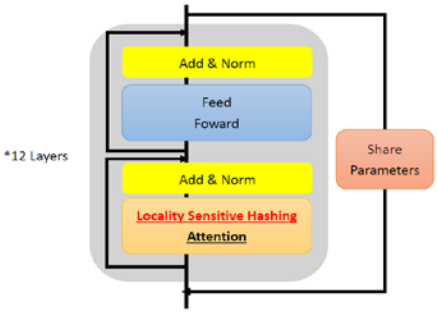
Abstract

1. Introduction

Space computational costs and time complexities are more and more important in the era of large language models. ALBERT greatly reduces the time and space needed for training a language model. Furthermore, Locality Sensitive Hashing attention mechanism also improves on language model's space complexity.



ALBERT-LSH Structure

Fig. 1

2. Purpose

We would like to experiment whether it will achieve a better result if the attention layer for ALBERT is replaced by Locality Sensitive Hashing.

3. Model selection

3.1. Locality Sensitive Hashing

Locality Sensitive Hashing is an attention mechanism that replaces the original dot-product attention and reduces the former space complexity of $O(N^2)$ to $O(N\ lg\ N)$. It randomly rotates Q vectors several rounds and hashes each q_i into several buckets:

$$\Pr_{A^{(1)},\dots A^{(I)}}[h_i(p) = r_{v_i}^{(i)}\ for\ all\ i \in [k] | A^{(i)}q = x^{(i)}]$$

This process finds all related q_is and computes them into attention-matrix with lower computational cost:

$$o_i = \sum_{j \in P_i} exp(q_i * k_i - z(i_i P_i)) v_i / \sqrt{d_k}$$

3.2. Model building

We connected locality sensitive hashing attention layer on ALBERT structure (fig 1). For training process, we use standard ALBERT pre-train model provided by Google for pre-training, then fine-tune it to specific tasks.

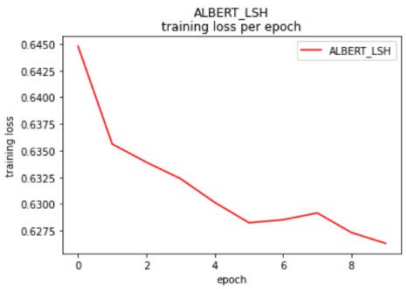


Fig. 2

4. Result

4.1. Model Correctness

We can see the Loss is decreasing through epochs of training. (fig. 2)

4.2. Evaluate on Test Benches

	MRPC	MNLI	SST-2	SQuAD1.1	SQuAD2.0
ALBERT_LSH	0.78	0.35	0.51	13.93	50.07
ALBERT	0.88	0.84	0.93	78.45	72.63

5. Result

Implementing Locality Sensitive Hashing on natural language processing tasks is workable. It reduces space needed for module to compute therefore we can construct a larger module. It is a trade-off between space and time. Our ALBERT_LSH uses less space, but hashing loses some text feature and decreases accuracies. We need further experiments to determine whether our module needs more time to train, or it is already the limit of LSH attention mechanism.