

Unsupervised Learning of Sequence Representations by Autoencoders

Wenjie Pei, David M.J. Tax

Pattern Recognition Laboratory, Delft University of Technology
wenjiecoder@gmail.com, D.M.J.Tax@tudelft.nl

Abstract

Traditional machine learning models have problems with handling sequence data, because the lengths of sequences may vary between samples. In this paper, we present an unsupervised learning model for sequence data, called the Integrated Sequence Autoencoder (ISA), to learn a fixed-length vectorial representation by minimizing the reconstruction error. Specifically, we propose to integrate two classical mechanisms for sequence reconstruction which takes into account both the global silhouette information and the local temporal dependencies. Furthermore, we propose a stop feature that serves as a temporal stamp to guide the reconstruction process, and which results in a higher-quality representation. Extensive validation on real-world datasets shows that the learned representation is able to effectively summarize not only the apparent features, but also the underlying and high-level style information. Take for example a speech sequence sample: our ISA model can not only recognize the spoken text (apparent feature), but can also discriminate the speaker who utters the audio (more high-level style).

1 Introduction

Sequence data is ordered in space or time, which is common around us such as video, audio or text data. Traditional machine learning models for fixed length observations are challenged by sequence data. An important reason is that the length of sequence examples is variable. Typically, sequence models are specifically designed to deal with sequence data for extensive applications ranging from computer vision [Karpathy *et al.*, 2014; Pei *et al.*, 2017a] to natural language processing [Bahdanau *et al.*, 2015]. However, it is not an efficient way to redesign the entire existing single-observation models to cater to sequence data. It would be beneficial if we can transform sequence data to the standard fixed-length feature representation, which can then be readily accessed by any existing traditional machine learning models.

In this paper, we aim to learn a fixed-length representation given a sequence. To be applicable to both labeled and unlabeled sequence data, the representation should be learned

in an unsupervised way. Besides, the obtained representation should be able to capture the informative features contained in the sequence. One straightforward way is to employ the Dynamic Time Warping (DTW) algorithm [Vintsyuk, 1968; Sakoe and Chiba, 1978]. To be specific, we can first construct a sequence vocabulary for reference and then calculate the pairwise distances from a target sequence sample to all the samples in the vocabulary. The achieved distance vector can be used as the representation to describe this target sample. Since DTW is good at aligning two sequences which share a similar shape but may vary in speed, this representation can capture the shape information well. However, it is difficult for DTW to explore more latent features that is not related to the global shape. For instance, while DTW can easily identify the audio content, it is difficult for DTW to recognize the speaker who utters the audio.

Here we introduce an unsupervised learning model for sequences, called Integrated Sequence Autoencoder (ISA), to learn a fixed-length vectorial representation. Specifically, our model integrates seamlessly two classical mechanisms for sequence reconstruction that takes into account both the global and local features. The representation is learned by minimizing the reconstruction error and the whole model can be trained in an end-to-end manner efficiently. Extensive experiments demonstrated that the learned representation by our model can not only capture the apparent shape information as well as DTW, but also summarize the underlying and high-level style information that DTW does not capture.

One key difference that distinguishes our model from the classical sequence-to-sequence supervised models used for machine translation [Bahdanau *et al.*, 2015] or text generation [Li *et al.*, 2015] is that those models predict words out of a vocabulary as a classification task in decoding. An ending token is added into the vocabulary to indicate a full stop during the sequence generation (reconstruction). In contrast, our model reconstructs the input sequence as a regression task. We predict/reconstruct a continuous feature value for each time step. It is hard for our model to perceive the temporal progress during decoding. To address this issue, we propose a stop feature. This acts as a temporal stamp to guide the reconstruction process. It can potentially help the model to better align the reconstructed sequence with the input one, which improves the quality of the representation.

2 Related Work

It is straightforward to employ Dynamic Time Warping (DTW) algorithm [Vintsyuk, 1968; Sakoe and Chiba, 1978] to extract representation for a sequence by calculating the pairwise distance to a vocabulary of sequences. While the representation obtained in this way is able to capture the global shape information well due to the specialty of DTW, it is challenging for it to explore more latent features which is not related to shape information. Here we propose to combine the ideas from autoencoders and sequence-to-sequence learning. We discuss related work on both topics separately below.

Autoencoder is a type of neural network that is designed to learn a latent representation for samples by minimizing the reconstruction error [Rumelhart *et al.*, 1986]. It can be used for dimension reduction [Hinton and Salakhutdinov, 2006] or unsupervised pre-training [Erhan *et al.*, 2010]. Various variations of autoencoders have been proposed to improve the quality of learned representation, such as Denoising autoencoder [Vincent *et al.*, 2010], Sparse autoencoder [Ng, 2009] or Variational Autoencoder [Kingma and Welling, 2014]. Autoencoder is typically used for single-observation tasks, which means the length of input is fixed. In contrast, our model focuses on learning representation for sequence data whose length is variable.

