

Layer Flexible Adaptive Computation Time

1st Lida Zhang
dept. Computer Science & Engineering
Texas A&M University
College Station, USA
lidazhang@tamu.edu

2nd Abdolghani Ebrahimi
dept. Industrial Engineering
Northwestern University
Evanston, USA
ghani@u.northwestern.edu

3rd Diego Klabjan
dept. Industrial Engineering
Northwestern University
Evanston, USA
d-klabjan@northwestern.edu

Abstract—Deep recurrent neural networks perform well on sequence data and are the model of choice. However, it is a daunting task to decide the structure of the networks, i.e. the number of layers, especially considering different computational needs of a sequence. We propose a layer flexible recurrent neural network with adaptive computation time, and expand it to a sequence to sequence model. Different from the adaptive computation time model, our model has a dynamic number of transmission states which vary by step and sequence. We evaluate the model on a financial data set and Wikipedia language modeling. Experimental results show the performance improvement of 7% to 12% and indicate the model’s ability to dynamically change the number of layers along with the computational steps.

I. INTRODUCTION

Recurrent neural networks (RNN) are widely used in supervised machine learning tasks for their superior performance in sequence data, such as machine translation [1], [2], speech recognition [3], [4], image description generation [5], [6], and music generation [7]. The design of the underlying network is always a daunting task requiring substantial computational resources and experimentation. Many recent breakthroughs hinge on multilayer neural networks’ ability to increase model accuracy, [8]–[10], leading to the important decision in RNNs of the number of computational steps. First, the right choice requires running several very expensive training processes to try many different computational steps. Even if a reinforcement learning algorithm is used to determine a good computational steps, [11], [12], it still requires a substantial training effort. The second issue with the fixed structure RNNs is the fact that the same computational steps is applied to each input in a sequence. It is conceivable that some inputs are harder to classify than others and thus such harder inputs should employ more computational steps. A similar argument holds for steps, e.g., certain steps in a sample can bear less predictive power and thus should use fewer computational steps in order to decrease the computational burden. The goal of our work is to introduce a network that automatically determines the computational steps - and together with this the number of hidden vectors to use - in training and inference which is dynamic with respect to samples and step number.

To resolve the inherent problems of fixed structure neural networks, [13] addresses this by providing an Adaptive Computation Time (ACT) model for RNN. In Graves’ model, a sigmoidal halting unit is utilized to calculate a halting

probability for each intermediate round within a step, and a computation stops when the accumulated halting probability reaches or exceeds a threshold. ACT can utilize multiple computation rounds within each individual step and it can dynamically adapt to different samples and steps. The model is appealing due to its modeling flexibility and its advantage in increasing model accuracy [14]. With the ACT mechanism, when a step of computation is halted, all intermediate states and outputs are used to calculate one mean-field state and output. The mean-field state and output have drawbacks. The outputs of the deepest computational step are the most informative, and should be the final outputs. The output from early computational steps may cause errors in the mean-field result which calls for using the last output only. However in such a case all computational steps should benefit from transmissions from the previous time step. This is not offered by ACT since its design is based on mean-field states and is a key feature of the proposed model. To distinguish the roles among different computational steps, each one should obtain its computation ability and receive its state individually from the previous time step. Thus a more natural design should be a multilayer RNN with a flexible number of layers which is exactly what our proposed model offers. Our experimental results show that ACT has marginal benefits over basic RNN or sequence to sequence (seq2seq) models, indicating that ACT, with a single hidden vector, cannot always work well. This also motivates us to develop the layer flexible RNN model with adaptive computation time.

The novelty of our work is that the number of layers in our model is flexible, so that it can both achieve adaptive computation time and maintain the individual roles among different layers. Similar to Graves’ work, we also utilize a unit to determine the action of each computational step within a time step by calculating their halting probabilities. To obtain the optimal computation ability, each layer should learn from the previous time step individually, and there should be concepts to decide how much to learn from each layer in the previous time step. We face the challenge that the number of layers is different between two consecutive time steps, so that we cannot set specific constant rules of how to transmit the states. In our model, each time step produces multiple hidden states (one state per computational time within the step). These multiple hidden states are then combined into a different number of hidden states for the next step using

attention ideas [15], [16] (the number of new hidden states equals to the number of computational steps in the next step). The network can thus have a flexible number of layers with dynamic number of transmission states.

In this paper, we propose a layer flexible adaptive computation time (LFACT) model for RNNs. Each layer indicates a computational step, produces a hidden state and receives its own transmission state from the previous time step. We also extend the model to the seq2seq framework. Our experimental results show that LFACT offers significant improvements over ACT and RNN on different data sets and frameworks. With LFACT, there is no need to decide the specific structure of an RNN model through extensive experimentation, since LFACT can automatically make decisions of computational steps based on its inputs. LFACT is designed with a different logic in mind from ACT, and at the same time overcomes the problems of ACT, e.g. poor performance on certain data sets. Our model increases the accuracy of 7% to 8% on a financial data set and 12% on Wikipedia language modeling, which attests to its robustness.

The rest of the manuscript is structured as follows. In Section II we review the literature. In Section 3, the flexible layer adaptive computation time RNN model is presented, including all of the alternative options. In Section 4 we introduce the data sets and discuss all the experimental results.

