Ch8. Deep Neural Network Sequence-Discriminative Training

Dong Yu, Li Deng

Speaker: 孝宗
Outline

• 8.1 Sequence-Discriminative Training Criteria
• 8.2 Practical Consideration
• 8.3 Noise Contrastive Estimation
Training Criteria

- The model parameters should be trained to minimize the expected loss

\[ J_{EL} = \mathbb{E}(J(W, b; o, y)) = \int J(W, b; o, y)p(o)d(o) \]

  - Model parameters: \{W, b\}
  - Observation-label pair: \{o, y\}

- Mean square error, MSE (regression task)
- Cross-entropy criterion, CE (classification task)
- Negative log-likelihood, NLL
Cross-Entropy

\[ J_{CE}(\mathbf{W}, \mathbf{b}; \mathcal{S}) = \frac{1}{M} \sum_{m=1}^{M} J_{CE}(\mathbf{W}, \mathbf{b}; \mathbf{o}^m, \mathbf{y}^m) \]

where

\[ y_i = P_{empirical}(i|\mathbf{o}) \] (observed in the training set)

\[ v_i^L = P_{dnn}(i|\mathbf{o}) \]

\[ e_t^L = \nabla_{z_t^L} J_{CE}(\mathbf{W}, \mathbf{b}; \mathbf{o}, \mathbf{y}) = - \frac{\partial}{\partial z_t^L} \sum_{i=1}^{C} y_i \log \text{softmax}(z_t^L) = (v_t^L - y) \]
Negative log-likelihood

\[ J_{CE}(\mathbf{W}, \mathbf{b}; S) = - \sum_{i=1}^{c} y_i \log v_i^L \]

\[ y_i = \mathbb{I}(x) = \begin{cases} 1, & \text{if } x \text{ is true} \\ 0, & \text{otherwise} \end{cases} \]

例: \( \mathbf{y} = [0,0,0,0,1,0,0,0] \)

\[ \iff \quad J_{NLL}(\mathbf{W}, \mathbf{b}; S) = - \log v_c^L \]
Sequence-Discriminative Training Criteria

• “Sequence”
  • Speech recognition is a sequence classification problem

• “Discriminative”
  • minimizing the classification error probability
    = maximum a posterior

• Match the maximum a posterior decision rule of LVCSR
  • Considering sequence constrain from HMM, dictionary, language model
Maximum Mutual Information

- Maximizing the mutual information between the distribution of the observation sequence and the word sequence

\[
I(X; Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}
\]

\[
= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} - \sum_{x,y} p(x,y) \log p(y)
\]

\[
= \sum_{x,y} p(x)p(y|x) \log p(y|x) - \sum_{x,y} p(x,y) \log p(y)
\]

\[
= \sum_x p(x) \left( \sum_y p(y|x) \log p(y|x) \right) - \sum_y \log p(y) \left( \sum_x p(x,y) \right)
\]

\[
= -\sum_x p(x) H(Y|X = x) - \sum_y \log p(y)p(y)
\]

\[
= -H(Y|X) + H(Y)
\]

\[
= H(Y) - H(Y|X).
\]

\[
I(O, W) = H(W) - H(W|O)
\]

忽略\( H(W) \)

\[
H(W|O) = \text{已知O時W的混亂程度}
\]

\( H: \text{entropy} \)
Maximum Mutual Information

\[
J_{\text{MMI}} (\theta; \mathcal{S}) = \sum_{m=1}^{M} J_{\text{MMI}} (\theta; o^m, w^m)
\]

\[
= \sum_{m=1}^{M} \log P(w^m | o^m; \theta)
\]

\[
= \sum_{m=1}^{M} \log \frac{p(o^m | s^m; \theta)^k P(w^m)}{\sum_w p(o^m | s^w; \theta)^k P(w)}
\]

- \(s^m\): state
- In practice, however, the sum is constrained by the decoded speech lattice for utterance \(m\) to reduce the computational cost.
Maximum Mutual Information

• Gradient

\[ \nabla_{\theta} J_{MMI}(\theta; o^m, w^m) = \sum_m \sum_t \nabla_{z_{mt}} J_{MMI}(\theta; o^m, w^m) \frac{\partial z^L_{mt}}{\partial \theta} \]

\[ = \sum_m \sum_t \hat{e}^L_{mt} \frac{\partial z^L_{mt}}{\partial \theta} \]

Utterance m, frame t

\[ \hat{e}^L_{mt} = \nabla_{z_{mt}} J_{MMI}(\theta; o^m, w^m) \]

