Structured Support Vector Machines for Noise Robust Continuous Speech Recognition

S.-X. Austin Zhang and Mark Gales

31 August 2011

Cambridge University Engineering Department

Interspeech 2011 Florence
SVMs for Continuous Speech Recognition

• SVMs generate boundaries between classes
  – For isolated digit recognition, no problem (10 classes)
    sequence kernels: map to fixed length
  – Continuous speech? too many classes! (e.g., 6 digits $\Rightarrow 10^6$ classes)

• Simplest approach:
  1. Using HMM-based Segmentation
  2. For each segment, isolated classification

Problem: Restricted to one fixed segmentation

• Alternatively, incorporate structures into SVM classes – Structured SVM!
Structured SVMs for CSR

- Model whole utterance – observations $O$, word sequence $w$
  - *joint* features $\phi(O, w)$. How to build?
    - generative model to extract features $\Rightarrow$ model compensation

\[
\begin{bmatrix}
\log P(o; \lambda^{one}) \\
\vdots \\
\log P(o; \lambda^{zero})
\end{bmatrix}
\]

- **Target**: Learn discriminative parameters $\alpha$

Decoding by match score: $\arg\max_w \alpha^T \phi(O, w)$
Structured SVM Training

- Training $\alpha$, so that match score of correct reference $w_{\text{ref}}$ $\geq$ all competing $w$

\[
\min_{\alpha, \xi} \frac{1}{2} ||\alpha||^2 \\
\text{s.t.} \quad \begin{align*}
\alpha^T \phi \left( \begin{array}{c} 123 \\ O^{(1)} \end{array} \right) &\geq \alpha^T \phi \left( \begin{array}{c} 000 \\ O^{(1)} \end{array} \right) + 1, \\
\alpha^T \phi \left( \begin{array}{c} 123 \\ O^{(1)} \end{array} \right) &\geq \alpha^T \phi \left( \begin{array}{c} 001 \\ O^{(1)} \end{array} \right) + 1, \\
\vdots &\vdots \\
\alpha^T \phi \left( \begin{array}{c} 456 \\ O^{(n)} \end{array} \right) &\geq \alpha^T \phi \left( \begin{array}{c} 000 \\ O^{(n)} \end{array} \right) + 1, \\
\alpha^T \phi \left( \begin{array}{c} 456 \\ O^{(n)} \end{array} \right) &\geq \alpha^T \phi \left( \begin{array}{c} 001 \\ O^{(n)} \end{array} \right) + 1, \\
\alpha^T \phi \left( \begin{array}{c} 456 \\ O^{(n)} \end{array} \right) &\geq \alpha^T \phi \left( \begin{array}{c} 999 \\ O^{(n)} \end{array} \right) + 1,
\end{align*}
\]

- Hard-margin, “+1” means “0-1 loss”
Structured SVM Training

• To generalize the training
  – Replace “0-1 loss” as $\mathcal{L}(w_{\text{ref}}, w)$
  – Introduce slack variable $\xi_i$ for linearly nonseparable cases

\[
\begin{align*}
\min_{\alpha, \xi} & \quad \frac{1}{2} \|\alpha\|^2 + C \sum_{i=1}^{n} \xi_i \\
\text{s.t.} & \quad \alpha^T \phi(\text{"1 2 3"}) \geq \max_{w \neq \text{"1 2 3"}} \left\{ \alpha^T \phi(w_{\text{ref}}, w) + \mathcal{L}(\text{"1 2 3"}, w) \right\} - \xi_1
\end{align*}
\]

• Unconstrained form $\min_{\alpha} \mathcal{F}_{\text{ssvm}}(\alpha)$

\[
\begin{align*}
\frac{1}{2} \|\alpha\|^2 + C \sum_{r=1}^{n} \left[ -\alpha^T \phi(O^{(i)}, w_{\text{ref}}^{(i)}) + \max_{w \neq w_{\text{ref}}} \left\{ \mathcal{L}(w, w_{\text{ref}}^{(i)}) + \alpha^T \phi(O^{(i)}, w) \right\} \right] + \text{linear} + \text{convex}
\end{align*}
\]

(\xi_i \text{ is the part in } [\cdot]_+)
Structured SVM ≡ Large Margin Log Linear Model

- Example of log-linear models
  \[ P(w|O; \alpha) = \frac{1}{Z} \exp(\alpha^T \phi(O, w)) \]

- Margin: minimum distance between correct \( w_{\text{ref}} \) and competing \( w \)
  \[ \min_{w \neq w_{\text{ref}}} \left\{ \log \left( \frac{P(w_{\text{ref}}|O; \alpha)}{P(w|O; \alpha)} \right) \right\} \]

- Regularised, large margin criterion ≡ Structured SVM
  \[
  \frac{1}{2} ||\alpha||^2 + C \sum_{r=1}^{n} \left[ -\alpha^T \phi(O^{(i)}, w_{\text{ref}}^{(i)}) + \max_{w \neq w_{\text{ref}}} \{ L(w, w_{\text{ref}}^{(i)}) + \alpha^T \phi(O^{(i)}, w) \} \right] +
  \]
Best Segmentation

• Features depend on the segmentation $\theta$

• For efficiency, consider one “best” segmentation – How to define “best”?
  – Previously, $\phi(O, w) = \phi(O, w; \theta_{hmm})$
    Problem: Best segmentation in HMM may not the best in SSVM
  – What we want: $\max_{\theta} \alpha^T \phi(O, w; \theta)$
Can we optimize segmentation $\theta$ in Structured SVMs?

