Submodular Subset Selection for Large-scale Speech Training Data

Kai Wei1, Yuzong Liu1, Katrin Kirchhoff1, Chris Bartels2, Jeff Blimes1

1Department of Electrical Engineering, University of Washington, Seattle
2SRI International, Menlo Park, CA

Motivations

Problems with ever-increasing data set:
- Increased computational demands.
- Gains by new data are often sublinear.
- Redundant data is processed repeatedly (EM or back-propagation).

Problem scenario:
Given a large amount of acoustic data (>1000 hours),
- how to identify a subset of the data that provides the most information?
- what is the smallest degree of information loss given a drastic data set size reduction?

Applications:
- Data selection for human annotation (batch active learning).
- Faster model configuration tuning on a representative and small subset.

Submodular Functions for Speech Data Subset Selection

- A special class of set functions with diminishing returns property.
- Given a finite set V, a set function \( f : 2^V \to \mathbb{R} \) is submodular, if
  \[
  f(k \cup S) - f(S) \geq f(k \cup R) - f(R),
  \]
  \( \forall R \subseteq S \subseteq V, \forall k \in V \setminus S. \)
- Discrete analog of convexity.

Problem Formulation:
- A large set of speech training utterances: \( V = \{ v_1, v_2, \ldots, v_n \}. \)
- Each utterance has a cost: \( \{ c(v_1), c(v_2), \ldots, c(v_n) \}. \)
- The cost of a subset \( S \subseteq V: c(S) = \sum_{v \in S} c(v). \)
- A submodular set function \( f : 2^V \to \mathbb{R} \) represents the value of each subset of \( V. \)
- Amount of data to be selected: \( B. \)
- Training data subset selection becomes:
  \[
  \max_{S \subseteq V, |S| \leq B} f(S).
  \]

Graph-based Submodular Functions:
- Facility location function:
  \[
  f_{fac}(S) = \sum_{i,j \in S} \max_{k} w_{ij},
  \]
  where \( w_{ij} \) is the similarity measure between speech utterances \( i \) and \( j. \)

  - Saturated coverage function:
  \[
  f_{sat}(S) = \sum_{i \in V} \min\{ C_i(S), \beta C_i(V) \},
  \]
  where \( C_i(S) = \sum_{j \in S} w_{ij} \) and \( \beta \) is the saturation threshold.

Limitation of Graph-based Submodular Functions:
Requires a pair-wise similarity graph. Becomes infeasible when \( n \) is large (in the millions or billions).

Solution: Feature-based Submodular Function:
\[
\sum_{u \in U} g(m_u(S)).
\]

- \( g() \): a concave function.
- \( U \): a set of the features (e.g., phones, triphones, triphone-states).
- \( m_u(S) = \sum_{v \in S} m_u(v) \): the relevance score of the feature \( u \in U \) within the set \( S. \)
- Does not require a pair-wise similarity graph.
- Scalable to much larger data set.

Greedy Algorithm for Problem 1

Algorithm 1 Greedy algorithm for knapsack constrained submodular max \([1]\)

1. Input: a monotone submodular function \( f \), budget constraint \( B \), and a list of costs \( \{ c(v_1), \ldots, c(v_n) \}. \)
2. Initialization \( S \leftarrow \emptyset. \)
3. repeat
4. Pick an element \( v^* \in \arg \max_{v \in V \setminus S} \frac{f(v \cup S) - f(S)}{c(v)}. \)
5. Update \( S \leftarrow S \cup v^*. \)
6. until Reaching the budget, i.e., \( c(S) > B. \)

- Approximately solves Problem 1 with constant factor guarantee \( \frac{1}{2}(1 - 1/e). \)
- Empirically much better than \( \frac{1}{2}(1 - 1/e) \), and often close to 1.
- Can be sped up to almost linear-time complexity, thanks again to submodularity.

Data and Systems

Task: Speech data subset selection for training a word recognizer.

Set-up:
- Training data: Switchboard, Switchboard Cellular, and Fisher corpora (1300 hours in total).

Evaluate with two different acoustic model paradigms:
- GMM-HMM: SRI’s DECIPHER system.
- DNN: A system also developed at SRI.

Baseline selection methods:
- Random baseline: Randomly draw specified amount of the training data.
- Histogram-entropy baseline [2]: Choose the subset with a maximum-entropy distribution over linguistic units.

Empirical Results

<table>
<thead>
<tr>
<th></th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>52.1 ± 1.5</td>
<td>38.2 ± 0.2</td>
<td>35.1 ± 0.3</td>
<td>34.4 ± 0.2</td>
<td>31.0</td>
</tr>
<tr>
<td>HE (words)</td>
<td>49.6</td>
<td>36.5</td>
<td>34.8</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>HE (tripones)</td>
<td>47.5</td>
<td>37.6</td>
<td>34.2</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SM (tripones)</td>
<td>47.5</td>
<td>35.7</td>
<td>33.3</td>
<td>32.6</td>
<td>31.0</td>
</tr>
</tbody>
</table>

Word error rates for the HMM-GMM system, for subsets chosen by random (Rand), histogram-entropy (HE), and the submodular (SM) selection.

<table>
<thead>
<tr>
<th></th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
<th>20%</th>
<th>all</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rand</td>
<td>43.7 ± 0.5</td>
<td>34.3 ± 0.9</td>
<td>31.5 ± 0.5</td>
<td>29.6 ± 0.2</td>
<td>26.0</td>
</tr>
<tr>
<td>HE (tripones)</td>
<td>42.8</td>
<td>33.9</td>
<td>31.3</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>SM (tripones)</td>
<td>41.1</td>
<td>31.8</td>
<td>29.3</td>
<td>28.2</td>
<td>26.0</td>
</tr>
</tbody>
</table>

Word error rates for the DNN system.

Acknowledgements

This work is based on research sponsored by Intelligence Advanced Research Projects Activity (IARPA) under agreement number FA8650-12-2-7263.

References