Sequence-to-sequence Learning aims to learn a mapping function from an input sequence to an output sequence. It has been extensively studied in many tasks like machine translation [Bahdanau *et al.*, 2015; Cho *et al.*, 2014] or text generation [Li *et al.*, 2015; Graves, 2013]. It is generally achieved by first encoding the input sequence to a latent representation and then decoding it into an output sequence [Sutskever *et al.*, 2014]. Typically, the decoding phase is formulated as a supervised classification task, making it hard to predict real-valued sequences. Our ISA model tries to avoid this and uses a regression formulation.

3 Model

Given an input sequence of variable length, our goal is to learn a vectorial representation in an unsupervised way. The learned representation is required to: (1) have fixed length, which can be readily used for further processing tasks, (2) summarize high-level features covering not only the holistic information contained in the whole sequence, but also the temporal dependencies between observations. Our proposed model is composed of two core modules: a Holistic autoencoder and an Atomistic autoencoder. They are integrated seamlessly via a loss function and the whole model can be trained in an end-to-end manner efficiently. We will elaborate on these two modules first and then present our Integrated Sequence Autoencoder. The graphical structure of the model is illustrated in Figure 1.

3.1 Holistic Sequence Autoencoder

The goal of Holistic Sequence Autoencoder is to learn a hidden representation which is able to capture the holistic features for the given sequence. This is achieved by reconstructing the full sequence from only one hidden representation

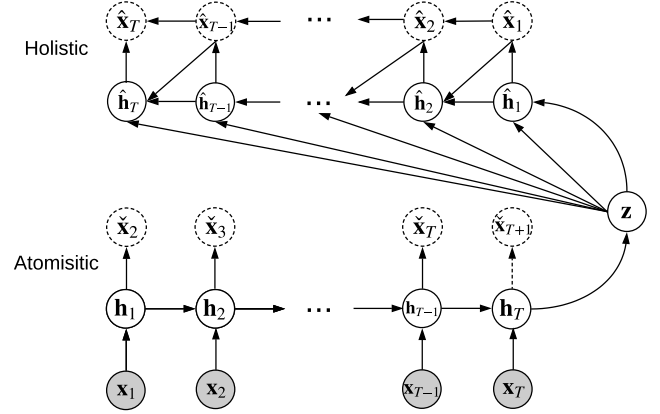


Figure 1: The graphical structure of our Integrated Sequence Autoencoder. Note that \tilde{x}_{T+1} is a virtual node not used in practice.

vector z . Similar to the classical autoencoder, The Holistic Sequence autoencoder consists of two parts: (1) the encoder which is used to encode the input data into a hidden representation z , and (2) the decoder for reconstructing the input sequence from the hidden representation. We employ Long Short-Term Memory (LSTM) [Hochreiter and Schmidhuber, 1997] as the backbone network for both the encoder and decoder due to its powerful modeling for the sequence data.

Encoder. The encoder of the Holistic Sequence Autoencoder is in line with the routine forward pass of LSTM model. Formally, given an input sequence $\mathbf{x}_{1,...,T} = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$ of length T in which $\mathbf{x}_t \in \mathbb{R}^D$ denotes the observation at the t -th time step of the sequence, the hidden state \mathbf{h}_t at time step t of the encoder is modeled as the element-wise product \odot between the output gate \mathbf{o}_t of the LSTM at time step t and the cell state \mathbf{C}_t (transformed by \tanh):

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{C}_t), \quad (1)$$

where the cell state \mathbf{C}_t is calculated by a weighted sum of the candidate cell state at the current time step $\tilde{\mathbf{C}}_t$ and the cell state in the previous step \mathbf{C}_{t-1} :

$$\mathbf{C}_t = \mathbf{f}_t \odot \mathbf{C}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{C}}_t. \quad (2)$$

Herein, \mathbf{f}_t and \mathbf{i}_t are respectively the forget gate and input gate to control the information flow from the previous and current time step. All three gates (\mathbf{i}_t , \mathbf{f}_t , \mathbf{o}_t) and the candidate cell state $\tilde{\mathbf{C}}_t$ are modeled based on the previous hidden state and current input observation in a similar nonlinear way. For instance, the output gate \mathbf{o}_t (at t -th time step) is computed by:

$$\mathbf{o}_t = \sigma(\mathbf{W}_o \cdot \mathbf{h}_{t-1} + \mathbf{U}_o \cdot \mathbf{x}_t + b_o). \quad (3)$$

