Sequence training; "HMM-Free" ASR

Steve Renals

Automatic Speech Recognition – ASR Lecture 18
17 March 2016

Training HMM/GMM acoustic models

- Use forward-backward algorithm to estimate the state occupation probabilities (E-step), which are used to re-estimate the parameters (M-step)
- Maximum likelihood estimation: estimate the parameters so that the model reproduces the training data with the greatest probability (maximum likelihood)
- Discriminative training: directly estimate the parameters so as to make the fewest classification errors (optimize the word error rate)
  - Focus on learning *boundaries* between classes
  - Consider incorrect word sequences as well as correct word sequences
  - This is related to direct optimisation of the posterior probability of the words given the acoustics $P(W \mid \mathbf{X})$

Hybrid HMM/NN acoustic models

- Neural networks are discriminatively trained at the **frame** level
- Consider a context-dependent DNN
  - Output is a softmax over HMM states
  - Training involves increasing the probability of the correct state – and hence decreasing the probabilities of the others, since probabilities sum to 1
  - Frame-level discrimination – the network learns to optimise discrimination at the frame level by choosing the best state at each time frame
- **Sequence discrimination** – train the system to select the best sequence of frames by increasing the probability of the best sequence and decreasing the probability of all competing sequences
- Can train both GMM and DNN based models using sequence discrimination

Hybrid HMM/NN acoustic models

- Neural networks are discriminatively trained at the **frame** level
- Consider a context-dependent DNN
  - Output is a softmax over HMM states
  - Training involves increasing the probability of the correct state – and hence decreasing the probabilities of the others, since probabilities sum to 1
  - Frame-level discrimination – the network learns to optimise discrimination at the frame level by choosing the best state at each time frame
- **Sequence discrimination** – train the system to select the best sequence of frames by increasing the probability of the best sequence and decreasing the probability of all competing sequences
- Can train both **GMM** and DNN based models using sequence discrimination

Recall: Maximum likelihood estimation (MLE)

- Maximum likelihood estimation (MLE) sets the parameters so as to maximize an objective function F_{MLE}:

$$F_{\text{MLE}} = \sum_{u=1}^{U} \log P_\lambda(\mathbf{X}_u \mid M(W_u))$$

  for training utterances $\mathbf{X}_1 \ldots \mathbf{X}_U$ where W_u is the word sequence given by the transcription of the uth utterance, $M(W_u)$ is the corresponding HMM, and λ is the set of HMM parameters

Maximum mutual information estimation

- Maximum mutual information estimation (MMIE) aims to directly maximise the posterior probability (sometimes called conditional maximum likelihood). Using the same notation as before, with $P(w)$ representing the language model probability of word sequence w:

$$
\begin{aligned}
F_{\text{MMIE}} &= \sum_{u=1}^{U} \log P_\lambda(M(W_u) \mid \mathbf{X}_u) \\
&= \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u)) P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w')) P(w')}
\end{aligned}
$$

Maximum mutual information estimation

- Maximum mutual information estimation (MMIE) aims to directly maximise the posterior probability (sometimes called conditional maximum likelihood). Using the same notation as before, with $P(w)$ representing the language model probability of word sequence w:

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log P_\lambda(M(W_u) \mid \mathbf{X}_u)$$

$$F_{\text{MLE}} = \sum_{u=1}^{U} \log \frac{P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)}{\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'))P(w')}$$

Maximum mutual information estimation

- **Numerator**: $P_\lambda(\mathbf{X}_u \mid M(W_u))P(W_u)$
  the likelihood of the data given the correct word sequence
- **Denominator**: $\sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'_u))P(w'_u)$
  the total likelihood of the data given all possible word sequences – obtained by summing over all possible word sequences estimated by the full acoustic and language models in recognition (M_{den}):

$$P(\mathbf{X} \mid M_{\text{den}}) = \sum_{w'} P_\lambda(\mathbf{X}_u \mid M(w'_u))P(w'_u)$$

  Estimate by generating lattices, and summing over all words in the lattice
- The objective function F_{MMIE} is optimised by making the correct word sequence likely (maximise the numerator), and all other word sequences unlikely (minimise the denominator)

