HYDRA: A Hybrid GPU-CPU Engine for LVCSR
Jungsuk Kim, Paul Märgner, Jike Chong, Ian Lane

A Hybrid GPU+CPU Speech Recognition Engine

- For intuitive Voice and Interactive Multimodal systems robust and responsive speech recognition is crucial
  - Robust
    - Acoustic robustness ➔ Large Acoustic Models
    - Linguistic robustness ➔ Large Vocabulary (1M+ words)
  - Responsive
    - Low latency ➔ Faster than real-time search
- Current state-of-the-art speech recognition systems are optimized for either robustness or responsiveness
  - Robustness: 5-10 x real-time >95% accuracy
  - Responsiveness: real-time 85% accuracy

On-The-Fly Partial Hypothesis Rescoring

<table>
<thead>
<tr>
<th>CPU Control</th>
<th>Manycore Control</th>
<th>GPU Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initialize data structures</td>
<td>Phase 0 Iteration Control</td>
<td>Prepare ActiveSet</td>
</tr>
<tr>
<td>On-The-Fly Rescoring (LM Lookup)</td>
<td>Phase 1 Compute Observation Probabilities</td>
<td>Phase 2 WFST Search</td>
</tr>
<tr>
<td>Collect Backtrack Info</td>
<td>Save Backtrack Log</td>
<td></td>
</tr>
<tr>
<td>Backtrack</td>
<td>Output Results</td>
<td></td>
</tr>
</tbody>
</table>

Decoding Process

1. Prepare Active Hypotheses Set
   - Gather active speech recognition hypotheses (word and phone sequences) from previous frame.
2. Compute Observation Probabilities
   - Compute likelihood of phonetic models (Gaussian Mixture Model) for current input feature.
3. On-The-Fly Partial Hypothesis Rescoring
   - On the CPU, rescore likelihoods of partial hypotheses using a higher order N-gram language model stored in main memory.
   - Partial Hypothesis rescoring and the observation probability computation can be performed concurrently.
4. Viterbi Search
   - Frame synchronous Viterbi search is performed on the GPU using WFST network composed using unigram language model.
   - Maintaining N-best paths during decoding to ensure good hypotheses are not pruned early.

Experimental Evaluation

- Acoustic Model
  - 51284 Data Set
  - 3000 tied state
  - 16 mixture Gaussians
  - 39th MFCCs features
- Language Model
  - Wall Street Journal 5k
  - 1-gram: 5k entries
  - 2-gram: 1.6M entries
  - 3-gram: 2.7M entries
- Evaluation Set
  - Nov. 92 ARPA WSI test set
  - 330 sentences
- NVIDIA GTX 680
  - Kepler architecture
  - 1536 CUDA cores

- 20x speed-up compared to standard WFST decoding on CPU at word accuracy of 93.80%
- 95.40% maximum accuracy is achieved.

- 1M vocab. network can be decoded on a modern GPU.
- Network size does not significantly affect decoding speed.

This work was partially supported by Samsung, Cisco and Nvidia.