II. LITERATURE REVIEW

A deep learning model and algorithm have many hyperparameters. In an RNN, one of the problems is deciding the computation amount of a certain input sequence. A simple solution is comparing different depths of networks and manually selecting the best option, but a series of expensive training processes is required to make the right decision. Hyperparameter optimization [17], [18] and Bayesian optimization [19]–[21] have been proposed to select an efficient architecture of a network. Based on these concepts, [12] and [11] propose mechanisms for network configuration using reinforcement learning. However, massive training efforts are still present. Another problem of such approaches is the assumption of a fixed structure of the network, irrespective of the underlying sample and step. The difficulty of classification varies in each data set and sample, and it is comprehensible that harder samples would require more computation. Therefore, applying networks with the same computational steps is inflexible and it cannot achieve the goal of flexible computation time among different samples. Conditional computation provides general ideas for alleviating the weaknesses of a fixed-structure deep network by establishing a learning policy [22], [23]. A halt neuron is designed and used as an activation threshold in self-delimiting neural networks [24], [25] to stop an ongoing computation whenever it reaches or exceeds the halting threshold. Work [26] shows that conditional computation helps the networks obtain adaptive depth and thus yield higher accuracy than fixed depth structures. An Adaptive Computation Time (ACT) mechanism for RNN is introduced by [13] to dynamically calculate each input step computation time and

determine their halting condition. These series of work focus on formulating the policies of halting conditions and use a single hidden vector in each cell; none of them contribute to designing flexible multilayer networks or study learning the rules of state transmission.

The ACT mechanism [13] is proved to improve performances and is applied in a few different problems. Universal Transformers [14] apply ACT on a self-attentive RNN to automatically halt computation. A dynamic time model for visual attention [27] is proposed to accelerate the processing time by adding a binary action at each step to determine whether to continue or stop. On their attempt to apply ACT on Residual Networks, [28] show that ACT can dynamically choose the number of evaluated computational steps and propose spatially adaptive computation time for Residual Networks for image processing to adapt the computation amount between spatial positions. Similarly, [29] extend ACT to a recognizing textual entailment task. In addition, ACT is also applied to reduce computation cost and calculate computation time in speech recognition [30], image classification [31], natural language processing [32], and highway networks [33]. These models simply apply the ACT mechanism on other models to achieve the abilities of adaptive halting computations. They focus on solving their specific problems but do not make any change to the structure of ACT cells. However, our work concentrates in the inner design of a layer flexible ACT cell for its ability of automatically and dynamically adapting the number of layers.

III. MODEL

We start with an explanation of RNN and ACT. A standard RNN contains three layers: the input layer, the hidden layer, and the output layer. The input layer receives input sequences x and transmits them to the hidden layer to compute the hidden states u . The output layer calculates the output y based on the updated state of each step. The equations are as follows:

$$u_t = f(x_t, u_{t-1}), \quad y_t = \sigma(W_o u_t + b_o). \quad (1)$$

In step t , input x_t from the input sequence x is delivered to the network. A cell in the hidden layer uses the input x_t and the state u_{t-1} from the previous step to update the hidden state u_t in the current step. Long Short-Term Memory (LSTM) [34] and Gated Recurrent Unit (GRU) [35] are frequently applied in the hidden layer cell f , which contain the dynamic computation information and the activation of the hidden cells. The output y_t is computed utilizing an output weight W_o , an output bias b_o , and an activation function σ .

ACT extends the standard RNN. The hidden layer contains several rounds of computation and each round produces an intermediate state and output. The representation of intermediate states u_t^n and intermediate outputs o_t^n are as follows:

$$u_t^n = \begin{cases} f(x_t^0, u_{t-1}), & n = 0 \\ f(x_t^n, u_t^{n-1}) & n > 0 \end{cases}, \quad x_t^n = (\delta_n, x_t), \quad (2)$$

$$o_t^n = \sigma(W_o u_t^n + b_o).$$

The first hidden cell, in step t , receives the state u_{t-1} from the previous step $t - 1$ and computes the first intermediate

state. All the following rounds of computation use the previous intermediate output u_t^{n-1} and produce an updated state u_t^n . To distinguish different rounds of computation, a flag δ_0 is augmented to the input x_t for the first round and another flag δ_n is added for all others. Each intermediate output o_t^n is computed based on the intermediate state u_t^n in the same round.

To determine the halting condition of a series of rounds of computation, units h_t^n are introduced in each computation round n as $h_t^n = \sigma(W_h u_t^n + b_h)$. Here W_h is the halting weight and b_h is the halting bias.

The total computation time N_t in a step is decided by the halting units and the maximum threshold L . Whenever the accumulated halting units' value in a step t is over 1 or the computation time reaches L , the computation halts. The definition of total computation time N_t is as follows:

$$N_t = \min\{\min\{n \mid \sum_{i=1}^n h_t^i \geq 1 - \epsilon\}, L\}, \quad (3)$$

where ϵ is a hyperparameter.

