Exploring the limits of Transfer learning with a Unified Text-to-Text Transformer

2.1. Model

- roughly equivalent to the original Transformer
- removing the Layer Norm bias
- placing the layer normalization outside the residual path
- using different position embedding scheme.

2.4. Input & Output Format

- "text-to-text" format
- provides a consistent training objective both for pre-training & fine-tuning

3.1. Baseline

- to reflect modern, typical practice
- pre-train a standard Transformer using a simple denoising objective
- separately fine-tune on each of downstream task.

3.1.2 Training

- "inverse square root" learning rate schedule

$$1 / \sqrt{\max(n, k)}$$

(n: current training iteration; k: number of warm-up steps (set to 10^4 in all experiments))

  - constant learning rate of 0.01 for the first 10^4 steps
  - then exponentially decays the learning rate until pre-training is over.

3.1.4. Unsupervised Objective

- mask consecutive spans of tokens & only predict dropped-out tokens
  → reduce computational cost of pre-training

3.1.5. Baseline Performance

- pre-training provides significant gains across almost all benchmarks
- inter-run variance
  - GLUE & SuperGLUE : average of scores of each benchmark
  - CoLA, CB, COPA : ↑ inter-run variance...
  - harder to compare models using GLUE & SuperGLUE scores alone.