\( z^L_{mt} \): softmax layer’s excitation (the value before softmax is applied)

\( z^L_{mt} \) is irrelevant to the training criterion.
Maximum Mutual Information

- $z_{mt}^L$ is irrelevant to the training criterion.
- The only difference the new training criterion introduces compared to the frame-level cross-entropy is $\bar{e}_{mt}^L$

$$
\bar{e}_{mt}^L(i) = \nabla_{z_{mt}(i)} J_{MMI}(\theta; o^m, w^m)
$$

$$
= \sum_r \frac{\partial J_{MMI}(\theta; o^m, y^m)}{\partial \log p(o_t^m | r)} \frac{\partial \log p(o_t^m | r)}{\partial z_{mt}^L(i)}
$$

$$
= \sum_r k \left( \delta(r = s_t^m) - \frac{\sum_{w:s_t=r} p(o^m|s)^k P(w)}{\sum_w p(o^m|s^w)^k P(w)} \right) \times \frac{\partial \log P(r|o_t^m) - \log P(r) + \log p(o_t^m)}{\partial z_{mt}^L(i)}
$$

$$
= \sum_r k \left( \delta(r = s_t^m) - \bar{\gamma}_{mt}^{DEN}(r) \right) \times \frac{\partial \log v_{mt}^L(r)}{\partial z_{mt}^L(i)}
$$

$$
= k \left( \delta(i = s_t^m) - \bar{\gamma}_{mt}^{DEN}(i) \right)
$$

$\delta(i = s_t^m)$: force-alignment

$\bar{\gamma}_{mt}^{DEN}(r)$: the posterior probability of being in state $r$ at time $t$, computed over the denominator lattices for utterance $m$
Boosted MMI

- To boost the likelihood of paths that contain more errors

\[ J_{BM}^{MMI}(\theta; S) = \sum_{m=1}^{M} J_{BM}^{MMI}(\theta; o^m, w^m) \]

\[ = \sum_{m=1}^{M} \log \frac{p(o^m|s^m; \theta)^k P(w^m)}{\sum_w p(o^m|s^w; \theta)^k P(w) e^{-bA(w,w^m)}} \]

\[ \gamma_{mt}^{DEN}(r) = \frac{\sum_{w:s_t=r} p(o^m|s)^k P(w) e^{-bA(w,w^m)}}{\sum_w p(o^m|s^w)^k P(w) e^{-bA(w,w^m)}} \]
MBR

• The MBR family of objective functions aims at minimizing the expected error corresponding to difference granularity of labels

\[
J_{MBR}(\theta; S) = \sum_{m=1}^{M} J_{MBR}(\theta; o^m, w^m)
\]

\[
= \sum_{m=1}^{M} \sum_{w} P(w|o^m)A(w, w^m)
\]

\[
= \sum_{m=1}^{M} \frac{\sum_{w} p(o^m|s^w; \theta)^k P(w)A(w, w^m)}{\sum_{w'} p(o^m|s^{w'}; \theta)^k P(w')}
\]

W=state => sMBR
W=phone => MPE
W=word => MWE
MBR

\[ \ddot{e}_{mt}(i) = \nabla_{z_{mt}(i)}^L J_{MBR}(\theta; o^m, w^m) \]

\[ = \sum_r \frac{\partial J_{MBR}(\theta; o^m, y^m)}{\partial \log p(o^m_t|r)} \frac{\partial \log p(o^m_t|r)}{\partial z^L_{mt}(i)} \]

\[ = \sum_r k \left( (\bar{A}^m(r = s^m_t) - \bar{A}^m) \dot{y}_{mt}^{DEN}(r) \right) \times \frac{\partial \log v^L_{mt}(r)}{\partial z^L_{mt}(i)} \]

\[ = k \left( (\bar{A}^m(i = s^m_t) - \bar{A}^m) \ddot{y}_{mt}^{DEN}(i) \right) \]

\( \bar{A}^m \): average accuracy of all paths in the lattice
\( \bar{A}^m(i = s^m_t) \): passes through state \( r \) at time \( t \)
sMBR

\[ \ddot{e}_{mt}^{L}(i) = \nabla_{z_{mt}(i)} I_{sMBR}(\theta; o^m, w^m) \]

\[ = k \left( (\bar{A}^m(i = s_t^m) - \bar{A}^m) \dot{y}_{mt}^{\text{DEN}}(i) \right) \]

\[ \dot{y}_{mt}^{\text{DEN}}(i) = \sum_s \delta(r = s_t)P(s | o^m) \]

\[ A(w, w^m) = A(s^w, s^m) = \sum_t \delta(s_t^w = s_t^m) \]

\[ \bar{A}^m(r = s_t^m) = E\{A(s, s^m) | s_t = r\} = \frac{\sum_s \delta(r = s_t) P(s | o^m) A(s, s^m)}{\sum_s \delta(r = s_t) P(s | o^m)} \]

\[ \bar{A}^m = E\{A(s, s^m)\} = \frac{\sum_s P(s | o^m) A(s, s^m)}{\sum_s P(s | o^m)} \].
A Uniformed Formulation