- How to learn $\alpha$ and $\theta$ jointly in training?
- How to find $w$ and $\theta$ in decoding?
Overall Training with Optimal Segmentation

• Previously, no variable $\theta$, $\min_\alpha$ (Convex Optimization)

$$\frac{1}{2}||\alpha||^2 + C \sum_{i=1}^{n} \left[ -\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}) + \max_{w \neq w_{\text{ref}}} \left\{ \mathcal{L}(w, w^{(i)}_{\text{ref}}) + \alpha^T \phi(O^{(i)}, w) \right\} \right]$$

• Now, Optimizing $\theta$ with $\alpha$ (Nonconvex Optimization)

$$-\max_{\theta} \alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}, \theta) + \max_{w \neq w_{\text{ref}}, \theta} \left\{ \mathcal{L}(w, w^{(i)}_{\text{ref}}) + \alpha^T \phi(O^{(i)}, w, \theta) \right\}$$
Structured SVMs for Noise Robust CSR

Overall Training with Optimal Segmentation

- Previously, no variable $\theta$, $\min_\alpha$ (Convex Optimization)

\[
\frac{1}{2} ||\alpha||_2^2 + C \sum_{i=1}^{n} \left[ \begin{array}{c}
\text{linear} \\
-\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}) + \max_{w \neq w_{\text{ref}}} \left\{ \mathcal{L}(w, w^{(i)}_{\text{ref}}) + \alpha^T \phi(O^{(i)}, w) \right\}
\end{array} \right]
\]

- Now, Optimizing $\theta$ with $\alpha$ (Nonconvex Optimization)

\[
\begin{aligned}
&\left[ \begin{array}{c}
\text{linear} \\
-\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}, \theta_{\text{ref}}) + \max_{w \neq w_{\text{ref}}, \theta} \left\{ \mathcal{L}(w, w^{(i)}_{\text{ref}}) + \alpha^T \phi(O^{(i)}, w, \theta) \right\}
\end{array} \right] + \\
&\alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}, \theta_{\text{ref}})
\end{aligned}
\]
Overall Training with Optimal Segmentation

- Algorithm (EM-style, guaranteed to converge):
  
  1. **Find reference segmentation:**
     \[
     \theta^{(i)}_{\text{ref}} \leftarrow \arg \max_{\theta} \alpha^T \phi(O^{(i)}, w^{(i)}_{\text{ref}}, \theta), \quad \forall i
     \]

  2. **Original SSVM training:**
     \[
     \alpha \leftarrow \min \mathcal{F}_{\text{ssvm}}(\alpha | \theta^{(i)}_{\text{ref}})
     \]

- Illustration of Algorithm (Concave-Convex Procedure):

![Illustration of Algorithm](image-url)
Decoding with Optimal Segmentation

• Decoding with optimal segmentation $\theta$

$$\arg \max_{w, \theta} \alpha^T \phi(O, w, \theta)$$

• In general, very difficult.
  – For our features, efficient search algorithm exists
Decoding with Optimal Segmentation

• Search optimal position of “black nodes”, and the label between them.

\[ \psi(t) = \max_{t_{st},w} \left\{ \psi(t_{st}) + \sum_{k=\text{"zero"}}^{\text{"one"}} \alpha_{k}^{(w)} \text{HMM}_k \right\} \]

  – Related to Factorial HMM inference
Implementation

• Overall framework:

- Loop 2 is parallelized training of original structured SVM
AURORA 2 Results

- Noise corrupted continuous digit task. SNRs from 0dB–20dB.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th>Decoding</th>
<th>Avg. all</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMM</td>
<td>—</td>
<td>—</td>
<td>9.5</td>
</tr>
<tr>
<td>Structured SVM</td>
<td>$\theta_{\text{hmm}}$</td>
<td>$\theta_{\text{hmm}}$</td>
<td>7.8</td>
</tr>
<tr>
<td></td>
<td>$\theta_{\text{hmm}}$</td>
<td>opt $\theta$</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>opt $\theta$</td>
<td>opt $\theta$</td>
<td>7.4</td>
</tr>
</tbody>
</table>

- **Structured SVMs with optimal segmentation:**
  - Gain over VTS-compensated HMMs: **22%**
  - Gain over Multi-Class SVMs: **10%**
  - Gain over original Structured SVMs: **4%**
Conclusion

• Examined Structured SVMs for noise robust speech recognition
  – equivalent to Large Margin Log Linear Models

• Enabled optimising segmentation in Structured SVMs *training*
  – use concave-convex procedure

• Enabled optimising segmentation in Structured SVMs *decoding*
  – proposed Viterbi-style efficient search

• Good Performance