A GPU-based WFST Decoder with Exact Lattice Generation

Zhehuai Chen\textsuperscript{1,3}, Justin Luitjens\textsuperscript{2}, Hainan Xu\textsuperscript{3}, Yiming Wang\textsuperscript{3}, Daniel Povey\textsuperscript{3}, Sanjeev Khudanpur\textsuperscript{3}

\textsuperscript{1}SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University
\textsuperscript{2}NVIDIA, USA
\textsuperscript{3}Center for Language and Speech Processing, Johns Hopkins University

chenzhehuai@sjtu.edu.cn, jluitjens@nvidia.com, dpovey@gmail.com, \{hxu31,yiming.wang, khudanpur\}@jhu.edu

Abstract

We describe initial work on an extension of the Kaldi toolkit that supports weighted finite-state transducer (WFST) decoding on Graphics Processing Units (GPUs). We implement token recombination as an atomic GPU operation in order to fully parallelize the Viterbi beam search, and propose a dynamic load balancing strategy for more efficient token passing scheduling among GPU threads. We also redesign the exact lattice generation and lattice pruning algorithms for better utilization of the GPUs. Experiments on the Switchboard corpus show that the proposed method achieves identical 1-best results and lattice quality in recognition and confidence measure tasks, while running 3 to 15 times faster than the single process Kaldi decoder. The above results are reported on different GPU architectures. Additionally we obtain a 46-fold speedup with sequence parallelism and multi-process service (MPS) in GPU.

Index Terms: ASR, Decoder, Parallel Computing, WFST, GPU, Lattice Processing, Confidence Measure

1. Introduction

Recent advances in deep learning based automatic speech recognition (ASR) invoke growing demands in large scale speech transcription and analysis. The main computation of a common speech transcription system includes: acoustic model (AM) inference and weighted finite-state transducer (WFST) decoding. To reduce computation of the AM inference, researchers have proposed a variety of efficient forms of AMs, including novel structures [1, 2], quantization [3, 4], frame-skipping [5, 6] and end-to-end systems [7, 8]. Meanwhile, algorithmic improvement, e.g. pruning [9, 10] and lookahead [11, 12] is the common way to speed up decoding.

GPU-based parallel computing is another potential direction which utilizes a large number of units to parallelize the computation. As most of the computation takes the form of matrix operation, the AM inference can be sped up by sequence batching [13] similar to that during training [14]. Nevertheless, GPU-based decoding is less prevalent, despite its success in small language model (LM) [15]. Two reasons hamper its wide application: i) it is hard to utilize large LMs since GPUs have smaller memory. ii) the solution is not general, e.g. with specific AM or LM requirements, and lack of lattice generation.

This work is an extension to Kaldi [16], applying GPU parallel computing in WFST decoding. It is a general-purpose offline decoder\textsuperscript{1}, which does not impose special requirements on the form of AM or LM, and works for all GPU architectures. To utilize large LMs in the 2nd pass and support rich post-processing, our design is to decode the WFSTs and generate exact lattices [17]. The work is open-sourced\textsuperscript{2}, and will be compatible with all released Kaldi recipes.

The rest of the paper is organized as follows. Prior work is compared in Section 2. Our main contributions are described in Sections 3: i) implement token recombination as an atomic GPU operation with no precision loss. ii) propose a load balancing strategy for token passing scheduling among GPU threads. iii) redesign parallel versions of exact lattice generation and pruning. Experiments are conducted on the Switchboard corpus in Section 4, followed by conclusion in Section 5.

