様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ８ 月 ２５ 日現在

機関番号：１２６０５
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８０６８

研究課題名（和文）A Sequence-to-sequence Model based Dissimilarity Measurement for Clustering Structural Data

研究課題名（英文）A Sequence-to-sequence Model based Dissimilarity Measurement for Clustering Structural Data

研究代表者
ＮＧＵＹＥＮＴＵＡＮ　ＣＵＯＮＧ（NGUYENTUAN, CUONG）

東京農工大学・工学（系）研究科（研究院）・特任助教

研究者番号：１０８１４２４６

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：手書き数式答案をクラスタリングするため，ニューラルネットワークのSeq2Seqモデルを利用し，時系列入力パータンの距離を計算する方法を提案した．この手法は，Deep Embedded ClusteringやSiamese Networksなどのグローバル特徴抽出手法より良い精度を確認した．提案手法も多段階の畳み込みニューラルネットワークの特徴抽出手法を向上することが出来た．オンライン手書き数式答案の編集距離と比べると提案した距離が優れているとの結果を得られた．引き続きこの方法を，予備試験から収集した大規模なオフライン手書き数式答案のデータベースに適用する．

研究成果の学術的意義や社会的意義
大規模な手書き数式回答をクラスタリングできると，同じ回答がグループ化され，採点する手間を削減し，採点の効率と信頼性を向上する．本研究は，クラスタリングするため，構造認識とそれらの関係を学習することの重要性を強調している．

研究成果の概要（英文）：We have finished applying the proposed generative sequence dissimilarity for clustering of handwritten mathematical answers. The method outperforms other global feature based clustering methods such as Deep Embedded Clustering and Siamese Networks. The method also superior to the hierarchical feature representations by Convolutional Neural Networks with Weakly Supervised learning. We have applied the method for clustering online handwritten mathematical expressions and show that the proposed metric is better than edit distance metric. We continue to apply the method for a large-scale database of offline handwritten mathematical answers collected from the preliminary examination.

研究分野：pattern recognition, machine learning

キーワード：clustering　online handwriting　offline handwriting　generative sequence　sequence to sequence

１．研究開始当初の背景

Clustering structural data is important for engineering, business and economics, financial technology, biology especially DNA sequence analyses and so on. A structural sample can be represented as a sequence. A tree-structured sample can be represented by a sequence through tree traversal [J. Morris, Information Processing Letters, 9(5), 197-200, 1979]. In a word sequence of text, the structure is implicitly represented and it could be identified as a syntax tree by grammatical tagging [E. Charniak, AI Magazine, 18(4), 33-44, 1997]. For a mathematical expression, whose structure is specified by 2-dimensional layout, is represented as an expression tree, which can be serialized into a sequence.

Conventional clustering algorithms face the difficulty of dealing with structural data due to the high complexity of samples and difficulty of defining the distance between samples.

Taking the complexity first, structural data are likely to be composed of many sub-structures and the relations of these sub-structures. For example, an image contains many objects as its sub-structures, a mathematical expression contains many expressions as sub-structures and relations among them. Successful approaches for dealing with structural data in classification apply the hierarchical analysis from low-level features of sub-structures to the high-level features of combining those sub-structures. Convolutional Neural Networks (CNN) have been state-of-the-art in Image Classification [A. Krizhevsky et al., NIPS, 1097-1105, 2012], Object Detection [S. Ren et al., NIPS, 91-99, 2015]. Unsupervised learning learns the features to represent samples in a lower dimensionality without requiring labeled data [G. Hinton et al., Science, 313(5786), 504-507, 2006]. On the other hand, supervised learning learns the sub-structures inside the samples using the labels associated to the sub-structures in samples, thus reduces the complexity of the structural samples. Semi-supervised learning [D. Kingma et al., NIPS, 3581-3589, 2014] performs learning using both the labeled data and unlabeled data without the large cost for labeling all the samples.

