Inverted HMM - a Proof of Concept

Patrick Doetsch, Stefan Hegselmann, Ralf Schlüter, Hermann Ney
Lehrstuhl für Informatik 6
RWTH Aachen University
Ahornstr. 55, 52056 Aachen
{doetsch,schlueter,ney}@cs.rwth-aachen.de

Abstract

In this work, we propose an *inverted* hidden Markov model (HMM) approach to automatic speech and handwriting recognition that naturally incorporates discriminative, artificial neural network based label distributions. Instead of aligning each input frame to a state label as in the standard HMM derivation, we propose to inversely align each element of an HMM state label sequence to a single input frame. This enables an integrated discriminative model that may be trained end-to-end from scratch or starting from an existing alignment path. The approach does not assume the usual decomposition into a separate (generative) acoustic model and a language model, and allows for a variety of model assumptions, incl. statistical variants of attention. Here, an initial proof-of-concept with experiments on the RIMES handwritten word recognition task is provided. For this initial experiment, model assumptions for the inverted HMM were chosen that are similar to those for a standard (hybrid) HMM and a similar LSTM-based network design was used. The experimental results show that this initial approach already performs similar to hybrid HMM, CTC, and a sequence-to-sequence approach using attention.

1 Introduction

Training and decoding in sequence classification systems like automatic speech recognition (ASR) or handwriting recognition should be consistent, i.e. should follow the same, global objective, which ideally would be word (or character) error rate on training and unseen test data, respectively. This motivates integrated system architectures that are trainable end-to-end. Current state-of-the-art large vocabulary continuous speech recognition systems are based on the hybrid HMM approach [2], e.g. as recently shown on the Switchboard task [22, 19]. In the standard HMM approach an explicit decomposition into separately trained acoustic (class-conditional) and language (class prior) models is performed. Acoustic models by derivation are generative models, but corresponding discriminative (sequence) training criteria are available, which do allow for end-to-end acoustic model training [20, 5, 17]. The hybrid HMM and the tandem [9] approach incorporating discriminative artificial neural networks into the originally generative acoustic models both are not fully consistent. The hybrid HMM approach utilizes approximate label priors, and both assume conditional independence for the input observations, although the currently best performing long-short-term memory (LSTM) [10] based models do consider full-utterance acoustic context. Bayes decision rule requires a posterior distribution over label sequences. Hence, a directly discriminative sequence-to-sequence approach seems more promising, which necessitates alternative solutions to the alignment problem, as proposed in connectionist temporal classification (CTC) [7], or in attention-based sequence-to-sequence approaches [3, 1, 14, 4]. The end-to-end properties of CTC are limited by a conditional independence assumption on label level, whereas attention is known to underperform for longer sequences [3].

In this work, we propose a fully stochastic, HMM-like decomposition of a label sequence posterior distribution that enables end-to-end training and does not rely on restrictive independence assumptions w.r.t. input observation or label sequence level. This approach somehow inverts the alignment problem

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

between HMM state labels and time frames/observations. In contrast to the attention approach, the resulting alignment problem remains part of the overall search problem, i.e. the local best recognized label and its corresponding alignment only is fixed when the search is finished, or when the local search space subsequently is reduced to a single best context. Due to its discriminative nature, the approach can consistently operate on the complete (utterance-level) acoustic context, thus accommodates e.g. bi-directional recurrent neural network without further assumptions.

We derive the inverted HMM, discuss potential model assumptions, and provide a first proof of concept on a small handwriting recognition task, which is based on our ASR software.

2 Alignment in Sequence Classification

The standard Bayes decision rule requires maximization of the class posterior given an input observation. In sequence classification, classes are discrete label sequences (e.g. words, characters, phonemes, generalized triphone states, etc.). Statistical sequence classification based on Bayes decision rule requires a model for the label sequence posterior probability distribution (or an equivalent discriminant function derived from it). The label sequence posterior distribution especially needs to cover the alignment problem between observation (sub-)sequences and labels.

