

ELMA: Encrypted Offloading for Embedded NLP Applications

Final Project Proposal

Group 3

Yihua Liu, Shuocheng Chen, Yiming Ju

UM-SJTU Joint Institute

October 20, 2021

Objectives

An architecture

- High-performance NLP model
- Restricted computational resources
- On embedded device

NLP

Definition

- Natural Language Processing
- Focus on the theories and methods of effective communication between human and computer in the field of natural language.

Applications

- Text classification
- Auto correct
- Machine translation
- Speech recognition
- ...

Computation offloading

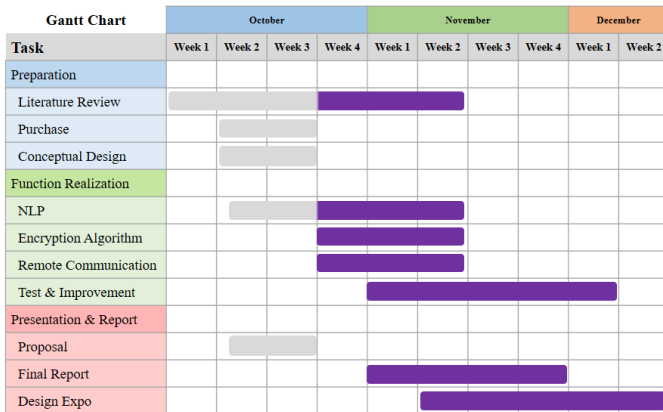
Definition

Migrate part of the data processing in the cloud to the local terminal devices.

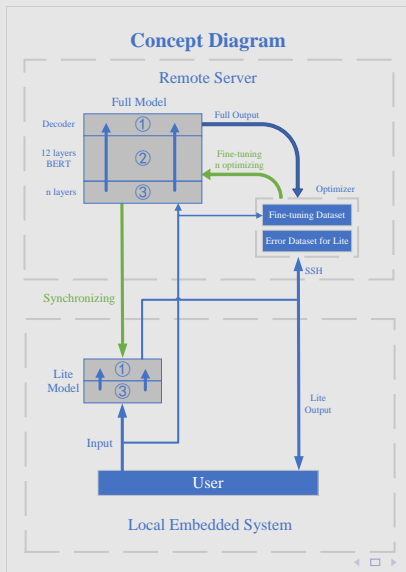
Advantages

- Accelerate computation
- Save energy
- Lower latency
- Protect privacy

Gantt Chart



Concept Design



Transfer Learning

Methodology

- Feature Extraction
- Fine-tuning

Benefits

- simplification of task completion
- better performance
- computational redistribution
- edge computing
- ...

BERT: Pre-trained State-of-the-Art Model

Features

- Fine-tuning based
- Universality for multiple tasks

Further Research

- ALBERT: a lite BERT
- EdgeBERT
- Robust multi-tasking fine-tuning for BERT
- Few-sample fine-tuning optimization
- ...

Question Answering

Definition

Question answering programs can construct answers through the query of a knowledge base or an unstructured collection of documents in a natural language.

Logical blocks

- Data source
- Information retrieval system
- Machine reading comprehension model

Question Answering Example

Text

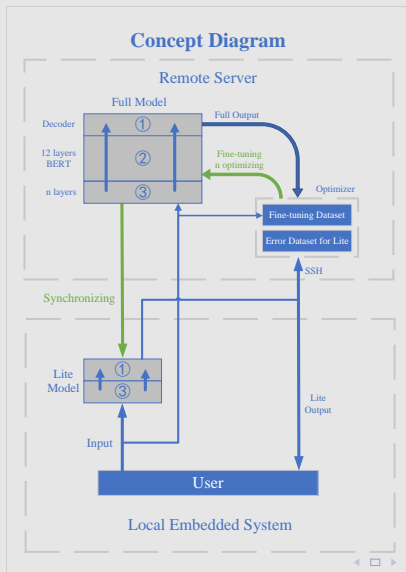
The UM-SJTU Joint Institute has **more than 100** talented and dedicated faculty and staff members working to pursue its mission. We place a high priority on creating an environment that enables faculty and staff to do their best work and values the contributions of all employees in making the JI a world-class educational and research insitute in **China**. Employee commitment is vital to our success. Once Jler, Jler forever.

