Machine Learning Test conference

Wednesday 21 February 2018 - Saturday 24 February 2018

Bonn

Book of Abstracts

Contents

1

Dropout: A Simple Way to Prevent Neural Networks from

Authors: Georey Hinton[1]; Iyer Sutskever[None]

[1] *University of Toronto*

Corresponding Authors: abc@gmail.com, efg@gmail.com

Deep neural nets with a large number of parameters are very powerful machine learning systems. However, overtting is a serious problem in such networks. Large networks are also slow to use, making it dicult to deal with overtting by combining the predictions of many dierent large neural nets at test time. Dropout is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of dierent \thinned" networks. At test time, it is easy to approximate the eect of averaging the predictions of all these thinned networks by simply using a single unthinned network that has smaller weights. This signicantly reduces overtting and gives major improvements over other regularization methods. We show that dropout improves the performance of neural networks on supervised learning tasks in vision, speech recognition, document classication and computational biology, obtaining state-of-the-art results on many benchmark data sets.

2

Distilling the Knowledge in a Neural Network

Authors: Oriol Vinyals[1]; Geoffrey Hinton[1]

[1] *Google*

Corresponding Authors: vinyals@google.com, geoffhinton@google.com

A very simple way to improve the performance of almost any mac
hine learning algorithm is to train many different models on the same data a
nd then to average their predictions [3]. Unfortunately, making predictions
using a whole ensemble of models is cumbersome and may be too computationally expen sive to allow deployment to a large number of users, especially if the indivi dual models are large neural nets. Caruana and his collaborators [1] have shown that it is possible to compress the knowledge in an ensemble into a single model which is much easier to deploy and we develop this approach further using a different compression technique. We achieve some surprising results on MNIST and w e show that we can significantly improve the acoustic model of a heavily used commercial systemby distilling the knowledge in an ensemble of models into a single model.

3

Large-scale Video Classification with Convolutional Neural Networks

Author: Andrej Karpathy[1]

[1] *Google*

Corresponding Author: efg@gmail.com

Convolutional Neural Networks (CNNs) have been established as a powerful class of models for image recognition problems. Encouraged by these results, we provide an extensive empirical evaluation of CNNs on largescale video classification using a new dataset of 1 million YouTube videos belonging to 487 classes. We study multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggest a multiresolution, foveated architecture as a promising way of speeding up the training. Our best spatio-temporal networks display significant performance improvements compared to strong feature-based baselines (55.3% to 63.9%), but only a surprisingly modest improvement compared to single-frame models (59.3%to 60.9%). We further study the generalization performance of our best model by retraining the top layers on the UCF- 101 Action Recognition dataset and observe significant performance improvements compared to the UCF-101 baseline model (63.3% up from 43.9%).

4

How transferable are features in deep neural

Author: Yosinski Jason[1]

[1] *Cornell University*

Corresponding Author: mka@gmail.com

Many deep neural networks trained on natural images exhibit a curious phenomenon
in common: on the first layer they learn features similar to Gabor filters
and color blobs. Such first-layer features appear not to be specific to a particular dataset or task, but general in that they are applicable to many datasets and tasks. Features must eventually transition from general to specific by the last layer of the network, but this transition has not been studied extensively. In this paper we experimentally quantify the generality versus specificity of neurons in each layer of a deep convolutional neural network and report a few surprising results.

5

Multi-Scale Orderless Pooling of Deep Convolutional Activation Features

Author: Gong Yunchao [1]

[1] *University of North Carolina*

Corresponding Author: y.gong@gmail.com

Deep convolutional neural networks (CNN) have shown their promise as a universal representation for recognition. However, global CNN activations lack geometric invariance, which limits their robustness for classication and matching of highly variable scenes. To improve the invariance of CNN activations without degrading their discriminative power, this paper presents a simple but eective scheme called multi-scale orderless pooling (MOP-CNN). This scheme extracts CNN activations for local patches at multiple scale levels, performs orderless VLAD
pooling of these activations at each level separately, and concatenates the
result. The resulting MOP-CNN representation can be used as a generic feature for either supervised or unsupervised recognition tasks, from image classication to instance-level retrieval; it consistently outperforms global CNN activations without requiring any joint training of prediction layers for a particular target dataset. In absolute terms, it achieves state-of-the-art results on the challenging SUN397 and MIT Indoor Scenes classication datasets, and competitive results on ILSVRC2012/2013 classification and INRIA Holidays retrieval datasets.