2. Comparisons with Prior Work

[18] introduced a parallel Viterbi algorithm in CPU, which concentrates on partitioning the WFSTs to trade off between load balancing and synchronization overhead. To cope with similar issues in the context of GPU architectures, We give general LM-independent solutions in Section 3.3. [15, 19] first utilized GPUs in ASR decoding and showed significant speedup. However, the limited GPU memory becomes a bottleneck for large vocabulary tasks. A series of works were proposed in [20, 21, 23] and attempted to cope with large LM issue by an on-the-fly rescoring [10] algorithm. Advantages of our method include: i) general solution and better performance, by generating an exact lattice in the first pass decoding followed by a second pass rescoring, which always obtains slightly better results [22]. The resultant lattice also benefits other speech processing tasks, e.g. confidence measure. ii) faster decoding speed. This work benefits from less synchronization overheads, better load balancing, and sequence parallelism. Moreover, it also requires less data transfer compared with [23]. Finally, to the best of our understanding, [24] is the only open-source project in this field, but it only implemented the basic Viterbi decoding without combining AM posteriors and beam [9], which cannot be applied to ASR.

3. Parallel Viterbi Beam Search

The proposed system works in a 2-pass decoding framework [25] to tackle the language model size problem and enable rich post-processing based on exact lattices.

\textsuperscript{1}Recent advances in CPU decoding make it easier to build a real-time decoder [2, 6] for online streaming applications. Thus our goal is to transcribe tons of offline audios.

\textsuperscript{2}https://github.com/chenzhehuai/kaldi/tree/gpu-decoder

This work was partially supported by DARPA LORELEI Grant No HR0011-15-2-0024 and NSF Grant No CRI-1513128. The first author was supported by the National Key Research and Development Program of China under Grant No.2017YFB1002102, the China NSFC project (No. 61573241) and the Shanghai International Science and Technology Cooperation Fund (No. 16550720300).

10.21437/Interspeech.2018-1339
3.1. System Outline
Figure 1 shows our framework, with two GPU concurrent streams performing decoding and lattice-pruning in parallel launched by CPU asynchronous calls. Namely, at frame \((t + 1)\), stream 2 prunes the lattice generated by stream 1 at frame \(t\). The procedure of decoding is similar to the CPU version [16] but works in parallel with specific designs discussed in the following sections. Load balancing controls the thread parallelism over both WFST states and arcs.

3.2. GPU-based Token Recombination

The synchronization overhead in Viterbi search comes from the token recombination. In ASR decoding, Viterbi search is reformulated to token passing algorithm [28], where a token represents a partial match up to frame \(t\) and each WFST state at frame \(t\) holds a single moveable token. At each frame, the Viterbi path is obtained by a \textit{token recombination} procedure, where a \textit{min} operation is performed on each state over all of its incoming arcs (e.g. state 7 in Figure 2 and the incoming arcs from state 2, 5 and 7), to compute the best cost and the corresponding predecessor of that state.

![Figure 1: Framework of Parallel Viterbi Beam Search with Exact Lattice Processing](image)

Algorithm 1 Thread-level Token Recombination (Inputs: accumulated cost, an out-going WFST arc and a current token)

1: procedure RECOMBINE(cost, arc, curTok)
2: \hspace{1em} oldTokPack = state2tokPack[arc.next_state]
3: \hspace{1em} curTokPack = packFunc(cost,arc.id) \triangleright pack into uint64
4: \hspace{1em} ret = atomicMin \(^3\)(oldTokPack,curTokPack)
5: \hspace{1em} if \hspace{0.5em} ret > curTokPack \hspace{1em} then \triangleright recombine
6: \hspace{1em} perArcTokBuf[arc.id] = *curTok \hspace{1em} \triangleright store token

The original CPU algorithm conducts recombination in serial. We discuss how to implement GPU-based token recombination in this section and leave how to parallelize it to the next section. A naive implementation is by adding critical sections [27] to recombine tokens in serial. Such design is inefficient and causes unexpected deadlocks on pre-Volta GPUs. [15] performs a per-state reduction on the token passing results, which requires extra burden of bookkeeping memory to eliminate write conflicts. [20] encodes all token data in 32 bits and does the recombination by atomic operations. This method loses precision and makes the decoder dependent on models.