Then, the second major problem is defining the distance. Since the number of sub-structures of each structural sample is different, it is necessary to transform structural samples to feature sequences and measure the distance between two sequences. A structural sample is transformed to a feature sequence by sequentially extracting the features of sub-structures. For clustering sequences, the distance calculated by matching sequences may not employ the structural information of the sequences. Model-based approaches use a parametric model such as a Hidden Markov Model (HMM) [P. Smyth, NIPS, 648-654, 1997] [M. Bicego, MLDM, 86-95, 2003] or a Poisson model [D. Witten et al., Annals of Applied Statistics, 5(4), 2493-2518, 2011] to model each pattern then use the model to measure dissimilarity between each pair of patterns. These approaches are costly since they need to use an individual model to model each sequence. For sequential data, Recurrent Neural Networks and Long-short Term Memory (LSTM), which are widely applied in sequence recognition including handwriting recognition [A. Graves et al., NIPS, 577-584, 2007], speech recognition [A. Graves et al., ICASSP, 6645-6649, 2013], show their advantages in processing these data.

Although many attempts have been made to tackle with the high complexity of samples and the difficulty of defining the distance between samples, there remains the problem of combining the solution of these two problems to deal with clustering structural data.

２．研究の目的

In this approach, we apply an end-to-end semi-supervised learning model to transcribe a structural sample into a label sequence. The model firstly transforms the structural sample into a feature sequence. For images, we apply a CNN to extract image features, then scan through the columns of image features to get the feature sequence. For sequential samples, we apply a LSTM through an input sequence to get the feature sequence. The feature sequence is put through a Seq2Seq learning model to learn how to transcribe a feature sequence into a label sequence. The model is learned by the semi-supervised learning method from a large scale of samples with limited labeled samples. This reduces the cost for labeling the whole training samples. For this approach, we could deal with the structural data which are either sequential or non-sequential. Therefore, the method is appropriate for both online (collections of pen coordination) and offline (images) handwriting samples. We also apply the attention mechanism which supplies the structural information by focusing on the sub-structures of the data.

３．研究の方法

To measure the structural dissimilarity between two inputs, we derive the probability of generating an input sequence from another input sequence by the learned Seq2Seq model. The overview of the method is illustrated in Fig. 1. Firstly, the two input samples S1 and S2 are transformed into feature sequences by a CNN if they are images or by a LSTM if they are sequences. The sequence encoder encodes these feature sequences into fixed-size feature vectors and the decoder decodes them into two label sequences L1 and L2. We feed L1 to the sequence decoder of S2 to obtain the probability $P(L1|S2)$ to generate L1 from S2 and vice versa to obtain the probability $P(L2|S1)$. To get the distance $d(S1, S2)$, we average these two conditional probabilities to make it symmetric.