6

Learning Sparse Gaussian Markov Networks using a Greedy Coordinate Ascent Approach

Author: Katya Scheinberg[1]

[1] *Columbia University*

Corresponding Author: katya.schein@gmail.com

In this paper, we introduce a simple but efficient greedy algorithm,
called SINCO, for the Sparse INverse COvariance selection problem, which is equivalent to learning a sparse Gaussian Markov Network, and empirically investigate the structure-recovery properties of the algorithm. Our approach is based on a coordinate ascent method which naturally preserves the sparsity of the network structure. We show that SINCO is often comparable to, and, in various cases, outperforms commonly used approaches such as glasso [7] and COVSEL [1], in terms of both structure-reconstruction error (particularly, false positive error) and computational time.

7

Machine Learning in Automated Text Categorization

Author: SEBASTIANI FABRIZIO [1]

[1] *Consiglio Nazionale delle Ricerche*

Corresponding Author: fab.sebas@gmail.com

The automated categorization (or classification) of texts into predefined categories has witnessed a booming interest in the last 10 years, due to the increased availability of documents in digital form and the ensuing need to organize them. In the research community the dominant approach to this problem is based on machine learning techniques: a general inductive process automatically builds a classifier by learning, from a set of preclassified documents, the characteristics of the categories. The advantages of this approach over the knowledge engineering approach (consisting in the manual definition of a classifier by domain experts) are a very good effectiveness, considerable savings in terms of expert labor power, and straightforward portability to different domains. This survey discusses the main approaches to text categorization that fall within the machine learning paradigm. We will discuss in detail issues pertaining to three different problems, namely, document representation, classifier construction, and classifier evaluation.

8

USING MACHINE LEARNING TO DESIGN AND INTERPRET

Authors: Molla Michael[1]; Page David [1]

[1] *University of Wisconsin*

Corresponding Authors: molla@cs.wins, david.pag@cs.edu

Gene-expression microarrays, commonly called gene chips, make it possible to
simultaneously measure the rate at which a cell or tissue is expressing
translating into a protein each of its thousands of genes. One can use these
comprehensive snapshots of biological activity to infer regulatory pathways in
cells, identify novel targets for drug design, and improve the diagnosis, prognosis, and treatment

planning for those suffering from disease. However, the amount of data this new technology produces is more than one can manually analyze.Hence, the need for automated analysis of microarray data offers an opportunity for machine learning to have a significant impact on biology and medicine. This article describes microarray technology, the data it produces, and the types of machine-learning tasks that naturally arise with this data. It also reviews some of the recent prominent applications of machine learning to gene-chip data, points to related tasks where machine learning may have a further impact on biology and medicine, and describes additional types of interesting data that recent advances in biotechnology allow biomedical researchers to collect.

9

Large-Scale Machine Learning at Twitter

Author: Lin Jimmy[1]

[1] *Twitter*

Corresponding Author: jim.lin@twitter.com

The success of data-driven solutions to dicult problems, along with the dropping costs of storing and processing massive amounts of data, has led to growing interest in large-scale machine learning. This paper presents a case study of Twitter's integration of machine learning tools into its existing Hadoop-based, Pig-centric analytics platform. We begin with an overview of this platform, which handles \traditional" data warehousing and business intelligence tasks for the organization. The core of this work lies in recent Pig extensions to provide predictive analytics capabilities that incorporate machine learning, focused specically on supervised classication. In particular, we have identied stochastic gradient descent techniques for online learning and ensemble methods as being highly amenable to scaling out to large amounts of data. In our deployed solution, common machine learning tasks such as data sampling, feature generation, training, and testing can be accomplished directly in Pig, via carefully crafted loaders, storage functions, and user-dened functions. This means that machine learning is just another Pig script, which allows seamless integration with existing infrastructure for data management, scheduling, and monitoring in a production environment, as well as access to rich libraries of user-dened functions and the materialized output of other scripts.