In automatic speech recognition (ASR) and related areas it has been common to assume decomposition of the sequence posterior into a label sequence prior, the so-called language model, and a generative class-conditional distribution, the so-called acoustic model. A hidden Markov model (HMM) then is used to model the generative sequence class-conditional distribution. It introduces a sequence of stochastic hidden variables, the HMM states, to model label sequences, i.e. provide a way to align observation sequences, thus enabling speaking rate normalization. The HMM derivation especially includes a conditional independence assumption between the observations over time, which is known not to be fulfilled [16]. The currently most common method to introduce neural networks to acoustic modeling is the so-called hybrid HMM approach [2], where emission probabilities directly are modeled by label posteriors. Hybrid HMMs work well with large temporal input context up to whole utterances, which is inconsistent with the HMM assumptions, though. In addition, usually approximative label priors (e.g. [15]) are used in the hybrid approach. However, w.r.t. recognition performance, the hybrid HMM approach still marks the current state-of-the-art in large-vocabulary continuous speech recognition, as e.g. shown recently on the Switchboard task [22, 19].

Connectionist temporal classification (CTC) [7] attempts to directly derive a discriminative model for the label sequence posterior, thus avoiding the conditional independence assumption of tandem and hybrid HMM. The alignment problem here is handled by introducing an additional blank symbol which may be inserted more or less arbitrarily between actual labels on a per-frame basis without changing the label sequence represented. In contrast to the hybrid HMM approach, no state transition probability distribution and no state prior are required for CTC. Note that in the CTC derivation a conditional independence of the labels is assumed. Depending on the label definition (e.g. syllables), this might be viable for acoustic modeling, but the label context dependency then needs to be considered by combination with a separate language model.

Hybrid-HMM and CTC approaches both operate on a frame-by-frame level, as the standard HMM-GMM approach: each observation frame is aligned to an HMM state (or CTC) label. In contrast to this, sequence-to-sequence approaches [21, 6] directly operate on label positions. Attention [3, 1, 14, 4], then allows to cover multiple frames per label directly. On top of an observation sequence encoder, attention assigns a local weighted average of its output as input to a label decoder. However, attention weights for a label position do not depend on the label to be classified in the same label position. Therefore, decisions in subsequent label positions cannot alter the attention for that label position. For labels, whose support in terms of the number of input frames varies strongly, it might be unsuitable to assume such locally uniform support (after encoding). Even though encoders based on recurrent neural networks in principle might be able to mitigate this to some extent, corresponding errors in the implicit alignment resulting from the attention process can be expected to propagate for longer sequences [3, 11]. As a consequence, e.g. in [11] a block-wise operation is suggested.

3 Inverted HMM

In attention-based sequence-to-sequence (seq2seq) models, the (implicit) alignment is not a fully stochastic process. The computation of the attention weights can be seen as a kind of partial decision on the local average observation (after encoding) to be considered for label classification in a given label context. Here, we try to generalize the HMM concept to both drop the conditional independence assumption and keep the alignment as part of the corresponding stochastic model. As for CTC and

seq2seq approaches, we start from the posterior of a label sequence α_1^S of length S, given an input observation sequence x_1^T of length T. Somewhat similar to attention, and in contrast to the standard HMM, we assume an *inverted* alignment of labels to observations (or encoder output), cf. Fig. 1, explicitly using latent variables t_1^S representing label position in time (e.g. label end/start time). In the following, a derivation with some exemplary model assumptions is presented[1]:

$$p(\alpha_1^S|x_1^T) = \sum_{t_1^S} p(\alpha_1^S, t_1^S|x_1^T) = \sum_{t_1^S} \prod_{s=1}^{S} p(\alpha_s, t_s|\alpha_1^{s-1}, t_1^{s-1}, x_1^T) \qquad \text{(general approach)}$$

