CAT: A CTC-CRF based ASR Toolkit Bridging the Hybrid and the End-to-end Approaches towards Data Efficiency and Low Latency

Keyu An, Hongyu Xiang, Zhijian Ou†

Speech Processing and Machine Intelligence (SPMI) Lab, Tsinghua University, China
aky19@mails.tsinghua.edu.cn, xiangyhl6@mails.tsinghua.edu.cn, ozj@tsinghua.edu.cn

Abstract

In this paper, we present a new open source toolkit for speech recognition, named CAT (CTC-CRF based ASR Toolkit). CAT inherits the data-efficiency of the hybrid approach and the simplicity of the E2E approach, providing a full-fledged implementation of CTC-CRFs and complete training and testing scripts for a number of English and Chinese benchmarks. Experiments show CAT obtains state-of-the-art results, which are comparable to the fine-tuned hybrid models in Kaldi but with a much simpler training pipeline. Compared to existing non-modularized E2E models, CAT performs better on limited-scale datasets, demonstrating its data efficiency. Furthermore, we propose a new method called contextualized soft forgetting, which enables CAT to do streaming ASR without accuracy degradation. We hope CAT, especially the CTC-CRF based framework and software, will be of broad interest to the community, and can be further explored and improved.

Index Terms: speech recognition, CRF, CTC, end-to-end, data-efficiency

1. Introduction

Deep neural networks (DNNs) of various architectures have become dominantly used in automatic speech recognition (ASR), which roughly can be classified into two approaches - the DNN-HMM hybrid and the end-to-end (E2E) approaches. Initially, the DNN-HMM hybrid approach was adopted [1], which is featured by using the frame-level loss (cross-entropy) to train the DNN to estimate the posterior probabilities of HMM states. A GMM-HMM training is firstly needed to obtain frame-level alignments and then the DNN-HMM is trained. The hybrid approach usually consists of an DNN-HMM based acoustic model (AM), a state-tying decision tree for context-dependent phone modeling, a pronunciation lexicon and a language model (LM), which can be compactly combined into a weighted finite-state transducer (WFST) [2] for efficient decoding.

Recently, the E2E approach has emerged [3, 4, 5, 6], which is characterized by eliminating the construction of GMM-HMMs and phonetic decision-trees, training the DNN from scratch (in single-stage) and, even ambitious, removing the need for a pronunciation lexicon and training the acoustic and language models jointly rather than separately. The key to achieve this is to define a differentiable sequence-level loss of mapping the acoustic sequence to the label sequence. Three widely-used E2E losses are based on Connectionist Temporal Classification (CTC) [3], RNN-transducer (RNN-T) [5], and attention based encoder-decoder [6] respectively.

When comparing the hybrid and E2E approaches (modularity versus a single neural network, separate optimization versus joint optimization), it is worthwhile to note the pros and cons of each approach. The E2E approach aims to subsume the acoustic, pronunciation, and language models into a single neural network and perform joint optimization. This appealing feature comes at a cost, i.e. the E2E ASR systems are data hungry, which require above thousands of hours of labeled speech to be competitive with the hybrid systems [7, 8, 9]. In contrast, the modularity of the hybrid approach permits training the AM and LM independently and on different data sets. A decent acoustic model can be trained with around 100 hours of labeled speech whereas the LM can be trained on text-only data, which is available in vast amounts for many languages. In this sense, modularity promotes data efficiency. Due to the lack of modularity, it is difficult for an E2E model to exploit the text-only data, though there are recent efforts to alleviate this drawback [10, 11]. In this paper, we are interested in bridging the hybrid and the E2E approaches, trying to inherit the data-efficiency of the hybrid approach and the simplicity of the E2E approach. A second motivation for such bridging is that low latency ASR has been addressed relatively easier and better in the hybrid approach than in the E2E approach, as will be discussed later in Section 2.