We propose Algorithm 1, which is a general method for GPU token recombination with no precision loss. The algorithm is performed per GPU thread, and runs in parallel for every arc, e.g. the arc from state 2 to state 7 in Figure 2 is processed by a thread following this algorithm (details on parallelism are in the following section). We pack the cost and the arc index into a uint64 to represent the token before recombination, with the former one in the higher bits for comparison purpose. In each frame, we save the token information in an array whose size is the number of arcs. This ensures there are no write conflicts between threads since each arc can be accessed at most once in each frame. After passing all tokens, we aggregate survived packed tokens, unpack them to get arc indexes, and store token information from the former array to token data structures exploiting thread parallelism.

![Figure 2: Example of Dynamic Load Balancing. The dashed box denotes a CUDA cooperative group and different groups are with different colors. Each group is controlled by Thread 0 of it. After processing all the forward links of a state, Thread 0 accesses the dispatcher and the next token is dynamically decided by atomic operation. Group 0 and 1 work in parallel.](image)

\(^3\)atomicMin(*address, val) [26]. Computes min(*address, val), writes the result to address, and returns the original *address.
3.3. Parallel Token Passing

Another issue in parallel algorithm is load balancing. For each state, we traverse their out-going arcs in parallel, till we reach a final state. Because WFST states might have different numbers of out-going arcs, the allocation of states and arcs to threads can result in load imbalance. Unlike in [15, 18] where the authors redesigned the WFST structure, this work does not require any structural change on AM, lexicon or LM.

We propose static load balancing, which starts off by assigning roughly equal number of arcs to each thread, before processing the corresponding arcs. This design requires the accumulated sum of the arc numbers of active tokens \( N \), which requires extra computation and kernel launch time.

Motivated by [28], we also introduce dynamic load balancing in Algorithm 2. We use a dispatcher in charge of global scheduling, and make \( N \) threads as a group (\( N = 32 \)) to process all arcs from a single token. When the token is processed, the group requests from the dispatcher a new token.

We implement task dispatching as an atomic operation. Figure 2 shows an example. We compare static and dynamic load balancing in Section 4.2.

Algorithm 2 Grid-level Token Passing (\( N=32 \); Inputs: the current active token vector)

1: procedure DYNAMIC LOAD BALANCING (toks)
2: \( group := \text{cooperative groups::} \text{tilled partition}(32) \)
3: if \( \text{group.thread.rank}() = 0 \) then \( \text{rank 0 in each group} \)
4: if \( i = \text{atomicAdd}(\text{global}_d, 1) \) then \( \text{allocate new tokens} \)
5: \( \text{group.shifft}(0) \) \( \text{broadcasts i to whole group} \)
6: if \( i != \text{sizeof(tok)} \) then return \( \text{all tokens processed} \)
7: for \( \text{arc in tok2arc(toks[i])} \) do \( \text{thread parallelism} \)
8: call Recombine(toks[i].cost+arc.cost, arc, toks[i])

3.4. Exact Lattice Generation

An exact lattice [17] is one that stores precise costs and state level alignments, which is crucial to LM rescoring performance, and enables rich post-processing. Implementing the lattice processing in GPU is non-trivial. [23] simply decodes in GPU and generates lattice in CPU, which significantly slows down the decoder and results in overheads of device-to-host (D2H) memory copy. We thoroughly solve this problem by redesigning parallel versions of lattice processing algorithms [17] in the following.

In lattice generation, for each token passing operation, a lattice arc needs to be stored in GPU memory. We design a GPU vector to store arcs using push_back function implemented by an atomic operation, where the memory is pre-allocated.

```cpp
// implementation of v.push_back(val)
int idx = atomicAdd(cnt_d, 1); // id = cnt_d++
mem_d[idx] = *val; // store data
```

To reduce the overhead of atomc operations, we pre-allocate \( K \) vectors and randomly select one to push_back the lattice arc into. After token passing, arcs from \( K \) vectors are aggregated. We process lattice nodes in a similar manner.\(^3\)

\(^3\)A GPU DeviceScan operation over around 10K integers.