ACT uses all the intermediate states and outputs to calculate one mean-field state u_t and output y_t (as represented in (4) and (5) below) for each step. A probability p_t^n produced by halting unit h_t^n is introduced into ACT for calculating the mean-field state and output according to the contribution of each intermediate computation round in a step. The updated mean-field state u_t is transmitted to the next input step and the output o_t is delivered to the output layer as the current step's output.

$$p_t^n = \begin{cases} h_t^n, & n < N_t \\ 1 - \sum_{i=1}^{N_t-1} h_t^i, & n = N_t \end{cases} \quad u_t = \sum_{i=1}^{N_t} p_t^i u_t^i \quad (4)$$

$$y_t = \sum_{i=1}^{N_t} p_t^i o_t^i \quad (5)$$

Given an input sequence x , the ACT model tends to compute as much as possible in each step to avoid making erroneous predictions and incurring errors. This can cause an extra computational expense and impede achieving the goal to adapt the computation time. Therefore, training the model to decrease the amount of computation becomes necessary. ACT introduces ponder cost $\mathcal{P}(x)$ as $\mathcal{P}(x) = N_t + p_t^{N_t}$ to represent the total computation time during the input sequence. The loss function $\mathcal{L}(x, gt)$ with gt being the ground truth is modified to encourage the network to also minimize $\mathcal{P}(x)$:

$$\hat{\mathcal{L}}(x, gt) = \mathcal{L}(y(x), gt) + \tau \mathcal{P}(x) \quad (6)$$

where τ is a hyperparameter time penalty that balances the ponder cost and prediction errors.

A. Layer Flexible Adaptive Computation Time Recurrent Neural Network

In this section, our Layer Flexible Adaptive Computation Time (LFACT) model is introduced. The main idea of LFACT

is dynamically adjusting the number of layers according to the imminent characteristic of different inputs and efficiently transmitting each layer's information to the same layer in the next step. Differing from ACT where only the mean-field state u_t in (4) is transmitted to the next step, which can be viewed as a single layer network, LFACT is designed for transmitting each layer's state individually between every consecutive step. In LFACT we compute N_t and N_{t+1} as in ACT. Each cell n (layer n) in step t takes x_t and \hat{u}_{t-1}^n as input and creates u_t^n for $n = 1, \dots, N_t$. Vector \hat{u}_t^n is computed from the output u_t^{n-1} of the previous cell and the hidden state \bar{u}_{t-1}^n from the previous step and same layer n . The problem is that at step t we produce u_t^n for $n = 1, \dots, N_t$ but for step $t+1$ we need \bar{u}_t^n for $n = 1, \dots, N_{t+1}$. The key of our model is to use the attention principle to create $\bar{u}_t^1, \bar{u}_t^2, \dots, \bar{u}_t^{N_{t+1}}$ from $u_t^1, u_t^2, \dots, u_t^{N_t}$. Fig. 1 depicts the model.

The representation of the LFACT model is as follows:

$$\begin{aligned} \hat{u}_{t-1}^n &= \begin{cases} g(\bar{u}_{t-1}^0, u_{t-1}^{N_t-1}), & n = 0 \\ g(\bar{u}_{t-1}^n, u_t^{n-1}) & n > 0 \end{cases}, \\ u_t^n &= f(x_t, \hat{u}_{t-1}^n) \quad n \geq 0, \\ o_t^n &= \sigma(W_o u_t^n + b_o). \end{aligned} \quad (7)$$

The LFACT model contains two types of states. One state u_t^n is the primary output of each hidden cell, which is the same as the states in standard RNN. The other state is the transmission state \bar{u}_t^n that is used for transmitting layer information to the next step. The primary state from previous layer u_t^{n-1} and the transmission state \bar{u}_{t-1}^n from the same layer in the previous time step are combined together through function g . The combined state is delivered to the current cell. Possible options for g are a multi-layer fully connected neural network, or an affine transformation of (x, y) followed by an activation function. In our experiments, we use $g(x, y) = \sigma(W_1 x + W_2 y + b)$.

In step t , the hidden layer cell f uses the input and the combined state from function g to compute and update the primary state u_t^n . The primary states are used to compute the transmission state \bar{u}_t^n for the next step. To avoid possible errors caused by the previous layer, input x_t is directly delivered to each layer as an input. For $n \leq N_{t+1}$, the equations governing the relationship between two transmission states read

$$\begin{aligned} \bar{u}_t^n &= \sum_{i=1}^{c_t^n} \alpha_{tin} u_t^i \\ \alpha_{tin} &= \frac{e^{\beta_{tin}}}{\sum_{j=1}^g e^{\beta_{tjn}}}, \\ \beta_{tin} &= V_n^T \cdot \sigma(W_Q u_{t+1}^i + V_Q u_t^i + b_Q) \quad i \leq c_t^n. \end{aligned} \quad (8)$$

To compute the transmission states \bar{u}_t^n , an attention unit α is introduced to represent the relationship between the primary states u_t^n in a certain layer n and the primary states in other layers. We propose two choices to select c_t^n :

$$c_t^n = \begin{cases} \min(N_t, n), & (a) \\ N_t. & (b) \end{cases} \quad (9)$$

