On Architectural Issues of Neural Networks in Speech Recognition

Hermann Ney
Human Language Technology and Pattern Recognition
RWTH Aachen University, Aachen, Germany

IEEE Distinguished Lecturer 2016/17
Human Language Technology (HLT)

Automatic Speech Recognition (ASR)

Statistical Machine Translation (SMT)

Handwriting Recognition
(Text Image Recognition)

tasks:
– speech recognition
– machine translation
– handwriting recognition

unifying view:
– input string → output string
– output string: natural language

we want to preserve this great idea
RWTH’s Joint Projects with InterACT: KIT, CMU or HKUST

- VERBMOBIL 1993-2000: funded by German BMBF
toy task (8000-word vocabulary): recognition and translation for appointment scheduling

- TC-STAR 2004-2007: funded by EU
  – real-life task, open domain, large vocabulary:
    first research system for speech translation (EU parliament)
  – partners: KIT Karlsruhe, FBK Trento, LIMSI Paris, UPC Barcelona, IBM-US Research, ...

- GALE 2005-2011: funded by US DARPA
  – emphasis on Chinese and Arabic speech and text
  – largest project ever on speech and language: 40 Mio USD per year

- BOLT 2011-2015: funded by US DARPA
  emphasis on colloquial text for Arabic and Chinese

- QUAERO 2008-2013: funded by OSEO France
  European languages, more colloquial speech, handwriting

- EU-BRIDGE 2012-2014: funded by EU
  emphasis on recognition and translation of lectures (TED, ...)

- BABEL 2012-2016: funded by US IARPA
  speech recognition for low-resource languages (and noisy audio!)
evaluations of ASR and SMT systems:

- project related evaluations:
  - VERBMOBIL
  - TC-STAR
  - QUAERO
  - EU-BRIDGE

- public evaluation campaigns:
  - NIST/LDC/DARPA
  - IWSLT (organized by InterACT members)
  - ACL WMT

- joint submissions with KIT/InterACT:
  system combination
Statistical Approach: No Alternative
(incl. Artificial Neural Networks!)

- Performance Measure (Loss Function)
- Probabilistic Models
  - Training Criterion
  - Optimization (Efficient Algorithm)
  - Bayes Decision Rule (Efficient Algorithm)
- Evaluation
- Training Data
- Test Data

H. Ney: Architecture ANN for ASR ©RWTH
InterACT25, KIT, Baden-Baden, 14/15-Jul-2016
Hidden Markov Models (HMM)

- fundamental problem in ASR: non-linear time alignment
- Hidden Markov Model:
  - linear chain of states \( s = 1, \ldots, S \)
  - transitions: forward, loop and skip
- trellis:
  - unfold HMM over time \( t = 1, \ldots, T \)
  - path: state sequence \( s_T^T = s_1...s_t...s_T \)
  - observations: \( x_T^T = x_1...x_t...x_T \)

general view:
- two sequences without synchronization: acoustic vectors and states (with labels)
- HMM: mechanism that takes care of the synchronization (=alignment) problem
Hidden Markov Models (HMM)

The acoustic model $p(X|W)$ provides the link between word sequence hypothesis $W$ and observations sequence $X = x_1^T = x_1...x_t...x_T$:

- acoustic probability $p(x_1^T|W)$ using hidden state sequences $s_1^T$:

$$p(x_1^T|W) = \sum_{s_1^T} p(x_1^T, s_1^T|W) = \sum_{s_1^T} \prod_t [p(s_t|s_{t-1}, W) \cdot p(x_t|s_t, W)]$$

- two types of distributions:
  - transition probability $p(s|s', W)$: not important
  - emission probability $p(x_t|s, W)$: key quantity
    realized by GMM: Gaussian mixtures models (trained by EM algorithm)

- phonetic labels (allophones, sub-phones): $(s, W) \rightarrow a = a_{sW}$

$$p(x_t|s, W) = p(x_t|a_{sW})$$

typical approach: phoneme models in triphone context:
decision trees (CART) for finding equivalence classes