\(^4\)Beside the confidence measure [29] examined in this paper, it also benefits tasks including minimum Bayes-risk decoding [30], system combination [31], discriminative training [32], etc.

\(^5\)In our preliminary experiments, this optimization speeds up lattice arcs by 10 times with \( K = 32 \). The speedup of lattice nodes is small.

3.5. Lattice Pruning

The parallel lattice pruning is based on the algorithm described in [33, 17] with necessary modifications for GPU parallelization. The original CPU version of lattice pruning iteratively updates extra costs of nodes and arcs until they stop changing, where extra cost is defined as the difference between the best cost including the current arc and the best overall path. Arcs with high extra costs are then removed, along with nodes that are not associated with any remaining arcs. In GPUs, we i) parallelize the iterative updating of nodes and arcs over GPU threads; ii) use a global arc vector to replace the linked lists in the old implementation, for its lack of random access features to enable parallel access; iii) implement the extra cost updating as an atomic operation to eliminate write conflicts among threads.

When a lattice arc is pruned, we do not physically remove the arc, as memory allocation is expensive. Instead, we do a final merging step to aggregate all remaining arcs using thread parallelism and the GPU vector proposed in Section 3.4. Memory consumption of this implementation is discussed in Section 4.3. Meanwhile, we do not prune lattice nodes because: i) we need a static mapping for each arc to trace the previous and the next nodes before and after D2H memory copy. We use frame index \( t \) and vector index \( i \) to trace a node, thus node positions in the vector cannot be changed. ii) the lattice is constructed in CPU by iterating remaining arcs, thus nodes are implicitly pruned. iii) node D2H copy is done in each frame asynchronously, which does not introduce overheads.

4. Experiments

4.1. Setup

In Switchboard 300 hours corpus, we have two acoustic model baseline setups: one is a “TDNN-LSTM-C” model described in [2] with lattice-free MMI (LF-MMI) objective [34] with subsampled frame rate. To eliminate the effect of sub-sampling, the other is a stacked bidirectional LSTM network [35] with cross entropy (CE) objective with the original frame rate.

Evaluation is carried out on the NIST 2000 CTS test set. A 30k-vocabulary tri-gram LM trained from the transcription of Switchboard corpus and a Fisher interpolated 4-gram LM are used for decoding and lattice rescoring respectively. Kaldi 1-best decoder and lattice decoder are taken as CPU baselines.\(^7\)

1-best results and lattice quality are both examined. For fair comparisons, lattice densities (lat.den., measured by arcs/frame) [25] are kept identical for CPU and GPU decoding results. We report word error rate (WER), lattice rescored WER (+rescored), lattice oracle WER (OWER) [36] in recognition tasks. Normalized cross entropy (NCE) [37] is reported to show the quality of word level confidence measure after lattice rescoring. A decision tree is trained from the posterior histogram to map the posterior probabilities to confidence scores [38]. Real-time factor (RTF) is taken to evaluate the decoding speed and the speedup factor (\( \Delta \)) versus corresponding CPU baselines. Because of the focus of this work, we report RTFs excluding AM inference in the tables. Overall RTFs are given when analyzing the tables. As the current goal is to build a fast speech transcription system, latencies of online streaming applications have not been taken into account. All experiments are conducted on E5-2686 v4 @ 2.30GHz; with 1 socket (8 threads). A Tesla V100 is used in default and different architectures from Kepler to Volta are examined in Section 4.3.

\(^7\)decode-faster-mapped and latgen-faster-mapped in Kaldi toolkit.
4.2. Performance and Speedup

Table 1 compares the proposed GPU lattice decoder with the CPU baseline under the same beam. All indicators are very close if not identical. The slight differences might be caused by different orders in which states are visited during decoding.