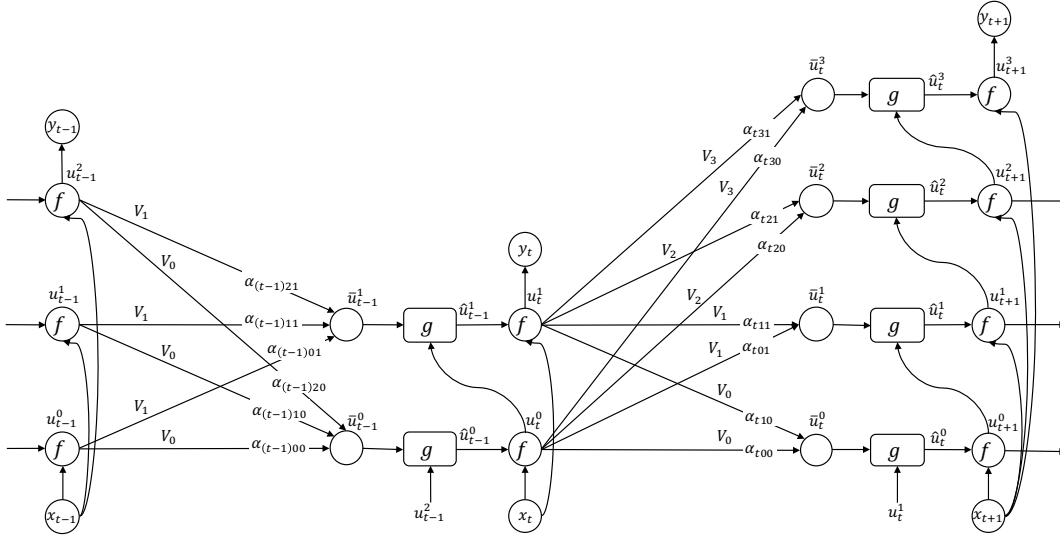


Fig. 1. LFACT model - an example of three consecutive steps. Step $t - 1$ has three layers, step t has two layers, and step $t + 1$ includes four layers.

Option (a) only considers the relationship between the state u_t^n of the current layer and the states u_t^i from the lower layers (i.e. $i \leq n$), called limited (LTD). Alternative (b) utilizes all computed transmission states (i.e. $i \leq N_t$), called ALL. When strategy LTD is applied and $N_{t+1} \leq N_t$, all primary states u_t^i in deeper layers (i.e. $i > N_t$) cannot be used. Strategy ALL aims to include the computed information of all the layers. To distinguish different layers, extra weights V_n are utilized to compute α . Weights V_n , W_Q and V_Q in (8) to compute α are vectors.

We use the same method as ACT to compute N_t (as represented in (3)), the computation time of each step. But unlike ACT, the halting unit is computed based on the output and transmission state of each layer as $h_t^n = \sigma(W_h u_t^n + V_h \bar{u}_{t-1}^n + b_h)$. In addition, instead of computing a mean-field output, we directly take the output of the deepest layer as one step output as $y_t = o_t^{N_t}$.

When applying loss function (6) to LFACT, the shallow layers have limited involvement in calculating gradients. Therefore, to get the prediction of each layer as accurate as possible, we introduce all of the intermediate outputs in the loss function, as

$$\tilde{\mathcal{L}}(x, gt) = \hat{\mathcal{L}}(x, gt) + \mu \sum_{i=0}^{N_t} \bar{\mathcal{L}}(o_t^i(x), gt). \quad (10)$$

In the experiments we use $\bar{\mathcal{L}} = \mathcal{L}$.

B. Sequence to Sequence Model with LFACT

In order to deal with sequence tasks, we propose a combination model using a seq2seq (encoder-decoder) model and our LFACT model (see Appendix A.1). In the seq2seq model, a cell in each step is replaced with our LFACT model to form a deep and flexible network. The seq2seq encoder part accepts a sequence input, and in the decoder part, we use the last ground truth as input.

IV. COMPUTATIONAL EXPERIMENTS

All the models are trained starting with random weights, i.e. no pretraining. Training the LFACT model takes 20% to 30% more time than a typical ACT model. Most experiments are based on a single seed, but in Section 4.2 we conclude that the variance is low if the seed is varied.

A. Financial Data Set

We test our LFACT models on a financial data set from [36]. The data set consists of the tick prices of twenty-two ETFs at five minute intervals. The data is labeled into five classes to represent the significance of the price changes, e.g., one class corresponds to the price being within one standard deviation. We have 22 softmax classification layers in each step. We have three test instances, and in each one we train our model on 50 weeks of returns (45,950 samples), use the next week (905 samples) as validation data to save the best performing weights, and test the model based on the saved weights using the following week (905 samples). Sequences have length 20. The financial data set is tested on both RNN and seq2seq frameworks.

a) *RNN Based Models*: RNN based models predict the next step price changes in each time step. The LFACT model utilizes option affine transformation for g ($g(x, y) = \sigma(W_1 x + W_2 y + b)$) and strategy ALL for computing transmission state \bar{u} ($c_t^n = N_t$). We test plain ACT and RNN, which have been tuned with respect to all hyperparameters as our baseline models, and compare them with the RNN based LFACT model. We apply 0.001 as our ponder time penalty ($\tau = 0.001$) for LFACT and ACT (the value is obtained by the general optimal τ value of the experiments from [13]), and use the Adam optimizer with 0.0005 learning rate to train the models. The maximum number of layers L is 5 and GRU cells with hidden vectors of size 128 are utilized in all the models.