Question

- How many faculty in UM-SJTU?
- Where is JI?

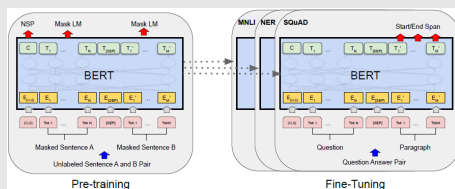
<https://visbert.demo.dataxis.com/>

Concept Design



Related Work

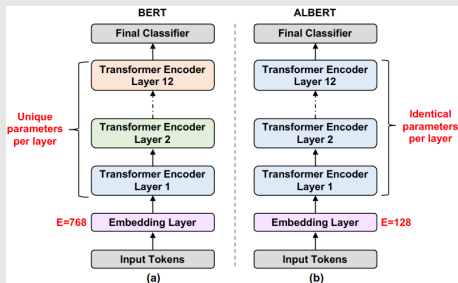
BERT: Bidirectional Encoder Representations from Transformers



The Bidirectional Encoder Representations from Transformers (BERT) [1] divides the NLP model into two steps: pre-training and fine-tuning. The transformer encoder is with Gaussian Error Linear Units (GELU) nonlinearities. BERT uses a Masked Language Model (MLM) for language model pre-training and a Next Sentence Prediction (NSP) task for text-pair representation pre-training. For fine-tuning, take question answering applications as an example, BERT takes question-passage pairs as input and feeds token and [CLS] representations to output layers.

Related Work

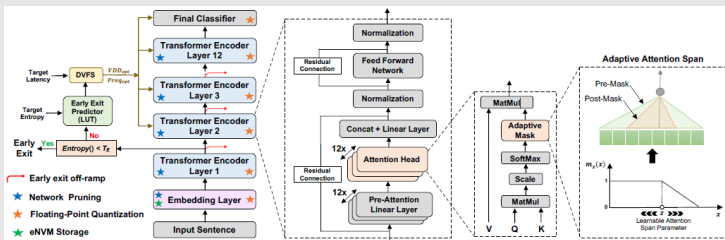
ALBERT: A Lite BERT



ALBERT [2] does 3 significant improvements on traditional BERT model: factorized embedded parameterization, cross-layer parameter sharing, and inter-sentence coherence loss (sentence-order prediction (SOP) loss). ALBERT achieves better performance than BERT given larger configurations and fewer parameters. There are different methods to do cross-layer parameter sharing. ALBERT shares all parameters across layers, but parameters to share can be customized, such as feed-forward network (FFN) or attention parameters only.




Related Work

EdgeBERT



EdgeBERT [3] is a HW/SW (more specifically, hardware/algorithm) co-design for multi-task NLP whose purpose is to reduce energy consumption as much as possible while achieving improvements of accuracy and speed. EdgeBERT proposes several main improvements, including entropy-based early exit (EE) and dynamic voltage-frequency scaling (DVFS). Our work is inspired by the strategies that EdgeBERT adopts to save energy, which is very important for embedded systems.

Reference

-  Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. [arXiv: 1810.04805 \[cs.CL\]](#).
-  Zhenzhong Lan et al. *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. 2020. [arXiv: 1909.11942 \[cs.CL\]](#).
-  Thierry Tambe et al. *EdgeBERT: Sentence-Level Energy Optimizations for Latency-Aware Multi-Task NLP Inference*. 2021. [arXiv: 2011.14203 \[cs.AR\]](#).

Q & A

Thanks!