Amazon
Global Hiring Science

Yizhen Egyn Zhu, PhD
Dawn Sepehr, PhD
04/20/2023

amazon

GLOBAL
HIRING SCIENCE

Agenda

- EDA & Data Processing
- Model Setup
- Alternative Approaches
  - What we have tried
  - What we haven’t tried

Explorative Data Analysis

- Descriptives of ratings
- Correlation matrix for all ratings
- Descriptives of text exercise
  - How many exercises did each candidate complete?
  - How many candidates completed each exercise?
  - What is the average length of text per candidate?

Data Processing

- Concatenate all text columns into one piece
  - Excluding the “final exercise”
- Rescale all ratings to 1 - 4

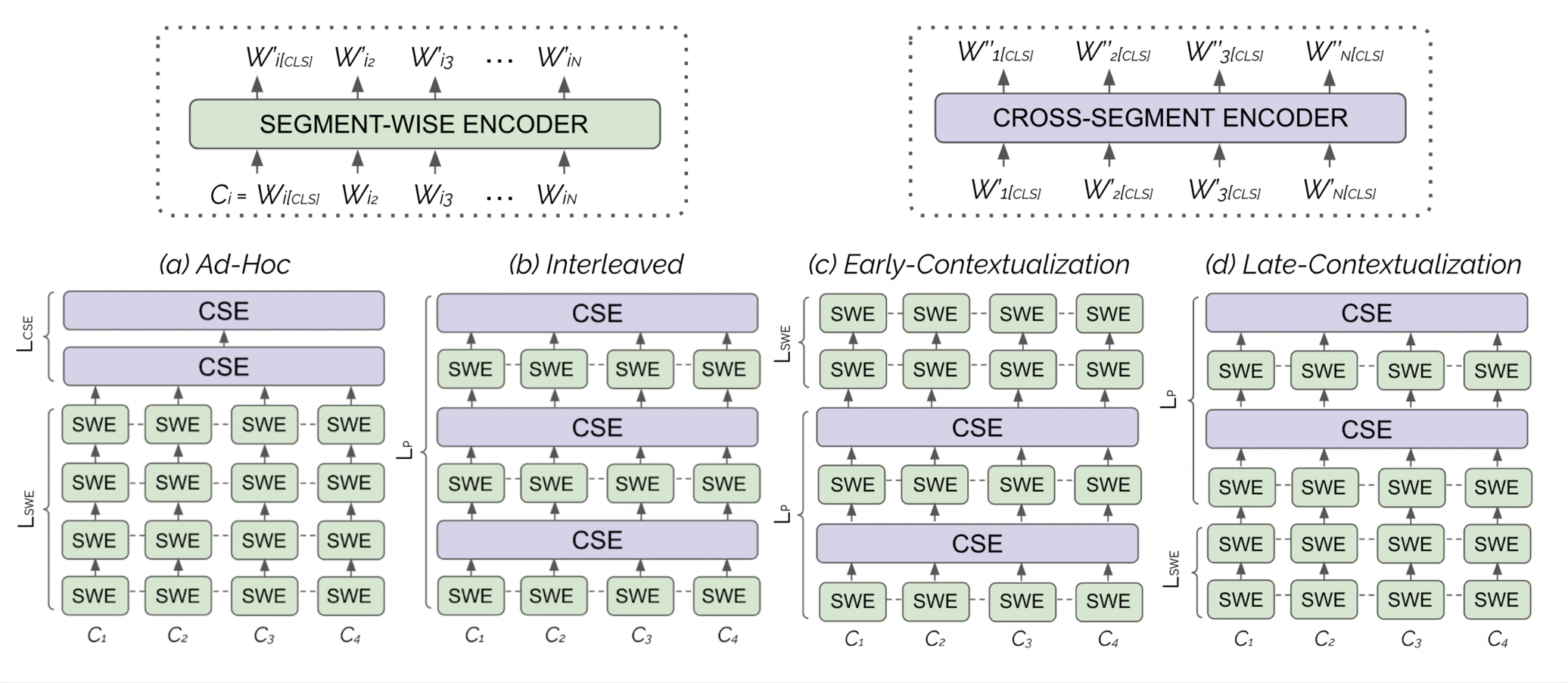
Model Setup

Model Architecture

- Hierarchical Attention Transformer (HAT; Chalkidis et al., 2022)
  - Segment-wise encoder
    - Contextualizes token-level representations per segment
  - Cross-segment encoder
    - Builds higher level segment-level representations
  - Equal/better performance, less computationally expensive than Longformer & BigBird

Chalkidis, I., Dai, X., Fergadiotis, M., Malakasiotis, P., & Elliott, D. (2022). An exploration of hierarchical attention transformers for efficient long document classification. *Computation and Language*. https://doi.org/10.48550/arXiv.2210.05529

Model Architecture



Embeddings

- Huggingface pre-trained weights: Hierarchical Attention Transformer (HAT) / kiddothe2b/adhoc-hierarchical-transformer-base-4096
  - Warm-start with RoBERTa weights (Liu et al., 2019)
  - Ad-hoc HAT (12 sentence + 4 document encoder blocks)
  - Pre-trained on Common Crawl's web crawl corpus
    - Note. This version of the embeddings is not fully pre-trained
  - Takes up to 4096 tokens

Liu Y., Ott, M., Goyal, N., Du, J., ... , & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *Computation and Language*. https://doi.org/10.48550/arXiv.1907.11692

Other Setups

- Target
  - Multi-label (7) regression task
- Loss
  - MSE
- Evaluation
  - 80/20 split on the train set
  - Weighted *r* for evaluation

Alternative Approaches

What We Have Tried

- Other Models
  - DistilBert
    - Cased & Uncased
    - Average of each prompt’s embeddings
  - Longformer
- Other Setups
  - Using MSE, r square, or unweighted r as evaluation metric
  - Other ways to scale or normalize target

What We Haven’t Tried

- Fully pre-trained ad-hoc HAT
- Other variations of HAT
- Using each prompt as a segment in HAT
- Fine-tune as an ordered classification task
- Anything else the audience can think of (:

Thank You!