Fig. 2a shows the F1 score improvements of LFACT and ACT over RNN. We test all models on three different instances

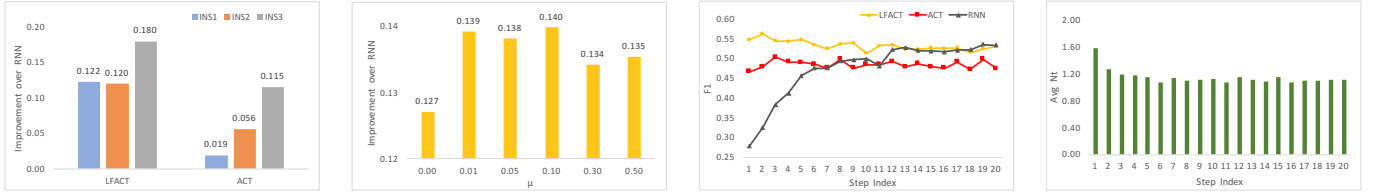


Fig. 2. Results of RNN based models (from left to right). (a) F1 score improvements over RNN on financial data set three instances (INS1, INS2 and INS3). (b) Average F1 improvements over RNN for different μ values on all three instances. (c) Average F1 score at each step on INS1. (d) Average computation time (N_t) for LFACT on INS1.

INS1, INS2, and INS3. Each bar indicates the average F1 score for all prediction steps in an instance. The results of LFACT are based on applying 0.1 to μ in loss (10). The F1 score of RNN is 0.475, 0.461, 0.447 for INS1, INS2, INS3, respectively. From fig. 2a, LFACT improves 14.1% over RNN on average, and ACT improves 6.3%. We introduce the new loss function (10) in order to directly update the weights of each layer from the intermediate outputs. Fig. 2b provides the performance comparison for different μ . The results are the average F1 score improvement over RNN for all three instances. The best range for μ in (10) is 0.01 to 0.1, and is better than the original one in (6) by 1.2%. The application of different μ values shows that our new loss function yields improvements.

Fig. 2c provides the F1 score distribution of steps 1 to 20 on INS1. LFACT consistently performs better than ACT, indicating that multiple layers of hidden vectors bring better effectiveness than a single one. The difficulty of a sequential prediction task is higher in early steps than in late ones, because the early steps have limited information from the input. LFACT and ACT both are stable in all prediction steps, but RNN acts poorly in early predictions. This benefit of LFACT and ACT implies that adaptive computation can contribute to hard tasks. Fig. 2d gives the average computation time (N_t) of each step on the test set of INS1. Higher average N_t of early steps proves LFACT’s ability of deeply computing on hard tasks, and further explains why LFACT is so effective on early predictions.

b) Seq2seq Based Models (10 Prediction Steps): In addition to the RNN framework, we also use the seq2seq version of models to predict the following ten steps. The raw sequence data with input length of 20 is delivered into seq2seq models as the inputs of the encoder part. All hyperparameters are the same as in the RNN based experiments, and the same strategies for g and c as in the RNN based LFACT are applied to the seq2seq framework. Considering that the encoder part does not have outputs, we apply loss function (6) in this task.

In fig. 3a, we present the F1 scores relative changes over seq2seq alone for each instance. The F1 scores of seq2seq are 0.439, 0.481, 0.447. The ACT model is worse than seq2seq on INS3, so the improvement here is negative. From the results, the seq2seq based LFACT improves F1 7.4% over seq2seq, and ACT acts similar to seq2seq. In fig. 3b, we provide the F1 scores for the ten prediction steps in the decoder individually on INS1. All three models decrease over time, but LFACT

and ACT are more stable than seq2seq. In seq2seq based models, the decoder part has constant input of last ground truth, and can cause information deterioration as time passes. Thus, the benefits of LFACT on late predictions over seq2seq alone imply better abilities of LFACT on information transmission and memorization. Surprisingly, the first prediction of seq2seq is better than LFACT, which conflicts the results from RNN. This may be caused by LFACT requiring delay when transforming from input to predictions since it has more trainable weights than seq2seq. However, the whole point of the seq2seq framework is multiple steps of predictions, and LFACT catches up very fast at the second prediction, so the disadvantage of LFACT should not be concerning.

Fig. 3 also presents the computation time (N_t) results for INS1: fig. 3c and 3d are the results of the training and validation process based on the optimized weights, and fig. 3e is for test. The result shows the change of N_t among the different steps, indicating that the LFACT model has the ability of adapting computation time dynamically according to its input. Because of the same input in the decoder, N_t values are the same from step 21 to 30 within each set. In addition, the low N_t values in test set imply that LFACT has low computation request in the decoder part. Thus, the multiple computation ability of LFACT is not the reason for the good performance in the seq2seq setting, as it is in the early predictions in the RNN setting. Comparing to seq2seq alone which contains only one computation time as well in the decoder, the significant benefits in late predictions for LFACT further confirm the conclusion that LFACT has the excellent abilities for information transmission and memorization.