- refinements:
  - augmented feature vector: context window around position $t$
  - subsequent LDA (linear discriminant analysis)
Hybrid Approach: HMM and ANN

consider modelling the acoustic vector \( x_t \) in an HMM:

- re-write the emission probability for annotation label \( a \) and acoustic vector \( x_t \)
  (strictly speaking: an approximation only):
  \[
  p(x_t|a) = p(x_t) \cdot \frac{p(a|x_t)}{p(a)}
  \]

  - prior probability \( p(a) \): estimated as relative frequencies
  - for recognition purposes: the term \( p(x_t) \) can be dropped

- result: model the label posterior probability by an ANN:
  \[
  x_t \rightarrow p(a|x_t)
  \]

  rather than the state emission distribution \( p(x_t|a) \)

- justification:
  - easier learning problem: labels \( a = 1, \ldots, 5000 \) vs. vectors \( x_t \in \mathbb{R}^{D=40} \)
  - well-known result in pattern recognition/machine learning;
    but ignored in ASR due to the mathematical beauty of the EM algorithm
History: ANN in Acoustic Modelling

- 1989 [Bridle 1989]: softmax operation for probability normalization in output layer
- 1990 [Bourlard & Wellekens 1990]:
  - for squared error criterion, ANN outputs can be interpreted as class posterior probabilities (rediscovered: Patterson & Womack 1966)
  - they advocated the hybrid approach: use the ANN outputs to replace the emission probabilities in HMMs
- 1993 [Haffner 1993]: sum over label-sequence posterior probabilities in hybrid HMMs
- 1994 [Robinson 1994]: recurrent neural network
  - competitive results on WSJ task
  - his work remained a singularity in ASR

Experimental situation:
- until 2011: ANNs were never really competitive with Gaussian mixture models
- after 2011: yes, deep learning [Deng & Hinton 2012]
more ANN approaches:

- **1994** [LeCun & Bengio \(^+\) 94]: convolutional neural networks
- **1997** A. Waibel’s team [Fritsch & Finke \(^+\) 97]: hierarchical mixtures of experts
- **1997** [Hochreiter & Schmidhuber 97]: long short-term memory neural computation with extensions [Gers & Schraudolph \(^+\) 02]

renaissance of ANN: concepts of deep learning and related ideas:

- **2000** [Hermansky & Ellis \(^+\) 00]: tandem approach: multiple layers of processing by combining Gaussian model and ANN for ASR
- **2002** [Utgoff & Stracuzzi 02]: many-layered learning for symbolic processing
- **2006** [Hinton & Osindero \(^+\) 06]: introduced what he called *deep learning (belief nets)*
- **2008** [Graves 08]: good results on LSTM RNN for handwriting task
- **2012** Microsoft Research [Dahl & Yu \(^+\) 12]:
  - combined Hinton’s deep learning with hybrid approach
  - significant improvement by deep MLP on a large-scale task
- since **2012**: other teams confirmed significant reductions of WER
TDNN: Time Delay Neural Network
[Waibel & Hanazawa+ 88]

TDNN: feed-forward multi-layer perceptron with special properties:
– long temporal context
– weight sharing
TDNN: Time Delay Neural Network

- first (?) publication: [Waibel & Hanazawa$^+$ 88] at ICASSP 1988, New York
  - 2036 citations (Google Scholar)
  - 1116 citations on 3 more papers on TDNN 1989/90
- recent work by Dan Povey’s team [Peddinti & Povey$^+$ 15] at Interspeech 2015:
  improvements over widely used deep MLP approach
  - on many of the standard ASR tasks (WSJ, Switchboard, Librispeech, ...)
  - on ASPIRE challenge (IARPA, March 2015):
    reverberant speech in farfield speech recognition
Today vs. 1988-94: What is Different?