(exemplary model assumptions)

$$= \sum_{t_1^S} \prod_{s=1}^{S} p(\alpha_s|\alpha_1^{s-1}, t_{s-1}^s, x_1^T) \cdot p(t_s|\alpha_1^{s-1}, t_{s-1}) \qquad \text{(limit alignment context)} \quad (1)$$

$$= \sum_{t_1^S} \prod_{s=1}^{S} p\big(\alpha_s|\alpha_1^{s-1}, a(y_{t_{s-1}+1}^{t_s}(x_1^T)\big) \cdot p(t_s|\alpha_1^{s-1}, t_{s-1}) \quad \text{(parametric attention)} \quad (2)$$

$$= \sum_{t_1^S} \prod_{s=1}^{S} p(\alpha_s|\alpha_{s-m+1}^{s-1}, t_{s-1}^s, x_1^T) \cdot p(t_s|t_{s-1}) \quad \text{(label-level Markov assumption)} \quad (3)$$

$$\approx \max_{t_1^S} \prod_{s=1}^{S} p(\alpha_s|t_s, x_1^T) \cdot p(t_s - t_{s-1}) \quad \text{(restrictive model assumptions used here)} \quad (4)$$

The last line shows the basic model chosen here for a proof of concept, with fairly restrictive model assumptions (i.e. conditional label independence, similar to CTC; stationary length model) and maximum approximation. However, less restrictive model assumptions are conceivable. E.g. a recurrent neural network (RNN) based decoder would require full label context dependency as in Eq. (1). Also, an attention operation a on top of a (potentially downsampled) encoder output sequence $y_1^{T'}$ parametrized by start/end times (t_{s-1}, t_s) is conceivable, as suggested in Eq. (2). In this case, attention might be realized by a unidirectional RNN operating on encoder output between t_{s-1} and t_s, which we investigate in further work. For a decoder component, also Markov assumptions on label level are possible as suggested in Eq. (3), which would enable dynamic programming search. The decoder can be implemented either label-synchronous, or time-synchronous, where the latter implies keeping predecessor hypotheses for more than the direct predecessor time frame. Provided the maximum step size $t_s - t_{s-1}$ is limited, the corresponding time complexity would still be linear in the length of the input T.

In addition to the label posterior, a length model is part of the inverted HMM approach, i.e. a distribution over the next time t_s a label α_s in position s would be aligned to. For the proof of concept provided here, we reduce the dependency of this model strongly to the direct predecessor label position in time, t_{s-1}, and assume stationarity. The inverted HMM does not generate any hypotheses for intermediate frames, where CTC hypothesizes the blank label, and where the hybrid approach coninuously hypothesizes the same target label (divided by the label prior). The idea of dropping the blank label was proposed earlier in [18, Sec. 2.5]), though therein even full-word labels still were modeled using CTC. Whereas state transitions in the hybrid HMM are restricted to loop, forward, and skip transitions, for the inverted HMM long skips over a number of frames are considered, cf. Fig. 1. The length model in the inverted HMM is therefore defined by the number of states per label and the maximum number of timesteps which may be skipped $t_s - t_{s-1}$, which must be set large enough to cover the speaking rate variation of the input signal in order to get reasonable performance.

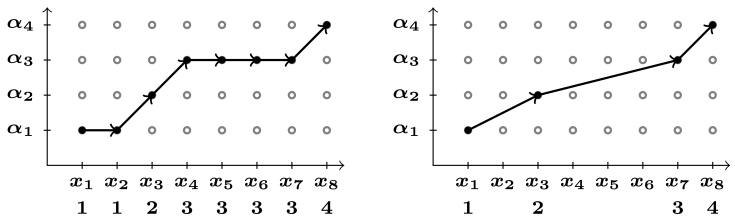

Figure 1: Alignment paths and target labels (below) for standard (left) and inverted HMM (right).

[1]Sequence notation used here: z_a^b denotes a sequence $z_a, z_{a+1}, \ldots, z_b$.

4 Experimental Results

We investigate the proposed inverted HMM from Eq. (4) on the isolated word (character sequence) recognition task of the RIMES database [8]. The RIMES task comprises unconstrained French handwriting and was used for the ICDAR 2011 handwriting recognition competition. The dataset contains 51,738 training word images and 7,464 testing word images. The recognition vocabulary contains 81 characters. A moment-based feature extraction as described in [13] was used to generate the observation sequence.