10

Noise reduction through Compressed Sensing

Author: J. F. Gemmeke,[1]

[1] *Radboud University*

Corresponding Author: gemmeke.j@gmail.com

We present an exemplar-based method for noise reduction using missing data imputation: A noise-corrupted word is sparsely represented in an over-complete basis of exemplar (clean) speech signals using only the uncorrupted time-frequency elements of the word. Prior to recognition the parts of the spectrogramdominated by noise are replaced by clean speech estimates obtained by projecting the sparse representation in the basis. Since at low SNRs individual frames may contain few, if any, uncorrupted coefficients, the method tries to exploit all reliable information that is available in a word-length time window. We study the effectiveness of this approach on the Interspeech 2008 Consonant Challenge (VCV) data as well as on AURORA-2 data. Using oracle masks, we obtain obtain accuracies of 36-44% on the VCV data. On AURORA-2 we obtain an accuracy of 91% at SNR -5 dB, compared to 61% using a conventionalframe-based approach, clearly illustrating the great potential of the method.

11

Online Group-Structured Dictionary Learning

Author: Szabó Zoltán[1]

[1] *Eötvös Loránd University*

We develop a dictionary learning method which is (i) online, (ii) enables overlapping group structures with (iii) non-convex sparsity-inducing regularization and (iv) handles the partially observable case. Structured sparsity and the related group norms have recently gained widespread attention in group-sparsity regularized problems in the case when the dictionary is assumed to be known and fixed. However, when the dictionary also needs to be learned, the problem is much more difficult. Only a few methods have been proposed to solve this problem, and they can handle two of these four desirable properties at most. To the best of our knowledge, our proposed method is the first one that possesses all of these properties. We investigate several interesting special cases of our framework, such as the online, structured, sparse non-negative matrix factorization, and demonstrate the efficiency of our algorithm with several umerical experiments.

12

Robust Face Recognition via Sparse Representation

Authors: John Wright[1]; Arvind Ganesh[1]

[1] *IEEE*

Corresponding Authors: ganesh.arvind@gmail.com, wright.john@gmail.com

We consider the problem of automatically recognizing human faces from frontal views with varying expression and illumination, as well as occlusion and disguise. We cast the recognition problem as one of classifying among multiple linear regression models and argue that new theory from sparse signal representation offers the key to addressing this problem. Based on a sparse representation computed by ‘1-minimization, we propose a general classification algorithm for (image-based) object recognition. This new framework provides new insights into two crucial issues in face recognition: feature extraction and robustness to occlusion. For feature extraction, we show that if sparsity in the recognition problem is properly harnessed, the choice of features is no longer critical. What is critical, however, is whether the number of features is sufficiently large and whether the sparse representation is correctly computed. Unconventional features such as downsampled images and random projections perform just as well as conventional features such as Eigenfaces and Laplacianfaces, as long as the dimension of the feature space surpasses certain threshold, predicted by the theory of sparse representation. This framework can handle errors due to occlusion and corruption uniformly by exploiting the fact that these errors are often sparse with respect to the standard (pixel) basis. The theory of sparse representation helps predict how much occlusion the recognition algorithm can handle and how to choose the training images to maximize robustness to occlusion. We conduct extensive experiments on publicly available databases to verify the efficacy of the proposed algorithm and corroborate the above claims.

13

Random Projections for Manifold Learning

Author: G. Baraniuk Richard[1]

[1] *Rice University*

Corresponding Author: richard.baraniuk@gmail.com

We propose a novel method for linear dimensionality reduction of manifold modeled data. First, we show that with a small number M of random projections of sample points in RN belonging to an unknown K-dimensional Euclidean manifold, the intrinsic dimension (ID) of the sample set can be estimated to high accuracy. Second, we rigorously prove that using only this set of random projections, we can estimate the structure of the underlying manifold. In both cases, the number of random projections required is linear in K and logarithmic in N, meaning that $K < M \ll N$. To handle practical situations, we develop a greedy algorithm to estimate the smallest size of the projection space required to perform manifold learning. Our method is particularly relevant in distributed sensing systems and leads to significant potential savings in data acquisition, storage and transmission costs.