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

$$F_{\mathrm{MPE}} = \sum_{u=1}^{U} \log \frac{\sum_{W} P_\lambda(\mathbf{X}_u \mid M(W)) P(W) A(W, W_u)}{\sum_{W'} P_\lambda(\mathbf{X}_u \mid M(W')) P(W')}$$

- $A(W, W_u)$ is the phone transcription accuracy of the sentence W given the reference W_u

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

$$F_{\text{MMIE}} = \sum_{u=1}^{U} \log \frac{\sum_{W} P_\lambda(\mathbf{X}_u \mid M(W_{\underline{u}}))P(W_{\underline{u}})A(W, W_u)}{\sum_{W'} P_\lambda(\mathbf{X}_u \mid M(W'))P(W')}$$

- $A(W, W_u)$ is the phone transcription accuracy of the sentence W given the reference W_u

MPE: Minimum phone error

- Basic idea adjust the optimization criterion so it is directly related to word error rate
- Minimum phone error (MPE) criterion

$$F_{\text{MPE}} = \sum_{u=1}^{U} \log \frac{\sum_W P_\lambda(\mathbf{X}_u \mid M(W))P(W)A(W, W_u)}{\sum_{W'} P_\lambda(\mathbf{X}_u \mid M(W'))P(W')}$$

- $A(W, W_u)$ is the phone transcription accuracy of the sentence W given the reference W_u
- F_{MPE} is a weighted average over all possible sentences w of the raw phone accuracy
- Although MPE optimizes a phone accuracy level, it does so in the context of a word-level system: it is optimized by finding probable sentences with low phone error rates

Example: meeting speech recognition

WER for HMM/GMM system

System	Training criterion	WER/%
Baseline	ML	28.7
SAT	ML	27.6
SAT	MPE	24.5

Sequence training of hybrid HMM/DNN systems

- Can train HMM/NN systems using a MMI-type objective function (e.g. Bridle and Dodd, 1991)
- Forward- and back-propagation equations are structurally similar to forward and backward recursions in HMM training
- Was not used in practice, for another 20 years...
- Now used for DNN systems (e.g. Vesely et al, 2013)
- The tricky parts are in the optimisation and in the use of lattices to compute the denominator term...

Limitations of HMMs

- Sequence trained HMM/NN systems have limitations
  - Markov assumption – current state depends on only the previous state
  - Conditional independence assumptions – dependence on previous acoustic observations encapsulated in the current state
- RNNs are powerful sequence models
  - recurrent hidden state much richer history representation than HMM state
  - can learn representations
  - can directly model dependences through time
- But HMM/RNN systems only use RNNs to model time within a phone / HMM state...

“End-to-end” (“HMM-Free”) RNN speech recognition

- **Can RNNs replace the HMM sequence model?**
- Yes – active research topic. On approach is to use an **RNN encoder-decoder** model
- The **encoder** maps the the input sequence vector into a sequence of RNN hidden states
- The **decoder** maps the RNN hidden states into an output sequence
- Input and output sequences may be different lengths
  - Input sequence of frames
  - Output sequence of phones or letters or words!
- Mapping to directly to words results in a joint acoustic and language model

RNN Encoder-Decoder

- The overall task is to compute the probability of an output sequence given an input sequence, $P(\mathbf{y}_1, \ldots, \mathbf{y}_O | \mathbf{x}_1, \ldots, \mathbf{x}_T) = P(\mathbf{y}_1^O | \mathbf{x}_1^T)$
- **Encoder:** compute a *context* $\mathbf{c}_o$ for each output $\mathbf{y}_o$
- **Decoder:** compute