Herein, \mathbf{W}_o and \mathbf{U}_o are the transformation matrices and b_o is the bias term. The hidden state \mathbf{h}_T at the last time step is employed as the summary representation z for the sequence that we aim for:

$$z = \mathbf{h}_T. \quad (4)$$

Decoder. To optimize the learning of the hidden representation z , our Holistic Sequence Autoencoder attempts to reconstruct the whole input sequence $\mathbf{x}_{1,...,T}$ from z and minimize

the reconstruction error to make the reconstructed sequence as close as possible to the original sequence. As a result, the learned representation \mathbf{z} is encouraged to preserve all the informative features contained in the sequence. Specifically, we employ another LSTM model as our decoder to reconstruct the sequence step by step. The reconstructed observation $\hat{\mathbf{x}}_t$ at time step t is modeled as:

$$\hat{\mathbf{x}}_t = \mathbf{M} \cdot g(\mathbf{A} \cdot \mathbf{z} + \mathbf{B} \cdot \hat{\mathbf{x}}_{t-1} + \mathbf{E} \cdot \hat{\mathbf{h}}_t), \quad (5)$$

which takes into account the summary representation \mathbf{z} , the hidden state $\hat{\mathbf{h}}_t$ in current step and the reconstructed observation in the previous step $\hat{\mathbf{x}}_{t-1}$. Here \mathbf{M} , \mathbf{A} , \mathbf{B} and \mathbf{E} are mapping matrices and a rectified linear unit (ReLU) function [Nair and Hinton, 2010] is used as the activation function g . The hidden state $\hat{\mathbf{h}}_t$ is derived similarly to the hidden state \mathbf{h}_t of the Encoder (Equation 1 and 2). It should be noted that the calculation of the three gates ($\hat{\mathbf{i}}_t$, $\hat{\mathbf{f}}_t$, $\hat{\mathbf{o}}_t$) and candidate cell state $\hat{\mathbf{C}}'_t$ is slightly different from Equation 3, since they also rely on the summary representation \mathbf{z} apart from the hidden state $\hat{\mathbf{h}}_t$ and reconstructed observation $\hat{\mathbf{x}}_{t-1}$. For any $\hat{\mathbf{s}} \in \{\hat{\mathbf{i}}_t, \hat{\mathbf{f}}_t, \hat{\mathbf{o}}_t, \hat{\mathbf{C}}'_t\}$, it is modeled as:

$$\hat{\mathbf{s}} = \sigma(\hat{\mathbf{W}} \cdot \hat{\mathbf{h}}_{t-1} + \hat{\mathbf{U}} \cdot \hat{\mathbf{x}}_{t-1} + \hat{\mathbf{V}} \cdot \mathbf{z} + \hat{\mathbf{b}}). \quad (6)$$

Parameter Learning. Suppose we are given a training set $\mathcal{D} = \{\mathbf{x}_{1, \dots, T^{(n)}}^{(n)}\}_{n=1, \dots, N}$ containing N sequences of length $T^{(n)}$. We learn all parameters involved in a Holistic Sequence Autoencoder jointly by minimizing the reconstruction error. In particular, we minimize the mean squared error between the reconstructed sequence and the original input sequence for each training sample in \mathcal{D} :

$$\mathbf{L}_h = \sum_{n=1}^N \left(\frac{1}{T^{(n)}} \sum_{t=1}^{T^{(n)}} \|\hat{\mathbf{x}}_t^{(n)} - \mathbf{x}_t^{(n)}\|^2 \right). \quad (7)$$

3.2 Atomistic Sequence Autoencoder

In contrast to the Holistic Sequence Autoencoder, which specializes in capturing the holistic features, the Atomistic Sequence Autoencoder is designed to focus on modeling the local temporal dependencies between adjacent steps in a sequence. It also employs a LSTM model as its backbone, but the difference from Holistic Sequence Autoencoder is that it performs encoding and decoding using only one LSTM module. As illustrated in Figure 1, we first encode the input observation \mathbf{x}_t at time step t into the hidden state \mathbf{h}_t , which is exactly same as the Holistic Sequence Autoencoder (Equation 1, 2 and 3). Then the hidden state \mathbf{h}_t is decoded to reconstruct the observation $\hat{\mathbf{x}}_{t+1}$ for the next time step:

$$\hat{\mathbf{x}}_{t+1} = \mathbf{P} \cdot g(\mathbf{F} \cdot \mathbf{h}_t), \quad (8)$$

where \mathbf{P} and \mathbf{F} are linearly mapping matrices. Similar to the Holistic Sequence Autoencoder, all the parameters of Atomistic Sequence Autoencoder are learned by minimizing the reconstruction error for the training samples:

$$\mathbf{L}_a = \sum_{n=2}^N \left(\frac{1}{T^{(n)}} \sum_{t=1}^{T^{(n)}} \|\hat{\mathbf{x}}_t^{(n)} - \mathbf{x}_t^{(n)}\|^2 \right). \quad (9)$$

Since the observation at each time step is reconstructed from the hidden state in the previous time step, the Atomistic Sequence Autoencoder tends to focus on the accuracy of local transitions between time steps in the reconstructed sequence rather than the global outline.