14

ImageNet Classification with Deep Convolutional

Author: Alex Krizhevsky[1]

[1] *University of Toronto*

Corresponding Author: alex.krxi@cs.toronto.edu

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists f five convolutional layers, some of which are followed by max-pooling layers, nd three fully-connected layers with a final 1000-way softmax. To make training aster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout"that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

15

Visualizing and Understanding Convolutional Networks

Author: D. Zeiler Matthew[1]

[1] *New York University*

Corresponding Author: matt.d@eu.ny

Large Convolutional Network models have recently demonstrated impressive classication performance on the ImageNet Benchmark (Krizhevsky et al., 2012). However there is no clear understanding of why they perform so well, or how they might be improved. In this paper we address both issues. We introduce a novel visualization technique that gives insight into the function of intermediate feature layers and the operation of the classier. Used in a diagnostic role, these visualizations allow us to nd model architectures that outperform Krizhevsky et al. on the ImageNet classication benchmark. We also perform an ablation study to discover the performance contribution from different model layers. We show our ImageNet model generalizes well to other datasets: when the softmax classier is retrained, it convincingly beats the current state-of-the-art results on Caltech-101 and Caltech-256 datasets.

16

VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION

Author: Simonyan Karen [1]

[1] *University of Oxford*

Corresponding Author: karen.simonyan@gmail.com

In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3×3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16–19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve state-of-the-art results. We have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.

17

Going Deeper with Convolutions

Author: Szegedy Christian [1]

[1] *Google Inc.*

Corresponding Author: christian.szaegy@google.com

We propose a deep convolutional neural network architecture codenamed Inception that achieves the new state of the art for classification and detection in the ImageNet Large-Scale Visual Recognition Challenge 2014 (ILSVRC14). The main hallmark of this architecture is the improved utilization of the computing resources inside the network. By a carefully crafted design, we increased the depth and width of the network while keeping the computational budget constant. To optimize quality, the architectural decisions were based on the Hebbian principle and the intuition of multi-scale processing. One particular incarnation used in our submission for ILSVRC14 is called GoogLeNet, a 22 layers deep network, the quality of which is assessed in the context of classification and detection.

18

Deep Residual Learning for Image Recognition

Author: He Kaiming [1]

[1] *Microsoft*

Corresponding Author: he.kaiming@microsoft.com

Deeper neural networks are more difficult to train. We present a residual learning framework to ease the training of networks that are substantially deeper than those used previously. We explicitly reformulate the layers as learning residual functions with reference to the layer inputs, instead of learning unreferenced functions. We provide comprehensive empirical evidence showing that these residual networks are easier to optimize, and can gain accuracy from considerably increased depth. On the ImageNet dataset we evaluate residual nets with a depth of up to 152 layers—8 deeper than VGG nets [41] but still having lower complexity. An ensemble of these residual nets achieves 3.57%

error on the ImageNet test set. This result won the 1st place on the ILSVRC 2015 classification task. We also present analysis on CIFAR-10 with 100 and 1000 layers. The depth of representations is of central importance for many visual recognition tasks. Solely due to our extremely deep representations, we obtain a 28% relative improvement on the COCO object detection dataset. Deep residual nets are foundations of our submissions to ILSVRC & COCO 2015 competitions1, where we also won the 1st places on the tasks of ImageNet detection, ImageNet localization, COCO detection, and COCO segmentation.

19

Rich feature hierarchies for accurate object detection and semantic segmentation

Author: Girshick Ross[1]

[1] *UC Berley*

Corresponding Author: ross.girshick@eecs.berlay.edu

Object detection performance, as measured on the canonical PASCAL VOC dataset, has plateaued in the last few years. The best-performing methods are complex ensemble systems that typically combine multiple low-level image features with high-level context. In this paper, we propose a simple and scalable detection algorithm that improves mean average precision (mAP) by more than 30% relative to the previous best result on VOC 2012—achieving a mAP of 53.3%. Our approach combines two key insights: (1) one can apply high-capacity convolutional neural Networks (CNNs) to bottom-up region proposals in order to localize and segment objects and (2) when labeled training data is scarce, supervised pre-training for an auxiliary task, followed by domain-specific fine-tuning, yields a significantperformance boost. Since we combine region proposals with CNNs, we call our method R-CNN: Regions with CNN features. We also compare R-CNN to OverFeat, a recently proposed sliding-window detector based on a similar CNN architecture.

