Details)



- Batch Size: 128
- Length: 512
- Training size = 2^{18} steps = $2^{18} \times 2^{16}$ = 2^{34} tokens
- constant learning rate: 0.001
- 5,000 steps/checkpoint

Baseline Performance



	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	**83.28**	**19.24**	**80.88**	**71.36**	**26.98**	**39.82**	**27.65**
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	**39.77**	24.04

Average and standard deviation of scores achieved by our baseline model and training procedure vs training on each task from scratch without any pretraining.