AFRL-AFOSR-VA-TR-2019-0152

Sparse Factor Analysis for Information Extraction and Fusion

Richard Baraniuk
WILLIAM MARSH RICE UNIV HOUSTON TX

05/21/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
05-01-2019	Final Report	06-01-2015 to 05-31-2018

4. TITLE AND SUBTITLE
Sparse Factor Analysis for Information Extraction and Fusion

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-14-1-0088

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Richard G. Baraniuk
richb@rice.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
William Marsh Rice University
6100 Main St
Houston, TX 77005

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 N. Randolph, Ste.325
Arlington, Virginia 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Statistical models have been developed to incorporate preferences of each instructor in recommending personalized learning action. Such models can provide feedback to instructors to help them understand their preferences. In addition, novel methods have been proposed to use neural networks to design data-driven question generation models. A new novel criterion is also proposed to evaluate the performance and relevance of such models. Furthermore, we have proposed a method for automatic short answer grading. Finally, stronger generative models have been developed using DRM framework. Over the past year, we have made significant progress on education data processing in five directions: 1) Statistical models for instructor content preference analysis, 2) Data-driven question generation models, 3) Criteria for Neural Question Generation Models, 4) Meta-learning model for automatic short answer grading, 5) Generative models using deep learning

15. SUBJECT TERMS
human-in-the-loop machine learning, learning analytics, matrix factorization

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Richard G. Baraniuk
U	U	U	UU		19b. TELEPHONE NUMBER *(Include area code)* 713-348-5132

A Latent Factor Model for Instructor Content Preference Analysis

Existing personalized learning systems (PLSs) have primarily focused on learners' data analytics. In this paper, we extend the capability of current PLSs by proposing latent factor model that analyzes instructors' question preferences in a particular subject given question exclusion data from each instructor as well as each question's Bloom Taxonomy tag. Such analytics enable a PLS to recommend personalized learning actions that cater to not only individual learners' learning progress but also the preferences of each instructor. Additionally, such analytics enable a PLS to provide feedback to the instructors to help them better understand their preferences. We formulate the problem of characterizing instructors' question exclusion behavior as the matrix factorization problem, and explicitly incorporate expert-labeled Bloom's Taxonomy tags on each question as a factor in our statistical model. Experimental results on a real-world educational dataset demonstrates that the proposed model achieves superior performance compared to several other baseline algorithms commonly used in recommendation systems. Additionally, by explicitly incorporating Bloom's Taxonomy, the model can provide meaningful interpretations that help understand how and why instructors exclude certain questions.

QG-Net: A Data-Driven Question Generation Model for Educational Content

The ever-growing amount of educational content renders it increasingly difficult to manually generate sufficient practice or quiz questions to accompany it. This paper introduces QG-Net, a recurrent neural network-based model specifically designed for automatically generating quiz questions from educational content such as textbooks. QG-Net, when trained on a publicly available, general-purpose question/answer dataset and without further fine-tuning, is capable of generating high quality questions from textbooks, where the content is significantly different from the training data. Indeed, QG-Net outperforms state-of-the-art neural network-based and rules-based systems for question generation, both when evaluated using standard benchmark datasets and when using human evaluators. QG-Net also scales favorably to applications with large

amounts of educational content, since its performance improves with the amount of training data.

NLL-QA: A Simple, Effective Selection Criterion for Neural Question Generation Models

We propose a simple criterion, NLL-QA, to replace the negative log likelihood (NLL) criterion universally employed in state-of-the-art neural question generation models to select the final output question. NLL-QA incorporates an external question answering model that explicitly evaluates the relevance of a generated question to the input answer from which it is generated, a crucial characteristic of a good question that negative log likelihood fails to adequately capture. We evaluate the utility of our proposed criterion using qualitative, quantitative, and human evaluation metrics. Experimental results demonstrate that, compared to the traditional negative log likelihood criterion, NLL-QA enables neural question generation models to select better questions that are more relevant to the input answer without sacrificing grammatical correctness.

A Meta-Learning Augmented Bidirectional Transformer Model for Automatic Short Answer Grading

We introduce ml-BERT, an effective machine learning method for automatic short answer grading when training data, i.e., graded answers, is limited. Our method combines BERT (Bidirectional Representation of the Transformer), the state-of-the-art model for learning textual data representations, with meta-learning, a training framework that leverages additional data and learning tasks to improve model performance when labeled data is limited. Our intuition is to use meta-learning to help us learn an initialization of the BERT parameters in a specific target subject domain using unlabeled data, thus fully leveraging the limited labeled training data for the grading task. Experiments on a real-world student answer dataset demonstrate the promise of ml-BERT method for automatic short answer grading.

DEEP LEARNING

Generative models that can capture latent variations in data are difficult to design in complex domains such as natural images where there are large numbers of nuisance variables.
Given the success of the Convolutional Neural Networks (CNNs) on inference tasks in such domains, we aim to design generative models whose inference corresponds to the CNNs.
One such class is the Deep Rendering Model (DRM). The DRM generates images via multiple levels of abstraction, from coarse to fine scale, and introduces a small set of latent variables at each level. However, a number of simplifying assumptions are made in the DRM to derive the CNNs as the bottom-up inference algorithm which do not correspond to realistic variation in natural images. For instance, the latent variables at different scales of the DRM are assumed to be independent. We propose an extension to the DRM, termed the Neural Rendering Model (NRM), that enforces dependencies among the latent variables via a parametrized joint prior distribution. This joint prior yields a new form of regularization for training the CNNs---the Rendering Path Normalization (RPN). Under the NRM, we obtain consistent estimators for unsupervised/semi-supervised learning tasks and derive generalization bounds. Our bound suggests that the RPN regularization helps improve generalization, an observation that is corroborated in practice.

Conditional likelihood estimation in the NRM yields the cross-entropy loss for training the CNN. This loss function has some shortcomings: it is unable to capture the true uncertainty in classifying objects on a given image. This is because the cross-entropy function only updates information for the correct object categories on each training example, and fails to incorporate uncertainties about the incorrect categories. We propose a simple and efficient alternative to cross entropy, termed as Max-Min cross-entropy. This arises as a combination of cross-entropy of any standard CNN with cross-entropy of another CNN with suitable modifications. These two networks are co-trained with shared weights and we term it as a Max-Min neural network. The second CNN follows the same architecture as the given CNN, but uses minimum pooling in place of maximum pooling and negative rectified linear units (i.e., min (, 0)) in place of ReLUs. We show that the

resulting max-min cross entropy is able to simultaneously utilize information from both correct and incorrect labels, and hence, improves CNN training. We design the Max-Min cross entropy in a principled way: it arises as bounds on the max-marginal and min-marginal of a deep generative model known as the Neural Rendering Model (NRM). Our experiments demonstrate that the Max-Min cross-entropy improves CNN's performance for supervised learning and has significant advantage for semi-supervised learning when combined with NRM on benchmarks including SVHN, CIFAR10, CIFAR100, and ImageNet.