20

Generative Adversarial Nets

Author: Ian Goodfellow[1]

[1] *Universite de Montreal*

Corresponding Author: ian.goodfellow@gmail.com

We propose a new framework for estimating generative models via an adversarial process, in which we simultaneously train two models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training data rather than G. The training procedure for G is to maximize the probability of D making a mistake. This framework corresponds to a minimax two-player game. In the space of arbitrary functions G and D, a unique solution exists, with G recovering the training data distribution and D equal to 1 2 everywhere. In the case where G and D are defined by multilayer perceptrons, the entire system can be trained with backpropagation. There is no need for any Markov chains or unrolled approximate inference networks during either training or generation of samples. Experiments demonstrate the potential of the framework through qualitative and quantitative evaluation of the generated samples.

21

Spatial Transformer Networks

Author: Jaderberg Max[1]

[1] *Google DeepMind*

Corresponding Author: max@google.com

Convolutional Neural Networks define an exceptionally powerful class of models, but are still limited by the lack of ability to be spatially invariant to the input data in a computationally and parameter efficient manner. In this work we introduce a new learnable module, the Spatial Transformer, which explicitly allows the spatial manipulation of data within the network. This differentiable module can be inserted into existing convolutional architectures, giving neural networks the ability to actively spatially transform feature maps, conditional on the feature map itself, without any extra training supervision or modification to the optimisation process. We show that the use of spatial transformers results in models which learn invariance to translation, scale, rotation and more generic warping, resulting in state-of-the-art performance on several benchmarks, and for a number of classes of transformations

22

Deep Convolutional Neural Network for Image Deconvolution

Author: Xu Li [1]

[1] *Lenovo Research Technology*

Corresponding Author: xul@lenovo.com

Many fundamental image-related problems involve deconvolution operators. Real blur degradation seldom complies with an ideal linear convolution model due to camera noise, saturation, image compression, to name a few. Instead of perfectly modeling outliers, which is rather challenging from a generative model perspective, we develop a deep convolutional neural network to capture the characteristics of degradation. We note directly applying existing deep neural networks does not produce reasonable results. Our solution is to establish the connection between traditional optimization-based schemes and a neural network architecture where a novel, separable structure is introduced as a reliable support for robust deconvolution against artifacts. Our network contains two submodules, both trained in a supervised manner with proper initialization. They yield decent performance on non-blind image deconvolution compared to previous generative-model based methods.

23

Fast R-CNN

Author: Girshick Ross [1]

[1] *Microsoft Research Center*

Corresponding Author: rgb@microsoft.com

This paper proposes a Fast Region-based Convolutional Network method (Fast R-CNN) for object detection. Fast R-CNN builds on previous work to efficiently classify object proposals using deep convolutional networks. Compared to previous work, Fast R-CNN employs several innovations to improve training and testing speed while also increasing detection accuracy. Fast R-CNN trains the very deep VGG16 network 9× faster than R-CNN, is 213× faster at test-time, and achieves a higher mAP on PASCAL VOC 2012. Compared to SPPnet, Fast R-CNN trains VGG16 3× faster, tests 10× faster, and is more accurate.

24

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

Author: He Kaiming [1]

[1] *Microsoft*

Corresponding Author: he.kaimin@microsoft.com

State-of-the-art object detection networks depend on region proposal algorithms to hypothesize object locations. Advances like SPPnet [7] and Fast R-CNN [5] have reduced the running time of these detection networks, exposing region proposal computation as a bottleneck. In this work, we introduce a Region Proposal Network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposals. An RPN is a fully-convolutional network that simultaneously predicts object bounds and objectness scores at each position.