Specifically, we base on the recently developed CTC-CRF approach[12]. Basically, CTC-CRF is a CRF (Conditional Random Field) with CTC topology, which eliminates the conditional independence assumption in CTC and performs significantly better than CTC. It has been shown [12] that CTC-CRF has achieved state-of-the-art benchmarking performance with training data ranging from ~100 to ~1000 hours, while being end-to-end with a simplified pipeline (eliminating GMM-HMMs and phonetic decision-trees, training DNN-based AM in single-stage) and being data-efficient in the sense that cheaply available LMs can be leveraged effectively with or without a pronunciation lexicon.

In this paper we present CAT (CTC-CRF based ASR Toolkit) towards data-efficient and low-latency E2E ASR, which trains CTC-CRF based AMs in single-stage and uses separate LMs, with or without a pronunciation lexicon. On top of the previous work [12], the new contributions of this work are as follows.

1. CAT releases an full-fledged implementation of CTC-CRFs. A non-trivial issue in training CTC-CRFs is that the gradient is the difference between empirical expectation and model expectation. CAT contains efficient implementations of the forward-backward algorithm for calculating these expectations using CUDA C/C++ interface. CAT adopts PyTorch [13] to build DNNs and do automatic gradient computation, and so inherits the power of PyTorch in handling DNNs. In CAT, we can readily use the PyTorch DistributedDataParallel module to support training over multi-node and multi-GPU hardware.

2. We add the support of streaming ASR in the toolkit. To this end, we propose a new method called contextualized soft forgetting (CSF), which combines soft forgetting [14] and...
context-sensitive-chunk [15] in bidirectional LSTM (BLSTM). Extensive experiments show that: (a) CTC-CRF with soft for-
getting improves over CTC with soft forgetting significantly and
consistently; (b) With contextualized soft forgetting, the chunk
BLSTM based CTC-CRF with a latency of 300ms outperforms
the whole-utterance BLSTM based CTC-CRF.

3. CAT provides reproducible, complete training and
testing scripts for a number of English and Chinese bench-
marks, including but not limited to WSJ, Switchboard, Fisher-
Switchboard, and AISHELL datasets which are presented in
this paper. CAT achieves state-of-the-art ASR performance
on these datasets, which are comparable to the LF-MMI [18]
results in Kaldi (one of the strongest, fine-tuned hybrid ASR
toolkit) but with a much simpler training pipeline. Remark-
ably, compared to existing non-modularized E2E models, CAT
performs better on limited-scale datasets (with ~100 to ~2000
hours of training data), demonstrating its data efficiency.

2. Related Work

ASR toolkits. Roughly speaking, there are two approaches
to using DNNs in ASR - the DNN-HMM hybrid and the E2E
approaches. So does the classification of existing ASR tool-
kits. For the hybrid approach, Kaldi [19] may be the most
widely-used hybrid DNN-HMM based ASR toolkit. In Kaldi,
lattice-free maximum-mutual-information (LF-MMI) training
needs a multi-stage pipeline consisting of GMM-HMM training
and phonetic decision tree construction. There have emerged
even E2E ASR toolkits (e.g. ESPnet [20]/ESPRESSO [21],
Wav2letter++ [22], and Lingvo [23]), mostly focusing on using
attention-based encoder-decoder or hybrid CTC/attention.

EESEN [4] and E2E-LF-MMI [24, 25] seem to bridge the
hybrid and the E2E approaches, by using the sequence-level
loss (CTC and LF-MMI respectively) to train single-stage AMs
and employing WFST based decoding. EESEN is based on
CTC, which, different from CTC-CRF, is limited by its condi-
tional independence assumption and weak performance. E2E-
LF-MMI [24, 25] was developed with two versions of using
mono-phones or bi-phones, and bi-phone E2E-LF-MMI obtains
comparable results to hybrid LF-MMI. It is shown in our ex-
periments that mono-phone CTC-CRF performs comparably to
bi-phone E2E-LF-MMI but with a simpler pipeline. Bi-phone
CTC-CRFs is found to slightly improve over mono-phone CTC-
CRFs but will complicate the training pipeline. The differences
between E2E-LF-MMI and CTC-CRF are detailed in [12].