To examine the stability of the LFACT model, we further test the seq2seq based models with 5 prediction steps. The setting is the same as in the 10-prediction case except we have only 5 predictions. Fig. 5a shows the relative F1 scores for LFACT and ACT based on seq2seq alone. The F1 scores for seq2seq on the three instances are 0.492, 0.534, and 0.498. The seq2seq based LFACT performs better than both ACT and seq2seq in the 5-prediction task, and the benefit is significant over ACT. However, the improvement of LFACT over seq2seq is not as pronounced as in the 10-prediction task, and ACT is even worse than seq2seq. Fig. 5b is the F1 score distributions for the three models on INS1. The results match the 10-prediction task, and show that the advantage of LFACT is more likely to affect late predictions in the seq2seq framework.

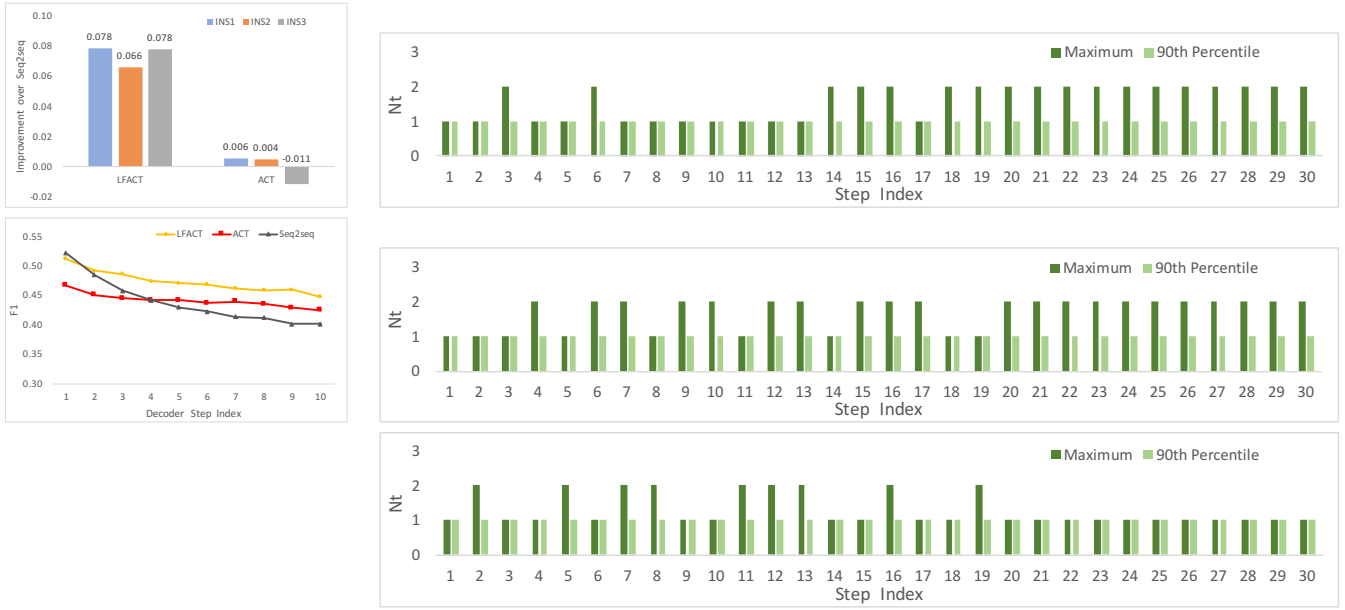


Fig. 3. Results of seq2seq based models. (a - top left) F1 score improvements over seq2seq on financial data set three instances (INS1, INS2 and INS3). (b - middle left) Average F1 score at each step on INS1. (c d e - right column) computation time (N_t) distributions based on optimized LFACT weights on INS1: X-axis is the step index; 1 to 20 indicate encoder; 21 to 30 are from the decoder part.

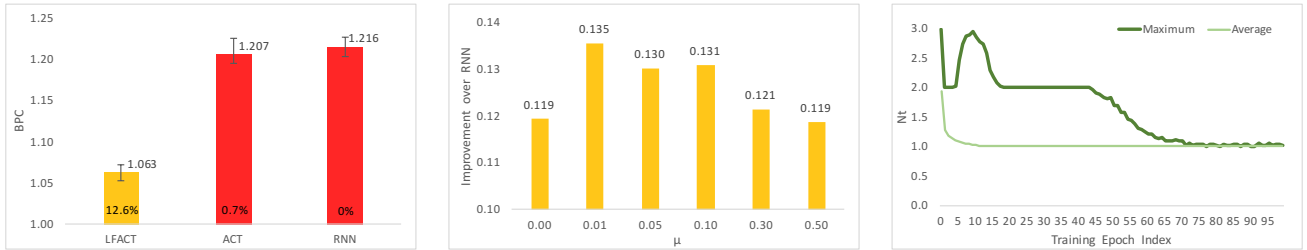


Fig. 4. Results of Wikipedia language modeling task (from left to right). (a) Models performances. Numbers above bars are BPC values and the percentage inside of bars are the relative changes over RNN. (b) Average F1 improvements over RNN for different μ values. (c) Computation time (N_t) change during LFACT model training process.