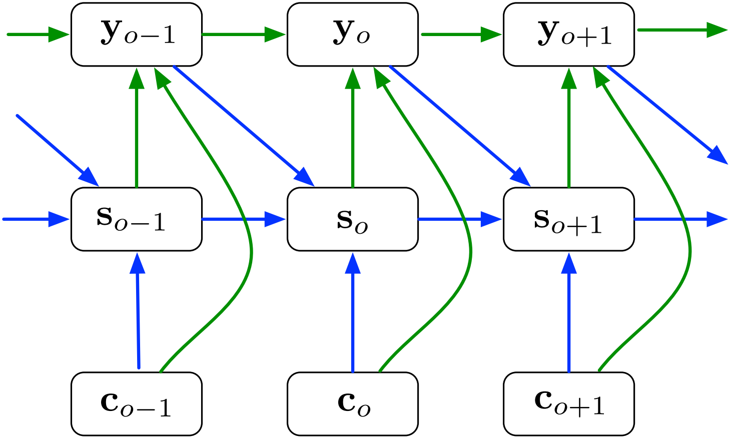
$$P(\mathbf{y}_1^O | \mathbf{x}_1^T) = \prod_o \underbrace{P(\mathbf{y}_o | \mathbf{y}_1^{o-1}, \mathbf{c}_o)}_{\text{RNN}}$$

$$P(\mathbf{y}_o | \mathbf{y}_1^{o-1}, \mathbf{c}_o) = \mathrm{softmax}(\mathbf{y}_{o-1}, \mathbf{s}_o, \mathbf{c}_o)$$

$$\mathbf{s}_o = f(\mathbf{y}_{o-1}, \mathbf{s}_{o-1}, \mathbf{c}_o)$$

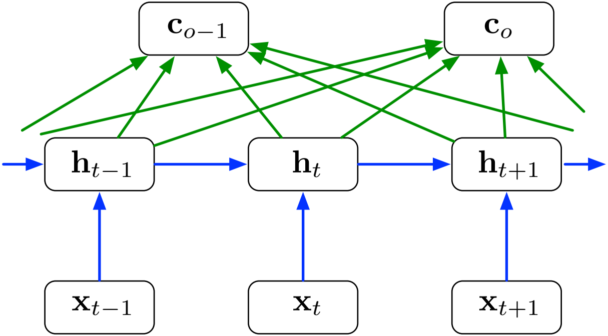
  - $\mathbf{y}_{o-1}$ is the previous output
  - $\mathbf{s}_o$ is the decoder state (recurrent hidden layer)
  - $\mathbf{c}_o$ is the encoder context

RNN decoder



$\mathbf{y}_{o-1}$
$\mathbf{y}_{o}$
$\mathbf{y}_{o+1}$
$\mathbf{s}_{o-1}$
$\mathbf{s}_{o}$
$\mathbf{s}_{o+1}$
$\mathbf{c}_{o-1}$
$\mathbf{c}_{o}$
$\mathbf{c}_{o+1}$

RNN encoder



$$\mathbf{c}_o = \sum_t \alpha_{ot} \mathbf{h_t}$$

$$\alpha_{ot} = \underbrace{\mathrm{softmax}(g(\mathbf{s}_{o-1}, \mathbf{h}_t))}_{\text{NN}}$$

RNN encoder-decoder

- Train all the parameters to maximise $\log P(\mathbf{y}_1^O | \mathbf{x}_1^T)$ using backprop through time
- The encoder is a bidirectional RNN
- Training/testing on Switchboard, directly mapping MFCCs to words (no pronunciation model, no language model) gives 49% WER
- Improved training scheme, FBANK features gives 37% WER
- Potential improvements
  - multiple recurrent layers in the encoder
  - incorporating a language model in the decoder
  - using character-based output sequence

Reading

- HMM discriminative training: Sec 27.3.1 of: S Young (2008), “HMMs and Related Speech Recognition Technologies”, in *Springer Handbook of Speech Processing*, Benesty, Sondhi and Huang (eds), chapter 27, 539–557. http://www.inf.ed.ac.uk/teaching/courses/asr/2010-11/restrict/Young.pdf
- NN sequence training: K Vesely et al (2013), “Sequence-discriminative training of deep neural networks”, Interspeech-2013, http://homepages.inf.ed.ac.uk/aghoshal/pubs/is13-dnn_seq.pdf
- RNN encoder-decoder: L Lu et al (2015), “A Study of the Recurrent Neural Network Encoder-Decoder for Large Vocabulary Speech Recognition”, Interspeech-2015, http://homepages.inf.ed.ac.uk/llu/pdf/liang_is15a.pdf