Low latency ASR. An important feature for a practical
ASR toolkit is its ability to do streaming ASR with low la-
tency. In the hybrid approach, chunk-based schemes have been
investigated in BLSTM [15, 26]. Time-delay neural networks
(TDNNs) with interleaving LSTM layers (TDNN-
LSTM) [27] has been developed in Kaldi to successfully limit the
latency while keeping the recognition accuracy. In con-
trast, it is challenging and more complicated for attention-based
encoder-decoders to do streaming ASR, which recently has re-
ceived increasing studies, such as monotonic chunkwise atten-
tion (MoChA) [28], triggered attention [29], or using limited fu-
ture context in the encoder [17]. RNN-T has some advantage for
streaming ASR but is data hungry, requiring large-scale training
data to work. The RNN-T result over the Fisher-Switchboard
data (2300 hours) [30] is worse than CAT, as shown in Table 4.

CAT, which, different from CTC-CRF, is limited by its condi-
tional independence assumption and weak performance. CTC-CRF
performs better on limited-scale datasets (with ~100 to ~2000
hours of training data), demonstrating its data efficiency.

3. CTC-CRF based ASR

CAT consists of separable AM and LM, which meets our ra-
nionale to be data efficient by keeping necessary modularity. In
the following we mainly describe our CTC-CRF based AMs.
CAT uses SRILM for LM training, and some code from Kaldi
and EESEN for data preparation, decoding graph compiling and
WFST based decoding. More details can be found in the toolkit.

Consider discriminative training of DNN-based AMs in
single-stage based on the loss defined by conditional maximum
likelihood [12]:

\[ \mathcal{L}(\theta) = -\log p_\theta(l|x) \]  (1)

where \( x \equiv x_1, \ldots, x_T \) is the speech feature sequence and
\( l \equiv l_1, \ldots, l_L \) is the label (phone, character, word-piece and etc)
sequence, and \( \theta \) is the model parameter. Note that \( x \) and \( l \) are in
different lengths and usually not aligned. To handle this, a hidden
state sequence \( \pi_1, \ldots, \pi_T \) is introduced; state topology refers to
the state transition structure in \( \pi \), which basically de-

\[ p_\theta(l|x) = \sum_{\pi \in S^T} p_\theta(l|\pi, x) \]  (2)

And the posteriori of \( \pi \) is further defined by a CRF:

\[ p_\theta(\pi|l, x) = \frac{\exp(\phi_\theta(\pi, x))}{\sum_{\pi'} \exp(\phi_\theta(\pi', x))} \]  (3)

Here \( \phi_\theta(\pi, x) \) denotes the potential function of the CRF, def-
de as:

\[ \phi_\theta(\pi, x) = \log p(\pi) + \sum_{t=1}^{T} \log p_\theta(\pi_t|x) \]

where \( l = B(\pi) \). \( \sum_{t=1}^{T} \log p_\theta(\pi_t|x) \) defines the node poten-
tial, calculated from the bottom DNN. \( \log p(\pi) \) defines the edge
potential, realized by an n-gram LM of labels and, for reasons to
be clear in the following, referred to as the denominator n-gram
LM. Remarkably, regular CTC suffers from the conditional in-
dependence between the states in \( \pi \). In contrast, by incorporat-
ing \( \log p(\pi) \) into the potential function in CTC-CRF, this draw-
back is naturally avoided. Combining Eq. (1)-(3) yields the
sequence-level loss used in CTC-CRF:

\[ \mathcal{L}(\theta) = -\log \frac{\sum_{\pi \in S^T} \exp(\phi_\theta(\pi, x))}{\sum_{\pi'} \exp(\phi_\theta(\pi', x))} \]  (4)

The gradient of the above loss involves two gradients calcu-
lated from the numerator and denominator respectively, which
essentially correspond to the two terms of empirical expecta-
tion and model expectation as commonly found in estimating
CRFs. Similarly to LF-MMI, both terms can be obtained via
the forward-backward algorithm. Specifically, the denominator
calculation involves running the forward-backward algorithm over the denominator WFST $T_{\text{den}}$. $T_{\text{den}}$ is a composition of the CTC topology WFST and the WFST representation of the n-gram LM of labels, which is called the denominator n-gram LM, to be differentiated from the word-level LM in decoding.