25

DLPaper2Code: Auto-generation of Code from Deep Learning Research Papers

Author: Sankaran Anush [1]

[1] *IBM Researcher*

Corresponding Author: anush.danuish@gmail.com

With an abundance of research papers in deep learning, reproducibility or adoption of the existing works becomes a challenge. This is due to the lack of open source implementations provided by the authors. Further, re-implementing research papers in a different library is a daunting task. To address these challenges, we propose a novel extensible approach, DLPaper2Code, to extract and understand deep learning design flow diagrams and tables available in a research paper and convert them to an abstract computational graph. The extracted computational graph is then converted into execution ready source code in both Keras and Caffe, in realtime. An arXiv-like website is created where the automatically generated designs is made publicly available for 5; 000 research papers.

26

Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Author: Saenko Kate[1]

[1] *Texas University*

Corresponding Author: saenko@gmail.com

Models based on deep convolutional networks have dominated recent image interpretation tasks; we investigate whether models which are also recurrent, or "temporally deep", are effective for tasks involving sequences, visual and otherwise. We develop a novel recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable, and demonstrate the value of these models on benchmark video recognition tasks, image description and retrieval problems, and

video narration challenges. In contrast to current models which assume a fixed spatio-temporal receptive field or simple temporal averaging for sequential processing, recurrent convolutional models are “doubly deep”in that they can be compositional in spatial and temporal “layers”.

27

Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling

Author: Sak Hasim [1]

[1] *Google*

Corresponding Author: hasim.sak@google.com

Long Short-Term Memory (LSTM) is a specific recurrent neural network (RNN) architecture that was designed to model temporal sequences and their long-range dependencies more accurately than conventional RNNs. In this paper, we explore LSTM RNN architectures for large scale acoustic modeling in speech recognition. We recently showed that LSTM RNNs are more effective than DNNs and conventional RNNs for acoustic modeling, considering moderately-sized models trained on a single machine. Here, we introduce the first distributed training of LSTM RNNs using asynchronous stochastic gradient descent optimization on a large cluster of machines. We show that a two-layer deep LSTM RNN where each LSTM layer has a linear recurrent projection layer can exceed state-of-the-art speech recognition performance.

28

Deep Sentence Embedding Using Long Short-Term Memory Networks

Author: Palangi Hamid[None]

Corresponding Author: hamid.palangi@yahoo.com

This paper develops a model that addresses sentence embedding, a hot topic in current natural language processing research, using recurrent neural Networks (RNN) with Long Short-Term Memory (LSTM) cells. The proposed LSTM-RNN model sequentially takes each word in a sentence, extracts its information, and embeds it into a semantic vector. Due to its ability to capture long term memory, the LSTM-RNN accumulates increasingly richer information as it goes through the sentence, and when it reaches the last word, the hidden layer of the network provides a semantic representation of the whole sentence.

29

DATA CLASSIFICATION USING SUPPORT VECTOR

Author: K. SRIVASTAVA DURGESH [None]

Corresponding Author: durgesh.g@gmail.com

Classification is one of the most important tasks for different application such as text categorization, tone recognition, image classification, micro-array gene expression, proteins structure predictions, data Classification etc. Most of the existing supervised classification methods are based on traditional statistics, which can provide ideal results when sample size is tending to infinity. However, only finite samples can be acquired in practice. In this paper, a novel learning method, Support

Vector Machine (SVM), is applied on different data (Diabetes data, Heart Data, Satellite Data and Shuttle data) which have two or multi class. SVM, a powerful machine method developed from statistical learning and has made significant achievement in some field. Introduced in the early 90' s, they led to an explosion of interest in machine learning. The foundations of SVM have been developed by Vapnik and are gaining popularity in field of machine learning due to many attractive features and promising empirical performance. SVM method does not suffer the limitations of data dimensionality and limited samples [1] & [2].

30

A Practical Guide to Applying Echo State Networks

Author: Lukosevicius Mantas [1]

[1] *Jacobs University Bremen*

Corresponding Author: mangtas@bremen.edu.de

Reservoir computing has emerged in the last decade as an alternative to gradient descent methods for training recurrent neural networks. Echo State Network (ESN) is one of the key reservoir computing \ avors". While being practical, conceptually simple, and easy to implement, ESNs require some experience and insight to achieve the hailed good performance in many tasks. Here we present practical techniques and recommendations for successfully applying ESNs, as well as some more advanced application-specic modications.