3.3 Integrated Sequence Autoencoder

To assimilate the merits of both Holistic Sequence Autoencoder and Atomistic Sequence Autoencoder, we propose to integrate the two autoencoders. The resulting model is called Integrated Sequence Autoencoder (ISA). Figure 1 shows that two autoencoders can be seamlessly integrated by sharing one LSTM module. The shared LSTM module is not only used for performing encoding of both two models, but also responsible for the decoding of the Atomistic Sequence Autoencoder. The Integrated Sequence Autoencoder is trained in an end-to-end manner by computing the weighted sum of Loss functions of two autoencoders:

$$\mathbf{L} = \alpha \cdot \mathbf{L}_h + (1 - \alpha) \cdot \mathbf{L}_a, \quad (10)$$

where $\alpha \in [0, 1]$ is a hyper-parameter that balances the impact of two sub-autoencoders.

3.4 Stop Feature

To help our model track the temporal progress during decoding, we propose a stop feature to guide the reconstruction process. Specifically, we concatenate a scalar value $v_t \in [0, 1]$ to the feature vector for each time step $[\mathbf{x}_t v_t]^T$. The value of the stop feature increases smoothly as the time step moves forward. Higher value (closer to 1) indicates a closer position to the end of the sequence. Hence the stop feature acts like a temporal stamp stick to each time step of a sequence. We investigate three different mechanisms for the stop feature (shown in Figure 2):

- **Linear** mechanism: the stop feature value increases linearly with time. Mathematically, the stop feature v_t at time step t of a sequence with total length T is defined as:

$$v_t = \frac{t}{T}. \quad (11)$$

- **Tanh** mechanism: the increase of v is rapid at the beginning of the sequence, but slows down at the end. The stop feature v_t at time step t is modeled as:

$$v_t = \tanh\left(\gamma \cdot \frac{t}{T}\right) + 1 - \tanh(\gamma), \quad (12)$$

where $\gamma > 0$ is a hyper-parameter to determine the gradient.

- **Exp** mechanism: the stop feature value increases exponentially. The stop feature v_t at time step t is modeled as (for $\gamma > 0$):

$$v_t = \exp\left(\gamma \cdot \frac{t - T}{T}\right). \quad (13)$$

Intuitively, the stop feature can potentially help the model to memorize the relative temporal location and thus better align the reconstructed sequence with the input sequence. Consequently, it leads to higher quality of learned representation, which is demonstrated by experiments in Section 4.4.

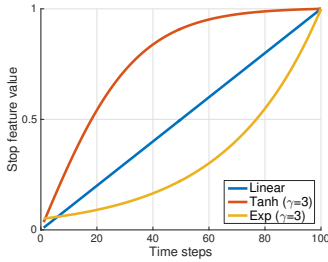


Figure 2: Three mechanisms for stop feature.

4 Experiments

We first conduct experiments to compare our ISA model to the DTW algorithm, in which we aim to investigate what is learned by the representations of both models. Then we perform comparisons between the Holistic, Atomistic models and our proposed ISA model, and validate our theoretical analysis regarding their respective advantages. Subsequently, experiments are performed to evaluate the effectiveness of the stop feature and its effect on sequence encoding. Finally, we evaluate our model by feeding the learned representation to a classifier (i.e., a Support Vector Machine, SVM) and compare the classification performance to other state-of-the-art sequence classification models.

4.1 Experimental Setup

Datasets

We conduct experiments on three datasets, selected to show generalization across different tasks and modalities: (1) a on-line handwritten character dataset (OHC) [Williams *et al.*, 2008], (2) a audio dataset of Arabic spoken digits [Hammami and Bedda, 2010] and (3) the Cohn–Kanade extended facial expression dataset (CK+) [Lucey *et al.*, 2010].

The online handwritten character (OHC) data set [Williams *et al.*, 2008] is a pen-trajectory sequence dataset that comprises three dimensions of features at each time step: the pen movement in the x and y directions, and the pen pressure. The dataset contains 2858 sequences with an average length of 120 time steps. Each sequence corresponds to a handwritten character that has one of 20 labels.