During training of the models we take 10% of the training data for cross-validation w.r.t. cross-entropy and frame error. We further partitioned all input sequences into chunks of size 100 and processed up to 25 chunks in a minibatch. Optimization was done using the Adam [12] optimizer with a learning rate of 10^{-3}. The general setup of the LSTM networks used for the inverted HMM from Eq. (4), hybrid HMM, CTC, and seq2seq encoder was composed of three bidirectional layers with 500 units in each direction. The softmax layer contained one unit for each of the 81 characters. In the attention-based seq2seq system we then initialized another unidirectional LSTM decoder with the final state of the last hidden layer and computed context vectors in each iteration as in [4].

Even though the inverted HMM allows for end-to-end training and more elaborate modeling, the focus here is a first proof of concept to show that the inverted HMM already works well even with restricted model assumptions and an initialization via existing Viterbi alignments from previously trained standard GMM-HMM models. Actual end-to-end training and more powerful model assumptions for this novel approach are under way and will be investigated in further work.

Recognition is performed using either the standard HMM (0,1,2) topology or an inverted alignment topology with a maximum time step (skip) of 7 per state and an overall 3 states per label. Furthermore, we used the initial models to create a new standard HMM and inverted HMM alignment of the training data, respectively. Any skipped time frame/observation in the inverted HMM alignment is labeled with a special skip-symbol which tells the network to avoid computing an error signal in the output layer. We repeated the realignment process three times to obtain the final results shown in Table 1. For further comparison we added the results for a system trained with CTC and an attention-based seq2seq system. Note that, in contrast to CTC, here only Viterbi training was used for the inverted HMM, yet.

The hybrid and inverted alignment systems were trained in an expectation-maximization fashion by reiterating training of the model and recomputation of the (standard/inverted) Viterbi alignment. However, the entire pipeline also is trainable from scratch, although this was not yet analyzed, here. During decoding, a unigram language model was used that was estimated on the training transcriptions of the RIMES corpus. In this proof-of-concept, we computed the score for all words in the known vocabulary and selected the one with the best score. The initial alignment was obtained by training an HMM with Gaussian mixture models as emission density distributions.

The results show that inverted alignments are suitable for recognition and that competitive results to traditional systems can be obtained. Similar to CTC, the recurrent structure allows the network to localize the relevant information about the current label at one single observation and a precise position is not even required. Overall the results are comparable to the results of the competing teams in the ICDAR 2011 competition, where a more elaborate multi-dimensional LSTM model trained with CTC reached the first place with a WER of 5.13% and an HMM system with 12.53% WER was ranked second.

The inverted alignment model is a promising candidate for end-to-end training and provides a statistical alternative to attention-based systems. It might help bridge the gap between traditional hybrid recognition systems and sequence-to-sequence learning. Our future work will concentrate on actual end-to-end training of this novel approach, as well as an investigation into more elaborate model assumption including stochastic attention on top of an encoder, as well as decoders including label context dependency.

Table 1: Comparison of the proposed inverted HMM alignment based recognition with a hybrid HMM, CTC, and sequence-to-sequence attention on the RIMES isolated word (character sequence) handwriting recognition dataset. Word (WER) and character error rates (CER) are provided.

Systems	Alignment	WER [%]	CER [%]
Hybrid	Viterbi	7.1	3.0
Inverted		7.5	2.9
seq2seq	Attention	7.7	4.1
CTC	Forward / Backward	6.7	2.8

References

[1] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio: "End-to-End Attention-Based Large Vocabulary Speech Recognition". In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4945–4949, Shanghai, China, Mar. 2016.

[2] H. A. Bourlard and N. Morgan: *Connectionist Speech Recognition: a Hybrid Approach*. Kluwer Academic Publishers, Norwell, MA, 1993.