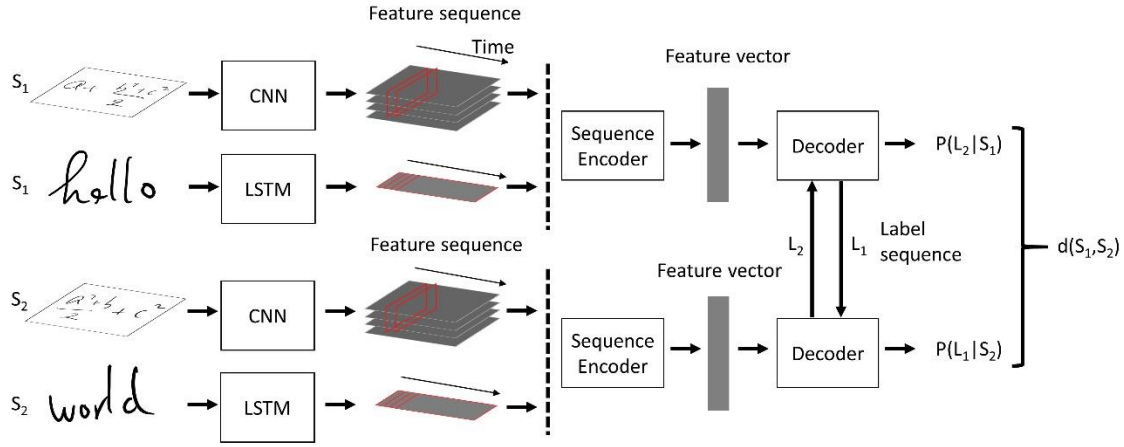

Figure 1. Distance of two input sequences

We illustrate the method to calculate the probability to generate a label sequence from an input sequence in Fig. 2. The feature sequence of the first input S1 is fed to the sequence encoder. The predicted probability to generate the output label sequence of the second input S2 from S1 is obtained by feeding L2 = XYZ as the input label sequence for the decoder. Here, X, Y, Z, are the labels of the label sequence L2. The probability outputs which are associated to L2 are P(X), P(Y), P(Z) and P(<eos>), which are P(X|S1), P(Y|X,S1), P(Z|XY,S1), P(<eos>|XYZ,S1), respectively. We obtain the conditional probability by the following formulae:

$$P(L_2|S_1) = P(XYZ < eos > |S_1)$$
$$= P(X|S_1)P(Y|X,S_1)P(Z|XY,S_1)P(< eos > |XYZ,S_1) \quad (1)$$

４．研究成果

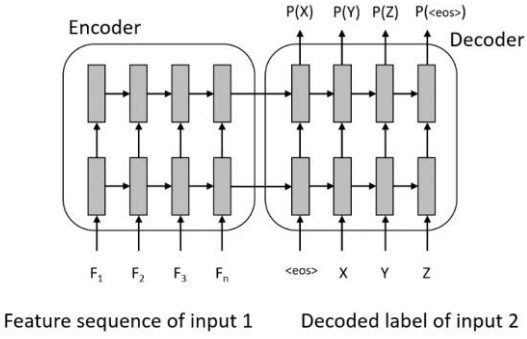

Figure 2. Probability to generate label sequence from feature sequence

We focus on the problem of clustering handwritten answers, for both online handwriting data (pen-traces) and offline handwriting data (image of handwriting). First, we proposed a CNN-based method to learn both localization and classification representations of mathematical symbols in handwritten formula images. Symbols in various scales are located and classified by multi-level features of multi-scaled CNN. We train the CNN networks by weakly supervised training and fine-tune them by symbols attention to enhance classification and location prediction. Multi-level spatial representations are extracted from the CNN for calculating the distance. Experiments on our collected datasets and the CROHME dataset show promising results. We also prepared a method for clustering online handwritten mathematical expressions using BLSTM-CTC for recognizing label sequence and pyramid histogram of characters for sequence embedding. Secondly, we have finished applying the proposed generative sequence dissimilarity for the clustering of handwritten mathematical answers. The method outperforms other clustering methods which do not focus on local features such as Deep Embedded Clustering and Siamese Networks. The method is also superior to the hierarchical feature representations by Convolutional Neural Networks with Weakly Supervised learning. We have applied the method for clustering online handwritten mathematical expressions and show that the proposed metric is better than the edit distance metric. We continue to apply the method for a large-scale database of offline handwritten mathematical answers collected from the preliminary examination. This dataset is challenging since the answers contain both handwritten text and handwritten mathematical expressions.

We have also improved the recognition performance of handwritten mathematical expressions. We have developed online and offline handwritten mathematical expression recognition using seq2seq with an attention mechanism and weakly supervised learning. Our recognition system is ranked 3rd in an official offline handwritten mathematical competition organized by the International Conference on Frontiers of Handwriting Recognition.

We will extend our research to other metric learning methods and unsupervised learning to improve the robustness of the clustering method. We also consider generalizing the proposed methods for other datasets rather than handwriting.