The Arabic spoken digit dataset [Hammami and Bedda, 2010] consists of 8800 utterances, which were collected by asking 88 Arabic speakers to utter each of 10 digits ten times. Each sequence comprises 13-dimensional Mel-Frequency Cepstral Coefficients (MFCCs), which were sampled at 11,025Hz, 16 b using a Hamming window. We use two different versions of this dataset: (1) a *digit* version in which the uttered digit is the class label (10 classes) and (2) a *voice* version in which the speaker is the class label (88 classes).

The Cohn–Kanade extended facial expression dataset (CK+) [Lucey *et al.*, 2010] consists of 593 image sequences (videos) from 123 subjects. Each video presents a single facial expression. We use a subset of 327 videos in our experiments, which have validated labels corresponding to one of seven emotions (anger, contempt, disgust, fear, happiness, sadness, and surprise). We adopt the shape features used in [van der Maaten and Hendriks, 2012] as the feature representation. Each frame is represented by the variation

of 68 feature point locations (x , y) with respect to the first frame [Lucey *et al.*, 2010], which results in 136-dimensional feature representation for each frame in the video.

Settings

We perform 5-fold cross validation on OHC dataset, and 10-fold cross validation on CK+ dataset across all the experiments. For the Arabic (digit) dataset, we use the same data division as Hammami and Bedda [Hammami and Bedda, 2010]: 6600 samples as training set and 2200 samples as test set. There is no speaker overlap between them. Regarding the Arabic (voice), we use the samples of the first 8 digits as training set and the left out samples as test set.

For all the recurrent networks mentioned in this work, the number of hidden units is tuned by selecting the best configuration from the candidate set $\{32, 64, 128\}$ with a validation set set side from the training set. The loss function weight α of Integrated Sequence Autoencoder (Equation 10) and the hyper-parameter γ of the Stop feature in Equation 12 and 13 are tuned by minimizing the reconstruction error on the validation set. We perform gradient descent optimization using RMSprop [Tieleman and Hinton, 2012] to train our model. The gradients are clipped in the interval $[-5, 5]$ [Bengio *et al.*, 2013]. As a way to evaluate our model and DTW, we perform classification on the learned representations using SVM classifier [Cortes and Vapnik, 1995; Chang and Lin, 2011] with radial basis kernel. The kernel parameter and regularization parameter of the SVM are tuned with the validation set.

4.2 Comparison to DTW

The most intriguing question to investigate is whether or not our model learns the same information as DTW. We perform experiments on the Arabic (digit) and Arabic (voice) datasets. Specifically, given a test sample, we employ DTW to calculate the pairwise distances from it to all training samples as its DTW-based representation, whose dimensionality is therefore the number of training samples. Then we perform comparisons between this DTW-based representation and the representation learned by our model, both in supervised and unsupervised ways.

Evaluation of Classification performance. Typically, a better representation results in better classification performance. We first evaluate the classification performance of both representations (from DTW and our model) by feeding them to the same classifier, i.e., an SVM with the same settings. The experimental results are presented in Figure 3. Both models perform quite well on Arabic (digit) dataset, which is a relatively easy task. It is not surprising for DTW since it is good at aligning two sequences which share a similar shape but may vary in speed, which fit well with digit classification. However, it fails to capture the discriminative information for voice (speaker) classification while our ISA model significantly outperforms DTW. It demonstrates that our model is able to learn the high-level representation which captures not only the global shape feature but also the underlying characteristics. Herein, classification on Arabic (voice) is much more challenging than Arabic (digit)* since there is no digit overlap between training and test sets for Arabic (voice).

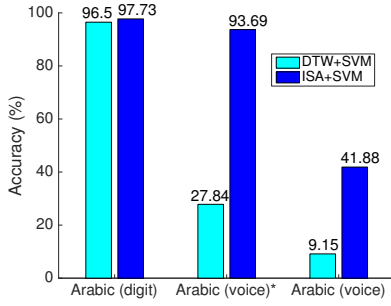


Figure 3: The comparison of classification performance of SVM on representations learned by DTW and Integrated Sequence Autoencoder (ISA) with stop feature on Arabic (digit) and Arabic (voice) datasets. Note that Arabic (voice) has no digit overlap between training and test set while all 10 digits can be seen in both training and test sets for Asteriskd Arabic voice* dataset.

Clustering Visualization. To obtain more insight into what representation both models have been learned on the Arabic (voice) dataset, we apply t-SNE [van der Maaten and Hinton, 2008]. Figure 4 shows t-SNE maps of 50 utterances of the same digit from 5 randomly selected speakers for both models. In the DTW two main clusters can be distinguished, but within a cluster speakers are confused. The ISA representation, on the other hand more strongly focus the individual speakers (although not perfectly: the cluster of the speaker indicated by ‘*’ misses two samples).

