Resource Constrained Structured Prediction

Tolga Bolukbasi, Kai-Wei Chang, Joseph Wang, Venkatesh Saligrama

AAAI 2017
Learning with adaptive budget

Input

Output

System cost

“cat”

cost

“cat”
Exploiting the structure for budget
Budgeted structured learning – setup

\[ R(X, Y) = \mathcal{L}(F(X), Y) \]

Modified Risk

\[ R(X, Y, S) = \mathcal{L}(F(X, S), Y) + \sum_{i=1}^{m} \sum_{k \in S_i} c_k \]

Feature subset for each part

\[ S = [s_1 \ldots s_m] \]
Budgeted structured learning – policy

Adaptive Structured Feature Selector

\[ \Pi: \mathcal{X} \rightarrow \mathcal{S} \]

Maps from “initial” part features to subsets of features adaptively

**Goal:** find \( \Pi \) to minimize expected loss

\[ E_{X,Y}[R(X,Y,\Pi)] \]

\[ R(X,Y,\Pi) = \sum_{S \in \mathcal{S}} R(X,Y,S)1_{\Pi(X)=S} \]
Challenges

- How to find the best feature states for training?
  - Combinatorial search space

- What should be the form of policy?
  - Varying sample shape/structure
  - Combinatorial output space
Finding the best feature states

\[ S^*(X, Y) = \min_{S \in S} R(X, Y, S) \]

- Greedy approximation with trajectory search
  - Start with initial state
  - Do greedy search over all single feature acquisitions
  - Pick the one that minimizes immediate loss
  - Repeat until all features are acquired
1-step policy

- Upper bound the risk \( R(X, Y, \Pi) = \sum_{S \in S} R(X, Y, S) 1_{\Pi(X) = S} \) with the worst case one \( R_{\text{max}}(X, Y) = \max_{S_1, S_2 \in S} (R(X, Y, S_1) - R(X, Y, S_2)) \)

\[
\Pi^* = \min E_{X,Y} \left[ R_{\text{max}}(X, Y) 1_{\Pi(X) \neq S^*(X,Y)} \right]
\]

**Theorem:** 1-step system is equivalent to cost-sensitive structured learning problem

Finding \( R_{\text{max}}(X, Y) \) is also combinatorial, approximate using greedy search as before
Anytime policy (sequential)

**Anytime policy:** Greedily find trajectory of feature subsets and train sequence of 1-step systems
Relevant work

- Feature cost dominates
- Reduction of structured prediction to sequential multiclass prediction
  - (emma)
- Input shape/structure is fixed
  - our own eccv paper
- Inference cost dominates

Previous work:
- Weiss and Taskar, 2013
- JW et al. 2014
Experiments – OCR

Approximately same performance, ~30% time reduction
Experiments – Dependency Parsing

Features:
- Part-of-speech tags
- Lexicons (Lemmas)
- Relationships to surrounding words

Words using relationship features
Thanks for listening!

Questions?

Features and outputs

<table>
<thead>
<tr>
<th>Structured labels</th>
<th>Features for each part</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Words</td>
<td>• RGB histograms</td>
</tr>
<tr>
<td></td>
<td>• HOG features</td>
</tr>
<tr>
<td>• Parse tree</td>
<td>• Part-of-speech tags</td>
</tr>
<tr>
<td></td>
<td>• Lemmas</td>
</tr>
<tr>
<td></td>
<td>• Word relationships</td>
</tr>
</tbody>
</table>