4. Contextualized Soft Forgetting towards Streaming ASR

To enable streaming ASR in CAT, we draw inspirations from soft forgetting [14] and context-sensitive-chunk [15] in using BLSTM. With the hypothesis that whole-utterance unrolling of the BLSTM leads to overfitting, soft forgetting, which is developed for CTC-based ASR, consists of three elements. First, the BLSTM network is unrolled over non-overlapping chunks. The hidden and cell states are hence forgotten at chunk boundaries in training. Second, the chunk duration is perturbed across training minibatches, which is called chunk size jitter. Third, the CTC loss is added with a twin regularization term, which is the mean-squared error between the hidden states of a pre-trained fixed whole-utterance BLSTM and the chunk-based BLSTM being currently trained. Since twin regularization promotes some remembering across chunks, this method is called soft forgetting. In streaming recognition, the hidden and cell states of the forward LSTM are copied over from one chunk to the next, and the backward LSTM hidden and cell states are reset to zero.

The idea of context-sensitive-chunk (CSC) is proposed in the BLSTM-HMM hybrid system to reduce the latency from a whole utterance to a chunk. In CSC, a chunk is appended with a fixed number of left and right frames as left and right contexts.

In CAT, we propose to apply soft forgetting to context-sensitive-chunks, which is called contextualized soft forgetting (CSF) as illustrated in Figure 1. First, we split an utterance into non-overlapping chunks. For each chunk, a fixed number of frames to the left and right of the chunk are appended as contextual frames except for the first and last chunk, where we use zeros as the left and right contexts respectively. Thus we form context-sensitive-chunks and run BLSTM over each CSC. The hidden and cell states of the forward and backward LSTM networks are reset to zeros at the left and right boundaries of each CSC in both training and inference. When calculating the sequence-level loss in CTC-CRF, we splice the neural network output from chunks into a sequence again, but excluding the network outputs from contextual frames. A pre-trained fixed whole-utterance BLSTM is used to regularize the hidden states of the CSC-based BLSTM, and the overall training loss is the sum of the CTC-CRF loss and the twin regularization loss with a scaling factor $\lambda$. Note that once the CSC-based BLSTM is trained, we can discard the whole-utterance BLSTM and perform inference over testing utterances without it.

5. Experiment Settings

The experiment consists of two parts. In the first part, we introduce the results on several representative benchmarks, including WSJ (80-h), AISHELL (170-h Chinese), Switchboard (260-h) and Fisher-Switchboard (2300-h) (the numbers in the parentheses are the size of training data in hours). The performances over these limited-scale datasets reveal the data efficiency of different ASR models. The second part presents the results for streaming ASR by the proposed contextualized soft forgetting with ablation study.

It should be noted that the results shown in this paper should not be compared with results obtained with heavy data augmentation (e.g. specAugment [31]), much larger DNNs, and model combination. When compared to results reported from other papers, unless otherwise stated, we cite those results under comparable conditions to the best of our knowledge.

5.1. Setup for benchmarking experiment

We compare CAT with state-of-the-art ASR systems on several benchmarks, as stated above. We apply speed perturbation for 3-fold training data augmentation, except on Fisher-Switchboard. Unless otherwise stated, 40 dimension filter bank with delta and delta-delta features are extracted. The features are normalized via mean subtraction and variance normalization per utterance, and sampled by a factor of 3.