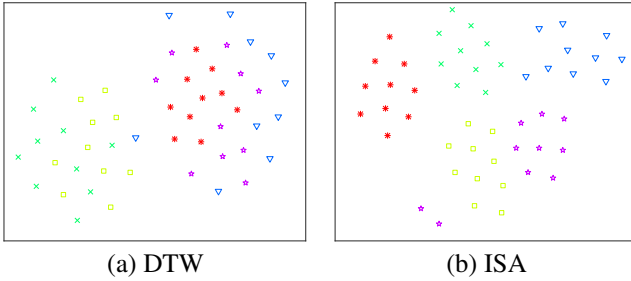


Figure 4: t-SNE maps of the Arabic (voice) test data for the same digit uttered by 5 randomly selected speakers (indicated by different colors and markers), constructed based on (a) DTW distance (to each training sample) and (b) our Integrated Sequence Autoencoder (ISA). The DTW distance measure fails to discriminate the samples between voice of different speakers while our Integrated Sequence Autoencoder is able to handle it well.

4.3 Comparison of Different Autoencoders

Next we compare the representations learned by the Holistic, Atomistic and Integrated sequence autoencoders. In particular, we perform classification on the learned representations with an SVM on four different tasks. The experimental results in Table 1 show that the Holistic autoencoder outperforms the Atomistic on the Arabic (digit) dataset whilst Atomistic exhibits better performance on voice classification. It is consistent with the theoretical analysis that Holistic Autoencoder specializes in capturing the global feature while the Atomistic autoencoder focuses on the local temporal information. Similarly, Holistic autoencoder performs much better than Atomistic autoencoder on OHC dataset whose discrimi-

native feature is mainly concerning the global shape of a sequence. The ISA, which is designed to assimilate the merits of both Holistic and Atomistic autoencoders, performs best among all the tasks.

Table 1: Classification accuracy (%) on 4 datasets by performing SVM on the learned representation of different autoencoders. The best performance per dataset is shown in bold.

Dataset	Model		
	Atomistic	Holistic	ISA
Arabic (digit)	68.36	96.23	97.41
Arabic (voice)	32.10	26.88	35.63
OHC	53.01	92.26	94.15
CK+	96.43	92.50	96.79

To qualitatively evaluate the performance of reconstruction, we visualize the reconstructed sequence of character examples from OHC dataset. Figure 5 presents two examples which are not well handled by Holistic autoencoder, probably because it is confused between similar handwritten ‘e’ and ‘o’, also between handwritten ‘a’ and ‘u’. It shows that the ISA is able to reconstruct more accurate sequences than Holistic autoencoder.

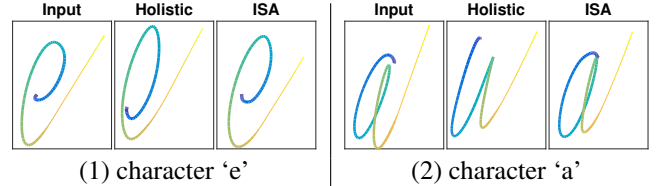


Figure 5: Visualization of some (hard) reconstructed characters (in OHC dataset) by Holistic autoencoder and Integrated Sequence Autoencoder (ISA). Both subfigures shows that ISA is able to reconstruct more accurate sequences than Holistic autoencoder. The pen trajectory moves from blue color to yellow color. Thicker line indicates higher pen pressure.

4.4 The Importance of the Stop Feature

Validation of functionality with synthetic dataset. We first validate the functionality of the stop feature with a synthetic dataset. We construct the synthetic data by drawing circles with the same radius but with a different number of loops (repetitions). Specifically, two classes of circle samples are constructed: the first class consists of 100 circle samples with 2 loops and the other class contains 100 samples with 3 loops. Each sample is represented as a sequence of 2-dimensional features (x and y coordinates) for each time step. The variable length of sequences is sampled from a uniform distribution within an interval [50, 200].

We learn the representations of the circle dataset using our Integrated Sequence Autoencoder, with and without stop features respectively. The representations are fed into the SVM classifier to perform classification. Table 2 shows that ISA without any stop feature performs poorly on discriminating between circles of 2 loops and circles of 3 loops. It is not surprising since the number of loops is the only discriminative information available, and this is hardly captured by the baseline ISA (without stop feature). In contrast, all three different stop features enable ISA to classify the circles of different loops perfectly.

Table 2: Classification accuracy (%) on 4 datasets by performing SVM on the learned representation of Integrated Sequence Autoencoder (ISA) with and without the stop features. The best performance per dataset is boldfaced.