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　2件／うちオープンアクセス　1件）

１．著者名 Ung Huy Quang、Nguyen Cuong Tuan、Phan Khanh Minh、Khuong Vu Tran Minh、Nakagawa Masaki	４．巻 146
２．論文標題 Clustering online handwritten mathematical expressions	５．発行年 2021年
３．雑誌名 Pattern Recognition Letters	６．最初と最後の頁 267～275
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patrec.2021.03.027	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 KHUONG Vu-Tran-Minh、PHAN Khanh-Minh、UNG Huy-Quang、NGUYEN Cuong-Tuan、NAKAGAWA Masaki	４．巻 E104.D
２．論文標題 Clustering of Handwritten Mathematical Expressions for Computer-Assisted Marking	５．発行年 2021年
３．雑誌名 IEICE Transactions on Information and Systems	６．最初と最後の頁 275～284
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2020EDP7087	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Ly Nam Tuan、Nguyen Cuong Tuan、Nakagawa Masaki	４．巻 136
２．論文標題 An attention-based row-column encoder-decoder model for text recognition in Japanese historical documents	５．発行年 2020年
３．雑誌名 Pattern Recognition Letters	６．最初と最後の頁 134～141
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patrec.2020.05.026	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nguyen Cuong Tuan、Khuong Vu Tran Minh、Nguyen Hung Tuan、Nakagawa Masaki	４．巻 131
２．論文標題 CNN based spatial classification features for clustering offline handwritten mathematical expressions	５．発行年 2020年
３．雑誌名 Pattern Recognition Letters	６．最初と最後の頁 113～120
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patrec.2019.12.015	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nguyen Cuong Tuan、Indurkhya Bipin、Nakagawa Masaki	４．巻 23
２．論文標題 A unified method for augmented incremental recognition of online handwritten Japanese and English text	５．発行年 2019年
３．雑誌名 International Journal on Document Analysis and Recognition (IJDAR)	６．最初と最後の頁 53～72
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s10032-019-00343-y	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

１．著者名 Nguyen Cuong Tuan、Nguyen Hung Tuan、Mita Kazuhiro、Nakagawa Masaki	４．巻 121
２．論文標題 Robust and real-time stroke order evaluation using incremental stroke context for learners to write Kanji characters correctly	５．発行年 2019年
３．雑誌名 Pattern Recognition Letters	６．最初と最後の頁 140～149
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patrec.2018.07.025	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Nguyen Hung Tuan、Nguyen Cuong Tuan、Ino Takeya、Indurkhya Bipin、Nakagawa Masaki	４．巻 121
２．論文標題 Text-independent writer identification using convolutional neural network	５．発行年 2019年
３．雑誌名 Pattern Recognition Letters	６．最初と最後の頁 104～112
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.patrec.2018.07.022	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計19件（うち招待講演　0件／うち国際学会　15件）

１．発表者名 Huy Quang Ung, Cuong Tuan Nguyen, Hung Tuan Nguyen, Masaki Nakagawa
２．発表標題 GSSF: A Generative Sequence Similarity Function based on a Seq2Seq model for clustering online handwritten mathematical answers
３．学会等名 Proceedings of International Conference on Document Analysis and Recognition（国際学会）
４．発表年 2021年

１．発表者名
Cuong Tuan Nguyen, Thanh-Nghia Truong, Hung Tuan Nguyen, Masaki Nakagawa

２．発表標題
Global Context for improving recognition of Online Handwritten Mathematical Expressions

３．学会等名
Proceedings of International Conference on Document Analysis and Recognition（国際学会）

４．発表年
2021年

１．発表者名
Thanh-Nghia Truong, Cuong Tuan Nguyen, Khanh Minh Phan, Masaki Nakagawa

２．発表標題
Improvement of End-to-End Offline Handwritten Mathematical Expression Recognition by Weakly Supervised Learning

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2020年

１．発表者名
Cuong Tuan Nguyen, Thanh-Nghia Truong, Huy Quang Ung, Masaki Nakagawa

２．発表標題
Online Handwritten Mathematical Symbol Segmentation and Recognition with Bidirectional Context