31

Musical Instrument Mapping Design with Echo State Networks

Author: Kiefer Chris [1]

[1] *University of London*

Corresponding Author: chris.ki@gold.ac.uk

Echo State Networks (ESNs), a form of recurrent neural network developed in the eld of Reservoir Computing, significant potential for use as a tool in the design of map- pings for digital musical instruments. They have, however, seldom been used in this area, so this paper explores their possible applications. This project contributes a new open source library, which was developed to allow ESNs to run in the Pure Data data ow environment. Several use cases were explored, focusing on addressing current issues in mapping research. ESNs were found to work successfully in scenarios of pattern classication, multiparametric control, explo- rative mapping and the design of nonlinearities and uncontrol. Un-trained behaviours are proposed, as augmentations to the conventional reservoir system that allow the player to introduce potentially interesting non-linearities and un- control into the reservoir.

32

Predictive Modeling with Echo State Networks

Author: Tino Peter [1]

[1] *University of Birmingham*

Corresponding Author: peter.tino@uc.bmg.edu.uk

A lot of attention is now being focused on connectionist models known under the name "reservoir computing". The most prominent example of these approaches is a recurrent neural network architecture called an echo state Network (ESN). ESNs were successfully applied in several time series modeling tasks and according to the authors they performed exceptionally well. Multiple enhancements to standard ESN were proposed in the literature. In this paper we follow the opposite direction by suggesting several simplifications to the original ESN architecture. ESN reservoir features contractive dynamics resulting from its'initialization with small weights. Sometimes it serves just as a simple memory of inputs and provides only negligible "extra-value"over much simple methods. We experimentally support this claim and we show that many tasks modeled by ESNs can be handled with much simple approaches.

33

Adaptive Nonlinear System Identification with Echo State Networks

Author: Jaeger Herbert [1]

[1] *International University Bremen*

Corresponding Author: herbert.jae@uc.bremen.de

Echo state networks (ESN) are a novel approach to recurrent neural network training. An ESN consists of a large, fixed, recurrent "reservoir" network, from which the desired output is obtained by training suitable output connection weights. Determination of optimal output weights becomes a linear, uniquely solvable task of MSE minimization. This article reviews the basic ideas and describes an online adaptation scheme based on the RLS algorithm known from adaptive linear systems. As an example, a 10-th order NARMA system is adaptively identified. The known benefits of the RLS algorithms carryover from linear systems to nonlinear ones; specifically, the convergence rate and misadjustment can be determined at design time.

34

Minimum Complexity Echo State Network

Author: Rodan Ali [None]

Corresponding Author: ali.rodan@gmail.com

Reservoir computing (RC) refers to a new class of state-space models with a fixed state transition structure (the "reservoir") and an adaptable readout form the state space. The reservoir is supposed to be sufficiently complex so as to capture a large number of features of the input stream that can be exploited by the reservoir-to-output readout mapping. The field of RC has been growing rapidly with many successful applications. However, RC has been criticized for not being principled enough. Reservoir construction is largely driven by a series of randomized model building stages, with both researchers and practitioners having to rely on a series of trials and errors.

35

Memory Capacity of Input-Driven Echo State Networks at the Edge of Chaos

Author: Barancok Peter [1]

[1] *University in Bratislava*

Corresponding Author: peter.barncok@outlook.com

Reservoir computing provides a promising approach to efficient training of recurrent neural networks, by exploiting the computational properties of the reservoir structure. Various approaches, ranging from suitable initialization to reservoir optimization by training have been proposed. In this paper we take a closer look at short-term memory capacity, introduced by Jaeger in case of echo state networks. Memory capacity has recently been investigated with respect to criticality, the so called edge of chaos, when the network switches from a stable regime to an unstable dynamic regime.

36

Imagining, Playing, and Coding with KIBO Using Robotics to Foster Computational Thinking in Young Children

Author: A. SULLIVAN Amanda [1]

[1] *DevTech Research Group*

Corresponding Author: amanda.e@tufts.edu

The KIBO robotics kit offers a playful and tangible way for young children to learn computational thinking skills by building and programming a robot. KIBO is specifically designed for children ages 4-7 years old and was developed by the DevTech research group at Tufts University through nearly a decade of research funded by the National Science Foundation. KIBO allows young children to become engineers by constructing robots using motors, sensors, and craft materials. Children also become programmers by exploring sequences, loops, and variables.

