Funded by FNSNF

idiap RESEARCH INSTITUTE EPFL UNIVERSITÉ DE GENÈVE

Fast Transformers with Clustered Attention

Apoorv Vyas [1 2], Angelos Katharopoulos [1 2], François Fleuret [2 3]

[1] Idiap Research Institute [2] Ecole Polytechnique Fédérale de Lausanne [3] University of Geneva, Switzerland

Overview

Goals

- Attention model with linear complexity with respect to sequence length.
- Backward compatibility.

Our approach

- Cluster the queries and use query centroids for attention approximation.
- For each query, recompute attention on a small subset of keys.

Main Results

- Linear complexity of attention computation for a fixed number of clusters.
- Softmax attention approximation without any fine-tuning.

Clustered Attention

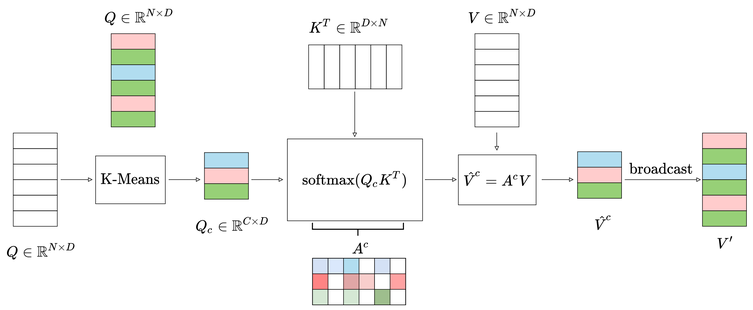

Figure 1: Demonstrating *clustered* attention computation for a sequence of length 8 and number of clusters set to 3. Colors represent the query groups and the computed centroids.

Attention computation is linear $O(NCD)$ for a fixed number of clusters.

Improved-Clustered Attention

For any query Q_i belonging to cluster j, we recompute the attention on the set of top-k keys with highest clustered attention weights A^c_j.

We efficiently compute new values $\hat{V}_i$ using following decomposition:

$$\hat{V}_i = \hat{V}^t_i + \hat{V}^b_i, \tag{1}$$

$\hat{V}^t_i$ is the weighted average of top-k values with recomputed attention weights.
$\hat{V}^b_i$ is the weighted average of rest of the values with clustered attention weights.

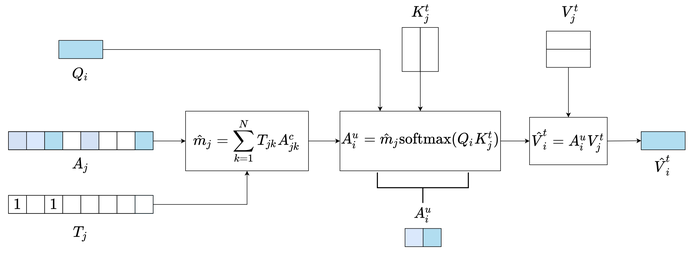

Figure 2: Efficient computation of $\hat{V}^t_i$. T_j identifies the top-k keys. K^t_j and V^t_j denote the set of top-k keys and values. Scaling with $\hat{m}_j$ ensures that the total attention weights sum to one.

For $k < C$ complexity for attention computation remains linear $O(NCD)$.

Results: Backward Compatibility

Approximating Pre-trained RoBERTa on GLUE and Squad:

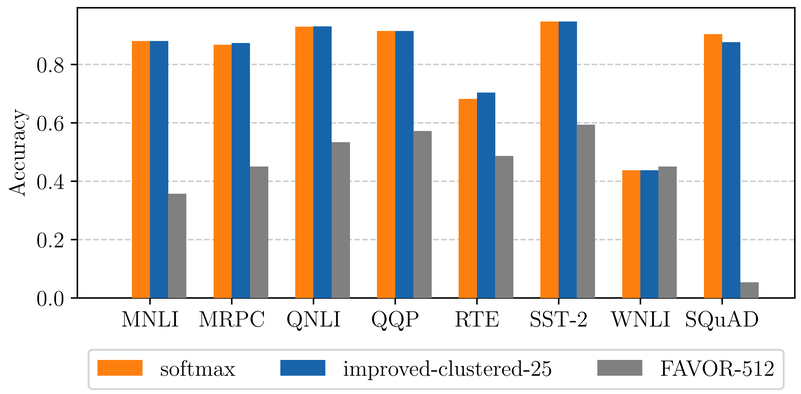

Figure 3: Approximation performance on GLUE and SQuAD benchmarks. For the GLUE tasks, the maximum sequence length is 128 while for SQuAD, it is 384. We use 25 clusters for improved-clustered and 512 projection dimensions for FAVOR.