Dataset	Model			
	ISA	ISA+Linear	ISA+Tanh	ISA+Exp
Circle (synthetic)	77.5	100	100	100
Arabic (digit)	97.41	97.05	97.73	96.86
Arabic (voice)	35.63	41.88	36.48	37.61
OHC	94.15	93.12	95.01	94.66

Classification performance on real datasets. Next we conduct similar classification experiments on 3 real datasets to evaluate the stop feature. The experimental results in table 2 show that overall the stop feature *Tanh* and *Linear* perform better than *Exp*.

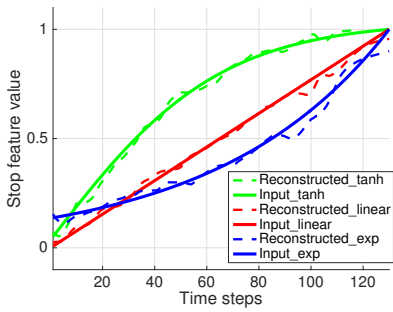


Figure 6: Visualization of the stop feature of a reconstructed character sample (in OHC dataset) by Integrated Sequence Autoencoder (ISA). The dashed lines indicate the reconstructed values and solid lines are the ground truth (input stop feature values).

To gain more insight, we visualize the stop feature values of a reconstructed character sample by the Integrated Sequence Autoencoder with three mechanisms in Figure 6. It shows that the reconstructed stop features are generally aligned well with the ground truth. Furthermore, the *Tanh* mechanism performs more stable and accurate than the other two, especially at the tail of the sequence. It is probably because the *Exp* grows exponentially, hence it is getting harder for the autoencoder to predict the values at tail part of the sequence. It is consistent with the quantitative results in Table 2 that the stop feature *Tanh* and *Linear* overall outperform *Exp*.

4.5 Comparison to State-of-the-art Sequence Classifiers

In the last set of experiments, we evaluate our model by feeding the learned representation to an SVM classifier and compare the classification performance to other state-of-the-art sequence classifiers. It is actually not completely fair because these sequence classifiers train their model using label information while our model has to learn the representation in an unsupervised way.

The sequence classifiers we use as baseline models are: (1) Hidden-Unit Logistic Model (HULM) [Pei *et al.*, 2017b], which is a graphical model for sequence classification employing binary hidden units to model latent structure in the

data, and (2) LSTM-based classifier, which is mounted a softmax layer and cross entropy loss upon the hidden representation of last time step of LSTM to perform classification. Furthermore, the DTW-based representation (with the SVM classifier) is also taken into comparison.

Table 3 presents the classification performance of 4 classifiers on 5 different tasks. Our model and DTW+SVM perform very well on Arabic (digit) task and even outperform the other two dedicated sequence classifiers. It indicates that both our model and DTW can capture the shape information of the digits properly. In contrast, DTW performs fairly poor on the Arabic (voice) task while our model achieves a comparable performance with LSTM model. It is quite a challenging task since the training and test set have no digit overlap, and the models have to generalize to new unseen digits. Our model (as well as DTW) has to learn and preserve both discriminative digit information and voice information simultaneously in one representation while the supervised sequence classifiers are taught by the labels to focus on the voice classification due to the steering of the label information.

We then investigate an easier case (Arabic (voice*)) in which all 10 digits appear in both the training and test set. In this scenario, the performance of our model is on par with the state-of-the-art (by HULM). Regarding the CK+ dataset whose discriminative information about facial expression is concerning both global and local features, our model achieves the best result and substantially outperforms other models while DTW does not perform well. It demonstrates the advantage of our model over DTW that our model is able to capture the features in both holistic and atomistic levels.

Table 3: Classification accuracy (%) on all 5 problems by different models. Note that there is no digit overlap between training and test set in Arabic (voice) data while all 10 digits appear in both training and test sets for Asterisk Arabic voice* dataset.

Dataset	Model			
	HULM	LSTM	DTW+SVM	ISA-Stop + SVM
Arabic (digit)	95.32	96.05	96.50	97.73
Arabic (voice)	56.14	43.75	9.15	41.88
Arabic (voice)*	94.55	93.01	27.84	93.69
OHC	97.66 (± 1.15)	98.08 (± 1.48)	97.88 (± 1.08)	95.01 (± 0.65)
CK+	93.56 (± 4.83)	94.64 (± 3.47)	80.36 (± 6.13)	96.43 (± 4.12)

Leveraging the unlabeled data for semi-supervised learning. Since our model works in unsupervised way, it can be potentially used in a semi-supervised learning scenario. Specifically, we can employ our model to learn high-quality sequence representations by training it on a large amount of unlabeled data which is easily obtained, and then train a classifier (like SVM) on an available labeled dataset which is typically small. Figure 7 presents a simulation of such semi-supervised learning scenario for our model. We consider 20% of the training data as the labeled data and the left 80% are unlabeled. We train our ISA model with increasing training data (without using labels!) while training the SVM classifier with the same subset of labeled 20% data constantly. We also list the classification performance for other classifiers (LSTM, HULM and DTW+SVM) as a reference. We observe that in-

creasing the (unlabeled) training data for our model can substantially improve the classification performance of the subsequent SVM, which indicates that leveraging the unlabeled data indeed improves the quality of learned representations by our model.