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2020年

１．発表者名
Kha Cong Nguyen, Cuong Tuan Nguyen, Seiji Hotta, Masaki Nakagawa

２．発表標題
A character attention generative adversarial network for degraded historical document restoration

３．学会等名
Proceedings of International Conference on Document Analysis and Recognition（国際学会）

４．発表年
2019年

１．発表者名
Nam Tuan Ly, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
An attention-based end-to-end model for multiple text lines recognition in japanese historical documents

３．学会等名
Proceedings of International Conference on Document Analysis and Recognition（国際学会）

４．発表年
2019年

１．発表者名
Nam Tuan Ly, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Attention Augmented Convolutional Recurrent Network for Handwritten Japanese Text Recognition

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2020年

１．発表者名
Trung Tan Ngo, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
A Siamese Network-based Approach For Matching Various Sizes Of Excavated Wooden Fragments

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2020年

１．発表者名
Kha Cong Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
A Semantic Segmentation-based Method for Handwritten Japanese Text Recognition

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2020年

１．発表者名
Hung Tuan Nguyen, Tsubasa Nakamura, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Online trajectory recovery from offline handwritten Japanese kanji characters of multiple strokes

３．学会等名
Proceedings of International Conference on Pattern Recognition, ICPR2020（国際学会）

４．発表年
2020年～2021年

１．発表者名
Kei Morizumi, Cuong Tuan Nguyen, Ikuko Shimizu, Masaki Nakagawa

２．発表標題
CNN and 2D BLSTM for Local Feature Extraction in Handwritten Mathematical Expression Recognition

３．学会等名
IEICE Technical Report, PRMU2020-56

４．発表年
2020年

１．発表者名
Huy Quang Ung, Vu Tran Minh Khuong, Anh Duc Le, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Bag-of-features for clustering online handwritten mathematical expressions

３．学会等名
International Conference on Pattern Recognition and Artificial Intelligent（国際学会）

４．発表年
2018年

１．発表者名
Vu Tran Minh Khuong, Huy Quang Ung, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Clustering Offline Handwritten Mathematical Answers for Computer-Assisted Marking

３．学会等名
International Conference on Pattern Recognition and Artificial Intelligent（国際学会）

４．発表年
2018年

１．発表者名
Hung Tuan Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Online Japanese Handwriting recognizers using Recurrent Neural Networks

３．学会等名
International Conference on Frontiers of Handwritting Recognition（国際学会）

４．発表年
2018年

１．発表者名
Nam Tuan Ly, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Training and End-to-End Model for Offline Handwritten Japanese Text Recognition by Generated Synthetic Patterns

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2018年

１．発表者名
Hung Tuan Nguyen, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
ICFHR2018-Competition on Vietnamese Online Handwritten Text Recognition using HANDS-VNOnDB

３．学会等名
Proceedings of International Conference on Frontiers in Handwriting Recognition（国際学会）

４．発表年
2018年

１．発表者名
Xiuyu Liang, Shinsuke Sasaki, Cuong Tuan Nguyen, Masaki Nakagawa

２．発表標題
Improvement of a Computer Automated Marking System for Online Handwritten Math Answers employing Machine Recognition

３．学会等名
IEICE Technical Report, PRMU2018-135

４．発表年
2019年

１．発表者名 佐藤旭，小林心，Nam Tuan Ly，Cuong Tuan Nguyen，北本朝展，中川正樹
２．発表標題 日本古典籍くずし字文書の文字列認識
３．学会等名 情報処理学会技術報告，Vol．2019-CH-119，No．4，pp．1-4
４．発表年 2019年

１．発表者名 Hung Tuan Nguyen，Cuong Tuan Nguyen，Masaki Nakagawa，Asanobu Kitamoto
２．発表標題 Text Segmentation for Japanese Historical Documents using Fully Convolutional Neural Network
３．学会等名 情報処理学会技術報告，Vol．2019-CH-119，No．4，pp．5-9
４．発表年 2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関