<table>
<thead>
<tr>
<th>system</th>
<th>lat. den.</th>
<th>WER</th>
<th>+rescored</th>
<th>OWER</th>
<th>NCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU 30.3</td>
<td>15.5</td>
<td>14.3</td>
<td>11.2</td>
<td>0.322</td>
<td></td>
</tr>
<tr>
<td>GPU 30.2</td>
<td>15.5</td>
<td>14.3</td>
<td>11.2</td>
<td>0.328</td>
<td></td>
</tr>
</tbody>
</table>

Decoding RTF speedups of both 1-best and lattice decoders are shown in Table 2. We obtain a 15-fold and 9.7-fold single-sequence speedups for 1-best and lattice decoders respectively. Most of the speedups come from parallel traversals of WFST states and arcs. The second most significant improvement is from lattice pruning. Beside the parallelism benefit similar to that of token passing, the GPU lattice pruning reduces the data amount of D2H memory copy and separate kernel launch costs. We obtain a 46-fold speedup if sequence parallelism is utilized with GPU MPS, which reduces context switching of multi-process [26]. For comparison, similar multi-process parallelism is applied in CPU but only results in a 1.8-fold speedup.

Although the speedup of atomic operation is not significant, it removes the critical section discussed in Section 3.2, which enables the use of older GPU architectures, shown in the next section. Meanwhile, dynamic load balancing is a comparable general solution versus both static load balancing, shown in the 4th row, and the graph partition in [18].

To make a fair comparison of the overall RTFs, we take a separate GPU to do single-sequence inference for both baseline and the GPU decoder. The single-sequence overall RTFs of baseline and the proposed 1-best decoder are 0.18 and 0.03

Deep integration of inference and decoding can be future topics.

Table 2: Speedup of the Proposed Method (beam=14).

<table>
<thead>
<tr>
<th>system</th>
<th>search RTF</th>
<th>lattice + RTF</th>
<th>search Δ</th>
<th>lattice Δ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPU 30.3</td>
<td>0.16</td>
<td>0.27</td>
<td>10X</td>
<td>10X</td>
</tr>
<tr>
<td>+ 8-sequence (1 socket)</td>
<td>0.035</td>
<td>0.028</td>
<td>46X</td>
<td>9.7X</td>
</tr>
<tr>
<td>GPU 30.2</td>
<td>0.011</td>
<td>0.028</td>
<td>15X</td>
<td>9.7X</td>
</tr>
<tr>
<td>+ atomic operation</td>
<td>0.015</td>
<td>0.075</td>
<td>11X</td>
<td>3.6X</td>
</tr>
<tr>
<td>+ dyn. load balancing</td>
<td>0.011</td>
<td>0.075</td>
<td>11X</td>
<td>3.6X</td>
</tr>
<tr>
<td>+ lattice prune</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>+8-sequence (MPS)</td>
<td>0.0035</td>
<td>0.0080</td>
<td>46X</td>
<td>34X</td>
</tr>
</tbody>
</table>

4.3. Analysis

Figure 3 shows the proposed decoder works well in varieties of ASR systems and consistently improves the speed.

• GPU architectures.

The proposed system can be used in GPU architectures older than Kepler. It achieves a 3-fold speedup in K20 versus CPU.

• Acoustic model frame rates.

\(^{9}\)OWER in this dataset is imprecise as we do not normalize the text in its scoring using glm. In LibriSpeech [34], we consistently observe that OWER is 23% of WER (3.4 v.s. 14.7) for both CPU and GPU.

\(^{8}\)The AM inference can be further improved by sequence batching [13], which speeds up 10 times in our preliminary trials.

\(^{10}\)The result is obtained from latgen-faster-mapped-parallel in Kaldi. 1-best decoder do not have such implementation.

5. Conclusions

In this work, we describe initial work on an extension of the Kaldi toolkit that supports WFST decoding on GPUs. We design parallel versions of decoding and lattice processing algorithms. The proposed system significantly and consistently speeds up the CPU counterparts in a varieties of architectures.

The implementation of this work is open-sourced. Future works include the deep integration with sequence parallel acoustic inference in GPUs.

\(^{11}\)Lattice generation statistics in CPU is included in token passing.
6. References