most popular and widely used:
  feed-forward multi-layer perceptron (FF MLP)
  – operations: matrix \cdot vector
  – nonlinear activation function

comparison for ASR: today vs. 1988-1994:
  • number of hidden layers:
    10 (or more) rather than 2-3
  • number of output nodes (phonetic labels):
    5000 rather than 50
  • optimization strategy:
    practical experience and heuristics,
    e.g. layer-by-layer pretraining
  • much more computing power

overall result:
  – huge improvement by ANN
  – WER is (nearly) halved !!
Recurrent Neural Network: String Processing

principle for string processing over time \( t = 1, \ldots, T \):
– introduce a memory (or context) component to keep track of history
– quantities: input = observation \( x_t \), memory \( h_{t-1} \), output distribution \( y_t \)

extensions:
– bidirectional variant [Schuster & Paliwal 1997]
– feedback of output labels
– long short-term memory [Hochreiter & Schmidhuber 97; Gers & Schraudolph\(^+\) 02]
– deep structure: several hidden layers
Direct Model of Label Sequence  
(spirit of CTC: connectionist temporal classification)

re-formulate the problem of speech recognition:

• sequence of phonetic labels (e.g. CART): $a_s, s = 1, \ldots, S$ 
  (which fully determines the sequence of words)

• key quantity: (local) label posterior probability calculated by an ANN

$$p_t(a|x_t^T) = p_t(a|x_{t+\delta})$$

• model localization effect by alignments, i.e. mappings from time to states:

$$t \rightarrow s = s_t$$
Direct Model of Label Sequence

sum over all hidden alignments $s_1^T$:

$$p(a_1^S|x_1^T) = \sum_{s_1^T} p(a_1^S, s_1^T|x_1^T) = \ldots$$

$$= \sum_{s_1^T} \prod_t p_t(a_{st}|x_t^T) = \sum_{s_1^T} \prod_t p_t(a_{st}|x_{t+\delta})$$

open issues:
- how to include the transition probabilities
- how to include the language model
- how to perform end-to-end training

requirement:
- avoid the global re-normalization as in discriminative/hybrid HMM
Comparison with Discriminative/Hybrid HMM

- **topology:**
  conventional HMM structure

- **important differences:**
  - no joint model \( p(a_1^S, x_1^T) \)
  - no global re-normalization (e.g. lattice)

- **open issues:**
  - transition probabilities
  - language model
  - consistent training criterion:
    sum over all alignments, end-to-end training, ...

---

goal: avoid joint probability \( p(a_1^S, x_1^T) \) as in discriminative/hybrid HMM
Comparison with CTC: connectionist temporal classification
[Graves & Fernandez+ 06]

characteristic properties of CTC:
- topology: for each symbol label: single state + blank state
- no transition probabilities
- training criterion: sum
- ANN structure: LSTM RNN or ...?

experiments for CTC and related neural network approaches:
- good results reported
- reason: LSTM RNN?
- direct comparison: to be done
Direct Model of Label Sequence: Inverted Alignments

– re-interpretation of ASR: segmentation and classification problem
– consider inverted alignments, i.e. from state $s$ to time $t$:
  $$s \rightarrow t = t_s$$
– sum over inverted alignments as hidden variables $t_1^S$:

$$p(a_1^S|x_1^T) = \sum_{t_1^S} p(a_1^S, t_1^S|x_1^T) = \ldots =$$

$$= \sum_{t_1^S} \prod_{s=1}^{S} p_{t_s}(a_s|x_1^{t_s}) = \sum_{t_1^S} \prod_{s=1}^{S} p_{t_s}(a_s|x_1^{t_s+\delta})$$

experiments: underway
Mechanism of Attention: Alignment by ANN
(originally introduced for MT [Bahdanau & Cho+ 15])

mechanism of attention:
ANN only

alignment direction:
from state $s$ to time $t$

occupation probabilities:
$\alpha(t|s)$

experiments:
going on work, many teams
Architectural Issues of ANN in ASR Systems:

- starting point: direct model of label sequences:
  - use ANN output as label posterior probability
  - (try to) avoid global re-normalization (no denominator/lattice)

- open questions:
  - how to include transition probabilities?
  - how to include language model?
  - end-to-end training: suitable training criterion

- some localization is needed: alignments
  - inverted alignments vs. traditional alignments
  - attention-based mechanism: alternative?

- experimental results: room for improvements
  - a large number of ongoing studies
  - clear conclusions: difficult
Congratulations to InterACT and Alex on 25 successful years!

Best wishes for the coming 25 years!
REFERENCES
References