- Loss functions can always be formulated as a ratio of values computed from two lattices
  - numerator lattice => the reference transcription (one path)
  - denominator => competing hypotheses

- The expected occupancies $\hat{\gamma}_m^\text{NUM}(i)$ and $\hat{\gamma}_m^\text{DEN}(i)$ for each state $i$ required by the extended Baum-Welch (EBW) algorithm
  - forward-backward
A Uniformed Formulation

• the gradient of the loss with respect to state log-likelihood is

\[
\frac{\partial J_{SEQ}(\theta; o^m, w^m)}{\partial \log p(o^m_t | r)} = k \left( \ddot{\gamma}_{mt}^{\text{DEN}}(r) - \ddot{\gamma}_{mt}^{\text{NUM}}(r) \right)
\]

• To posterior

\[
\frac{\partial J_{SEQ}(\theta; o^m, w^m)}{\partial P(r|o^m_t)} = \frac{\partial J_{SEQ}(\theta; o^m, w^m)}{\partial \log p(o^m_t | r)} \frac{\partial \log p(o^m_t | r)}{\partial \log P(r|o^m_t)} \frac{\partial \log P(r|o^m_t)}{\partial P(r|o^m_t)} = k \left( \ddot{\gamma}_{mt}^{\text{DEN}}(r) - \ddot{\gamma}_{mt}^{\text{NUM}}(r) \right) \frac{P(r|o^m_t)}{P(r|o^m_t)}
\]

Given that \( P(r|o^m_t) = \text{softmax}_r (z^L_{mt}) \) we get

\[
e^L_{mt}(i) = \frac{\partial J_{SEQ}(\theta; o^m, w^m)}{\partial P(r|o^m_t)} \frac{\partial P(r|o^m_t)}{\partial z^L_{mt}} = k \left( \ddot{\gamma}_{mt}^{\text{DEN}}(i) - \ddot{\gamma}_{mt}^{\text{NUM}}(i) \right)
\]

\[
e^L_t = \nabla_{z^L_t} J_{CE} (W, b; o, y) = - \frac{\partial \sum_{i=1}^{C} y_i \log \text{softmax}(z^L_t)}{\partial z^L_t} = (v^L_t - y)
\]

和原本的CE比較
A Uniformed Formulation

Compared with CE...

\[
e^L_{mt}(i) = \frac{\partial J_{SEQ}(\theta; o^m, w^m)}{\partial z^L_{mt}} = k \left( \hat{y}^{DEN}_{mt}(i) - \hat{y}^{NUM}_{mt}(i) \right)
\]

\[
e^L_t = \nabla_{z^L_t} J_{CE}(W, b; o, y) = - \frac{\partial \sum_{i=1}^{C} y_i \log softmax(z^L_t)}{\partial z^L_t} = (v^L_t - y)
\]

“Sequence” \(\Leftrightarrow\) state posterior on the lattices

numerator lattice => the reference transcription (one path)
denominator => competing hypotheses

Compared with RBM CD-1 training...

✓ Positive and negative sample of \(y\)

☐ Negative sample of \(o\)
Training Criterion Selection

- sMBR slightly outperforms other criteria
- it is thus suggested to use MMI if you need to implement one from scratch

| Table 8.2  The effect of different sequence-discriminative training criteria measured as WER on the Hub5’00 and Hub5’01 datasets when the 300h training set is used. (Summarized from [16]) |
|---------------------------------|-----------------|-----------------|
|                                | Hub5’00 SWB (%) | Hub5’01 SWB (%) |
| GMM BMMI                        | 18.6            | 18.9            |
| DNN CE                          | 14.2            | 14.5            |
| DNN MMI                         | 12.9            | 13.3            |
| DNN BMMI                        | 12.9            | 13.2            |
| DNN MPE                         | 12.9            | 13.2            |
| DNN sMBR                        | 12.6            | 13.0            |
Outline

• 8.1 Sequence-Discriminative Training Criteria
• 8.2 Practical Consideration
• 8.3 Noise Contrastive Estimation