The AM network, different from [12], is two blocks of VGG layers followed by a 6-layer BLSTM similar to [32]. We apply 1D max-pooling to the feature maps produced by VGG blocks on the frequency dimension only, since the input features have been sampled in time-domain and we find that max-pooling along the time dimension will deteriorate the performance. The first VGG block has 3 input channels corresponding to spectral features, delta, and delta-delta features. The BLSTM has 320 hidden units per direction for WSJ and AISHELL, and 512 for Switchboard and Fisher-Switchboard. The total number of parameters is 16M and 37M respectively, much smaller than most E2E models. Denominator 4-gram LMs are used throughout the experiments. In training, a dropout [33] probability of 50% is applied to the LSTM to prevent overfitting. Following [12], a CTC loss with a weight $\alpha$ is combined with the CRF loss to help convergence. We set $\alpha = 0.01$ by default and find in practice that the smaller $\alpha$ is, the better the final result will be.

5.2. Setup for streaming ASR experiment

To evaluate the effectiveness of contextualized soft forgetting, we first implement soft forgetting with the CTC-CRF loss. For a fair comparison, we adopt the same neural network architecture as in [14], which is a 6-layer BLSTM with 512 hidden units per direction. 40 dimension MFCC with delta and delta-delta are extracted, and the chunk size is set to 40. The whole-utterance BLSTM pre-trained on 260hr Switchboard obtain 14.3% WER on eval2000. For twin regularization, the scaling factor $\lambda$ is set to 0.005. In contextualized soft forgetting, the chunk size is also 40, with 10 left and 10 right frames appended.
6. Experimental results

6.1. Results for benchmarking experiment

The WER results on WSJ are shown in Table 1, including two test sets - dev93 and eval92. It can be seen that CTC-CRF, hybrid LF-MM and E2E-LF-MM, which all keep modularity, perform comparable, and much better than other E2E models.

The CER (Character Error rate) results on AISHELL are shown in Table 2. It can be seen that CTC-CRF obtains state-of-the-art performance on AISHELL dataset - the CER is much better than other E2E models and the hybrid LF-MM in Kaldi.

The WER results on Switchboard are shown in Table 3. The Eval2000 test set consists of two subsets - Switchboard (SW) and Callhome (CH). It can be seen that compared to bi-phone hybrid LF-MM and E2E-LF-MM, mono-phone CTC-CRF performs comparably but with a simpler pipeline. Remarkably, mono-phone CTC-CRF performs significantly better than other E2E models.

The WER results on Fisher-Switchboard are shown in Table 4. The performance of CTC-CRF, with no data augmentation, is on par with state-of-the-art hybrid and E2E models. Summing up the above results, we can see that on the limited-scale datasets (such as 80-h, 170-h, 260-h and 2300-h training data), the modularity of CTC-CRF clearly promotes data efficiency and achieve better results than other data hungry E2E models.

6.2. Results for streaming ASR experiment

First, we introduce different elements of soft forgetting [14] in steps to show their impact on WERs and also compare CTC and CTC-CRF. For this purpose, we follow [14] to report the non-streaming recognition results, as shown in Table 5. We start from training the basic chunk-based BLSTM networks with a fixed chunk size. It can be seen that CTC-CRF improves over CTC significantly under all experiment settings.

Then we examine the streaming recognition. It can be seen from Table 6 that CTC-CRFs trained with CSF improve significantly over CTC-CRFs with SF, and obtain comparable result with state-of-the-art TDNN-LSTM based hybrid model [27]. Remarkably, the CSF based streaming CTC-CRF (14.1%) even outperforms the whole-utterance CTC-CRF (14.3%), presumably because CSF alleviates overfitting in addition to realizing

 streaming ASR. This is in contrast to streaming ASR results by other E2E models, where streaming E2E models can hardly outperform their whole-utterance models [16, 17, 41].

7. Conclusion

This paper introduces an open source ASR toolkit - CAT, with the main features of data efficiency, simple pipeline, streaming ASR and superior results. We propose a new method called contextualized soft forgetting, which enables CAT to do streaming ASR without accuracy degradation. We hope CAT, especially the CTC-CRF based framework and software, will be of broad interest to the community, and can be further explored and improved, e.g., exploring different DNN architectures, different topologies of CRFs, and the application in more ASR tasks.
8. References