Approximating Pre-trained Wav2Vec Model:

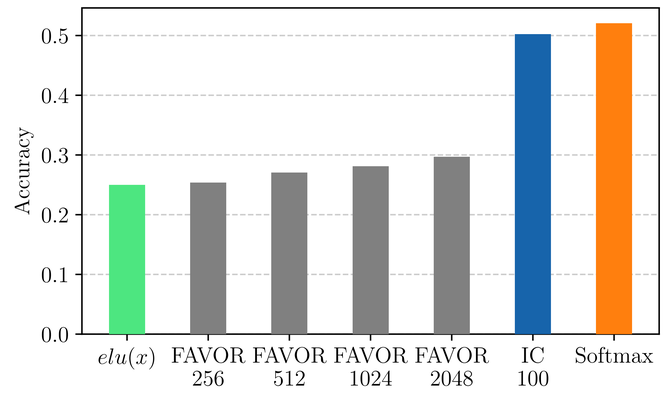

Figure 4: Approximation accuracy on the unsupervised task of the Wav2Vec model. The maximum sequence length is ~1400 on the dev set. We use 100 clusters for improved-clustered (IC) and vary the number of projection dimensions for FAVOR.

Improved-clustered attention can approximate arbitrarily complicated attention patterns with minimal loss in performance.

Results: Automatic Speech Recognition

Training Convergence Wall Street Journal :

	softmax	Reformer	clustered-100	i-clustered-100
PER (%)	5.03	8.59	7.50	5.61
Time/Epoch (s)	2514	2320	803	1325
Convergence Time (h)	87.99	210.09	102.15	72.14

Table 1: We report the test set Phone Error Rate (PER), the time per training epoch (in seconds) and the wall-clock time required for the convergence of each model (in hours).

Training Convergence Switchboard :

	softmax	clustered-100	i-clustered-100
WER (%)	15.0	18.5	15.5
Time/Epoch (h)	3.84	1.91	2.57
Convergence Time (h)	228.05	132.13	127.44

Table 2: We report the test set Word Error Rate (WER), the time per training epoch (in hours) and the wall-clock time required for the convergence of each model (in hours).

Time-Performance Tradeoff:

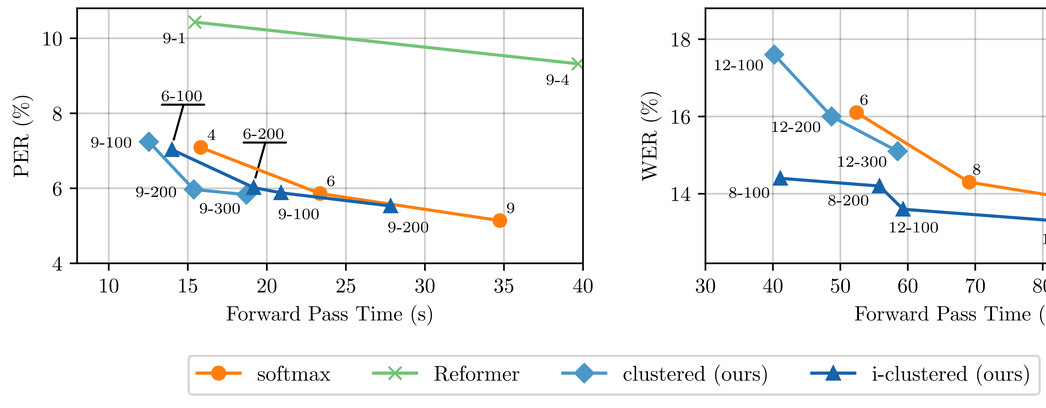

Figure 5: Performance comparison under an equalized computational budget. The numbers near the datapoints denote the number of layers and number of clusters or hashing rounds. Improved-clustered is consistently better than all baselines for a given computational budget.

Improved-clustered attention consistently outperforms baselines on both training convergence time and inference accuracy under a fixed computational budget.