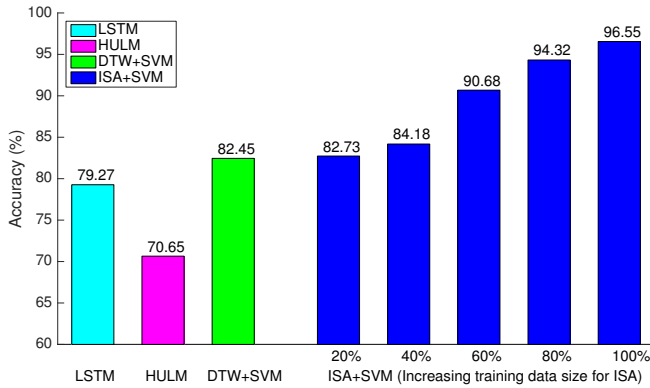


Figure 7: The Simulation of leveraging the unlabeled data for semi-supervised learning scenario on Arabic (digit) dataset with our ISA model. All classifiers (LSTM, HULM, SVM) are trained with 20% training data. To simulate the semi-supervised learning scenario, our ISA model is trained with increasing training data size (20% to 100%) without using labels while the associated SVM is trained with the same subset of 20% (labeled) training data constantly.

5 Conclusion

In this work, we present the Integrated Sequence Autoencoder (ISA), an unsupervised learning model to learn a fixed-length representation for varying-length sequences. The model combines the ideas from autoencoders and sequence-to-sequence learning, and integrates two classical mechanisms of sequence reconstruction. The representation learned by our model is able to capture both the global shape information and local temporal dependencies contained in the sequence data. Furthermore, we propose a stop feature to guide the reconstruction process, which helps the model to better align the reconstructed sequence with the input sequence and thus improves the quality of the learned representation. We show the generalization of our approach on three datasets that are across different tasks and modalities. As future work, the denoising of the sequence by reconstruction could be a promising direction.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Bengio *et al.*, 2013] Yoshua Bengio, Nicolas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *ICASSP*, pages 8624–8628, 2013.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines.

ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[Cortes and Vapnik, 1995] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[Erhan *et al.*, 2010] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660, 2010.

[Graves, 2013] Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013.

[Hammami and Bedda, 2010] N. Hammami and M. Bedda. Improved tree model for Arabic speech recognition. In *Int. Conf. on Computer Science and Information Technology*, pages 521–526, 2010.

[Hinton and Salakhutdinov, 2006] Geoffrey Hinton and Ruslan Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504 – 507, 2006.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.

[Kingma and Welling, 2014] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Li *et al.*, 2015] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. In *ACL (1)*, pages 1106–1115, 2015.

[Lucey *et al.*, 2010] P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.

[Nair and Hinton, 2010] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, pages 807–814, 2010.

[Ng, 2009] Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 2009.

[Pei *et al.*, 2017a] Wenjie Pei, Tadas Baltrusaitis, David M.J. Tax, and Louis-Philippe Morency. Temporal attention-gated model for robust sequence classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Pei *et al.*, 2017b] Wenjie Pei, Hamdi Dibeklioğlu, David M. J. Tax, and Laurens van der Maaten. Multivariate time-series classification using the hidden-unit logistic model.

- [Rumelhart *et al.*, 1986] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1. chapter Learning Internal Representations by Error Propagation, pages 318–362. 1986.
- [Sakoe and Chiba, 1978] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26:43–49, 1978.
- [Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. 2014.
- [Tieleman and Hinton, 2012] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural Networks for Machine Learning*, 4(2), 2012.
- [van der Maaten and Hendriks, 2012] Laurens van der Maaten and Emile Hendriks. Action unit classification using active appearance models and conditional random fields. *Cognitive Processing*, 13(2):507–518, 2012.
- [van der Maaten and Hinton, 2008] L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [Vincent *et al.*, 2010] Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408, 2010.
- [Vintsyuk, 1968] T.K. Vintsyuk. Speech discrimination by dynamic programming. *Kibernetika*, 4:81–88, 1968.
- [Williams *et al.*, 2008] B.H. Williams, M. Toussaint, and A.J. Storkey. Modelling motion primitives and their timing in biologically executed movements. In *NIPS*, pages 1609–1616, 2008.