[3] W. Chan, N. Jaitly, Q. V. Le, and O. Vinyals: "Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition". In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4960–4964, Shanghai, China, Mar. 2016.

[4] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio: "End-to-End Continuous Speech Recognition using Attention-Based Recurrent NN: First Results". *CoRR*, abs/1412.1602, Dec. 2014.

[5] G. Gosztolya, T. Grósz, and L. Tóth: "GMM-Free Flat Start Sequence-Discriminative DNN Training". In *Interspeech*, pages 3409–3413, San Francisco, CA, Sept. 2016.

[6] A. Graves: "Generating Sequences with Recurrent Neural Networks". *CoRR*, abs/1308.0850, Aug. 2013.

[7] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber: "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In *Intern. Conf. on Machine Learning (ICML)*, pages 369–376, New York, NY, June 2006.

[8] E. Grosicki and H. El-Abed: "ICDAR 2011 - French Handwriting Recognition Competition". In *Intern. Conf. on Document Analysis and Recognition (ICDAR)*, pages 1459–1463, Sept. 2011.

[9] H. Hermansky, D. P. Ellis, and S. Sharma: "Tandem Connectionist Feature Extraction for Conventional HMM Systems". In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, pages 1635–1638, Istanbul, Turkey, June 2000.

[10] S. Hochreiter and J. Schmidhuber: "Long Short-Term Memory". *Neural computation,* Vol. 9, No. 8, pages 1735–1780, 1997.

[11] N. Jaitly, Q. V. Le, O. Vinyals, I. Sutskever, and S. Bengio: "A Neural Transducer". *CoRR*, abs/1511.04868, Nov. 2015.

[12] D. P. Kingma and J. Ba: "Adam: A Method for Stochastic Optimization". *CoRR*, abs/1412.6980, Dec. 2014.

[13] M. Kozielski, P. Doetsch, and H. Ney: "Improvements in RWTH's System for Off-Line Handwriting Recognition". In *Intern. Conf. on Document Analysis and Recognition (ICDAR)*, pages 935–939, Washington, DC, Aug. 2013.

[14] M. Luong, H. Pham, and C. D. Manning: "Effective Approaches to Attention-Based Neural Machine Translation". *CoRR*, abs/1508.04025, Aug. 2015.

[15] V. Manohar, D. Povey, and S. Khudanpur: "Semi-Supervised Maximum Mutual Information Training of Deep Neural Network Acoustic Models". In *Interspeech*, pages 2630–2634, Dresden, Germany, Sept. 2015.

[16] S. H. K. Parthasarathi, S.-Y. Chang, J. Cohen, N. Morgan, and S. Wegmann: "The Blame Game in Meeting Room ASR: An Analysis of Feature Versus Model Errors in Noisy and Mismatched Conditions". In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 6758–6762, Vancouver, Canada, May 2013.

[17] D. Povey, V. Peddinti, D. Galvez, P. Ghahrmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur: "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI". In *Interspeech*, pages 2751–2755, San Francisco, CA, Sept. 2016.

[18] H. Sak, A. Senior, K. Rao, and F. Beaufays: "Fast and Accurate Recurrent Neural Network Acoustic Models for Speech Recognition". pages 1468–1472, Dreseden, Germany, Sept. 2015.

[19] G. Saon, T. Sercu, S. Rennie, and H.-K. J. Kuo: "The IBM 2016 English Conversational Telephone Speech Recognition System". In *Interspeech*, pages 7–11, San Francisco, CA, Sept. 2016.

[20] A. W. Senior, G. Heigold, M. Bacchiani, and H. Liao: "GMM-free DNN Acoustic Model Training". In *IEEE Intern. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 5602–5606, Florence, Italy, May 2014.

[21] I. Sutskever, O. Vinyals, and Q. V. Le: "Sequence to Sequence Learning with Neural Networks". *CoRR*, abs/1409.3215, Sept. 2014.

[22] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig: "Achieving Human Parity in Conversational Speech Recognition". *ArXiv e-prints*, abs/1610.05256, Oct. 2016.