B. Wikipedia Language Modeling

This task focuses on predicting characters from the Hutter Prize Wikipedia data set, which is also used in [13]. The original unicode text is used without any preprocessing. Every character is represented as one-hot, and presents one time step. In our experiment, 10,240 sequences including 512,000 characters in total are randomly selected as the training set, and 1,280 sequences with 64,000 characters in total are chosen as validation and test sets without repetition. Each sequence includes 50 consecutive characters, and the next character is predicted at each time step in this task (RNN setting). GRU cells with 128 hidden size are used to structure all models. The maximum number of layers L is set to 3, and a softmax layer with size 256 is added to each step in the decoder. We apply the optimized ponder time penalty (τ) 0.06 from Graves' experiments for this task. The models are evaluated using bit per character $BPC = E[\sum_t -\log_2 Pr(x_{t+1}|y_t)]$. Lower BPC values reflect better performances. All results are based on option affine transformation for g ($g(x, y) = \sigma(W_1x + W_2y + b)$) and strategy ALL ($c_t^n = N_t$).

In fig. 4a, we present the experimental results of LFACT and the two baseline models ACT and RNN on the language modeling task. The reported BPC values for LFACT are from different settings of hyperparameter μ in loss (10). Three different random seeds are applied for ACT and RNN to test the stability of the models. Maximum, minimum, and average BPC values are provided. The bars in fig. 4 represent average BPC values, and error bars indicate maximum and minimum BPC. From the experiment, ACT does not have a significant benefit over RNN, but LFACT improves 11.9% over ACT and 12.6% over standard RNN. From the error bars, LFACT has the smallest variance and ACT varies the most. Strong stability for LFACT reflects its better ability to deal with complex situations. To test the influence of the hyperparameter μ in loss function (10), we compare the different settings of μ in fig. 4b. When $\mu = 0$, the loss function is equal to the original one in (6). From fig. 4b, the best range for μ is from 0.01 to 0.1. However, when μ is set to be a larger value ($\mu > 0.3$), the new loss function does not bring any performance improvement over the original loss function.

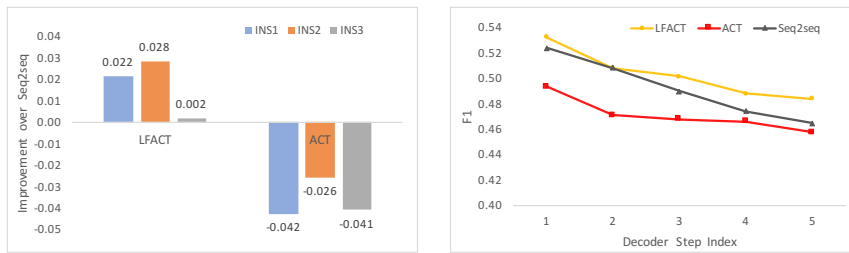


Fig. 5. 5 prediction steps of seq2seq (from left to right). (a) F1 relative changes over seq2seq on the financial data three instances (INS1, INS2, INS3). (b) Average F1 score at each prediction step for seq2seq based models on INS1.

In addition, we test the fully connected network option for $g(x, y)$ and strategy LTD ($c_t^n = \min(N_t, n)$). The fully connected network for g provides 1.074 BPC, and LTD gives 1.678. Neither of them are better than our experimental settings. Therefore, the affine transformation for g and ALL are better strategies for LFACT.

In fig. 4c, we provide the average maximum and average of each step computation time (N_t) during training of the Wikipedia language modeling task. We observe a clear decrease during the early training epochs, which eventually stabilizes. Note that during epochs 5 to 10, the maximum N_t increases but the average N_t still decreases. We postulate that the LFACT model has already obtained the ability to predict most samples during this period, and is putting more effort on the difficult samples. Appendix Appendix A.2 shows the Maximum N_t distributions of training, validation, and test based on the optimized weights. We only present the last 25 steps; the first 25 steps are all 1. The distributions show that the LFACT model is able to keep the computation time as low as possible, but also has the ability of deep computation for certain samples. With the optimized weights, only 0.03% of the sequences in the training set have more than one computation time, and validation and test sets have 0.24% and 0.16% of the sequences with multiple computation. This difference happens because the model is trained based on the training set, and the model should have learned the most efficient way to predict characters in the training set.

We test LFACT on different training sizes ranging from 100,000 characters in total to 10 million, as fig. 6 shows. As the training set size increases, our model achieves better performance and eventually gets around 0.99, which indicates scalability. We conclude that LFACT consistently has over 7% improvement on all of the training sizes over ACT and RNN. Due to the computational resource limitations, all the results in Section IV-B, including hyperparameter comparison, are based on 512,000 characters and 10,240 sequences training size, and 64,000 characters, 1,280 sequences test size.

V. CONCLUSION

Deciding the structure of recurrent neural networks has been a problem in deep learning applications, in particular the number of computational steps. A halting unit is applied in a previous work to adapt the computation time to inputs, but a single hidden vector structure leads to information

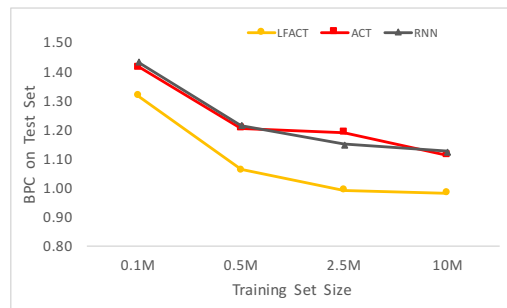


Fig. 6. Performance on different training size

transmission weaknesses. We propose LFACT which utilizes an attention strategy in designing an information transmission policy which leads to a flexible multilayer recurrent neural network with adaptive computation time. LFACT can automatically adjust computation time according to the computing complexity of inputs and has outstanding dynamic information transmission abilities between consecutive time steps. We apply LFACT in an RNN and a seq2seq setting and evaluate the model on a financial data set and Wikipedia language modeling. The experimental results show a significant improvement of LFACT over RNN and seq2seq and ACT on both data sets. The different number of layers in practice indicates LFACT’s ability of adapting computation time and information transmission.

VI. SOFTWARE AND DATA

If the paper is accepted, the source code for the Wikipedia experiments will be open sourced on github.

REFERENCES

- [1] M. Auli, M. Galley, C. Quirk, and G. Zweig, “Joint language and translation modeling with recurrent neural networks,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.
- [2] S. Liu, N. Yang, M. Li, and M. Zhou, “A recursive recurrent neural network for statistical machine translation,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014.
- [3] A. Graves, A.-r. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.
- [4] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates *et al.*, “Deep speech: Scaling up end-to-end speech recognition,” *arXiv preprint arXiv:1412.5567*, 2014.

- [5] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.
- [6] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-RNN)," *International Conference on Learning Representations*, 2015.
- [7] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription," *Proceedings of the 29th International Conference on Machine Learning*, 2012.
- [8] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
- [9] A.-r. Mohamed, G. E. Dahl, G. Hinton *et al.*, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.
- [10] R. K. Srivastava, K. Greff, and J. Schmidhuber, "Training very deep networks," in *Advances in Neural Information Processing Systems*, 2015.
- [11] B. Baker, O. Gupta, N. Naik, and R. Raskar, "Designing neural network architectures using reinforcement learning," *International Conference on Learning Representations*, 2017.
- [12] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *International Conference on Learning Representations*, 2017.
- [13] A. Graves, "Adaptive computation time for recurrent neural networks," *arXiv preprint arXiv:1603.08983*, 2016.
- [14] M. Dehghani, S. Gouws, O. Vinyals, J. Uszkoreit, and Ł. Kaiser, "Universal transformers," *arXiv preprint arXiv:1807.03819*, 2018.
- [15] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
- [16] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015.
- [17] J. S. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, 2011.
- [18] J. Bergstra, D. Yamins, and D. D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- [19] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems*, 2012.
- [20] H. Mendoza, A. Klein, M. Feurer, J. T. Springenberg, and F. Hutter, "Towards automatically-tuned neural networks," in *Workshop on Automatic Machine Learning*, 2016.
- [21] S. Saxena and J. Verbeek, "Convolutional neural fabrics," in *Advances in Neural Information Processing Systems*, 2016.
- [22] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.
- [23] Y. Bengio, P.-L. Bacon, J. Pineau, and D. Precup, "Conditional computation in neural networks for faster models," *International Conference on Learning Representations*, 2015.
- [24] J. Schmidhuber, "Self-delimiting neural networks," *arXiv preprint arXiv:1210.0118*, 2012.
- [25] R. K. Srivastava, B. R. Steunebrink, and J. Schmidhuber, "First experiments with powerplay," *The 2nd Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 2013.
- [26] C. Ying and K. Fragkiadaki, "Depth-adaptive computational policies for efficient visual tracking," in *International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, 2017, pp. 109–122.
- [27] Z. Li, Y. Yang, X. Liu, F. Zhou, S. Wen, and W. Xu, "Dynamic computational time for visual attention," *IEEE International Conference on Computer Vision Workshop*, 2017.
- [28] M. Figurnov, M. D. Collins, Y. Zhu, L. Zhang, J. Huang, D. P. Vetrov, and R. Salakhutdinov, "Spatially adaptive computation time for residual networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [29] M. Neumann, P. Stenetorp, and S. Riedel, "Learning to reason with adaptive computation," *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*, 2016.
- [30] M. Li and M. Liu, "End-to-end speech recognition with adaptive computation steps," *arXiv preprint arXiv:1808.10088*, 2018.
- [31] S. Leroux, P. Molchanov, P. Simoons, B. Dhoedt, T. Breuel, and J. Kautz, "Iamnn: Iterative and adaptive mobile neural network for efficient image classification," *Workshop on International Conference on Learning Representations*, 2018.
- [32] A. W. Yu, H. Lee, and Q. Le, "Learning to skim text," *The 55th Annual Meeting of the Association for Computational Linguistics*, 2018.
- [33] H. Park and C. D. Yoo, "Early improving recurrent elastic highway network," *arXiv preprint arXiv:1708.04116*, 2017.
- [34] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [35] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.
- [36] M. Harmon and D. Klabjan, "Dynamic prediction length for time series with sequence to sequence networks," *arXiv preprint arXiv:1807.00425*, 2018.