37

Breaking the Curse of Dimensionality with Convex Neural Networks

Author: Bach Francis [None]

Corresponding Author: francis.b@ens.fr

We consider neural networks with a single hidden layer and non-decreasing positively homogeneous activation functions like the rectified linear units. By letting the number of hidden units grow unbounded and using classical non-Euclidean regularization tools on the output weights, they lead to a convex optimization problem and we provide a detailed theoretical analysis of their generalization performance, with a study of both the approximation and the estimation errors. We show in particular that they are adaptive to unknown underlying linear structures, such as the dependence on the projection of the input variables onto a low-dimensional subspace.

38

Beyond Word Importance: Contextual Decomposition to Extract Interactions from LSTMs

Author: W. James Murdoch[None]

Corresponding Author: murodch.jamws@outlook.com

The driving force behind the recent success of LSTMs has been their ability to learn complex and non-linear relationships. Consequently, our inability to describe these relationships has led to LSTMs being characterized as black boxes. To this end, we introduce contextual decomposition (CD), an interpretation algorithm for analysing individual predictions made by standard LSTMs, without any changes to the underlying model. By decomposing the output of a LSTM, CD captures the contributions of combinations of words or variables to the final prediction of an LSTM. On the task of sentiment analysis with the Yelp and SST data sets, we show that CD is able to reliably identify words and phrases of contrasting sentiment, and how they are combined to yield the LSTM's final prediction. Using the phrase-level labels in SST, we also demonstrate that CD is able to successfully extract positive and negative negations from an LSTM, something which has not previously been done.

39

Generating Wikipedia by Summarizing Long Sequences

Author: Ahmad Saleh Mohammad[None]

Corresponding Author: mh.salah@gmail.com

We show that generating English Wikipedia articles can be approached as a multi- document summarization of source documents. We use extractive summarization to coarsely identify salient information and a neural abstractive model to generate the article. For the abstractive model, we introduce a decoder-only architecture that can scalably attend to very long sequences, much longer than typical encoder- decoder architectures used in sequence transduction. We show that this model can generate fluent, coherent multi-sentence paragraphs and even whole Wikipedia articles. When given reference documents, we show it can extract relevant factual information as reflected in perplexity, ROUGE scores and human evaluations.

40

RNN Approaches to Text Normalization: A Challenge

Author: Sproat Richard [1]

[1] *Google*

Corresponding Author: rcih.sproay@gmail.com

This paper presents a challenge to the community: given a large corpus of written text aligned to its normalized spoken form, train an RNN to learn the correct normalization function. We present a data set of general text where the normalizations were generated using an existing text normalization component of a text-to-speech system. This data set will be released open-source in the near future. We also present our own experiments with this data set with a variety of different RNN architectures. While some of the architectures do in fact produce very good results when measured in terms of overall accuracy, the errors that are produced are problematic, since they wouldconvey completely the wrong message if such a system were deployed in a speech application. On the other hand, we show that a simple FST-based filter can mitigate those errors, and achieve a level of accuracy not achievable by the RNN alone.

41

Analyzing Language Learned by an Active Question Answering Agent

Author: Buck Christian [1]

[1] *Google*

Corresponding Author: buck.sa@google.com

We analyze the language learned by an agent trained with reinforcement learning as a component of the ActiveQA system [Buck et al., 2017]. In ActiveQA, question answering is framed as a reinforcement learning task in which an agent sits between the user and a black box question-answering system. The agent learns to reformulate the user's questions to elicit the optimal answers. It probes the system with many versions of a question that are generated via a sequence-to-sequence question reformulation model, then aggregates the returned evidence to find the best answer.

42

Approaches for Neural-Network Language Model Adaptation

Author: Biadsy Fadi [1]

[1] *Google Inc.*

Corresponding Author: fadi.biad@google.com

Language Models (LMs) for Automatic Speech Recognition (ASR) are typically trained on large text corpora from news articles, books and web documents. These types of corpora, however, are unlikely to match the test distribution of ASR systems, which expect spoken utterances. Therefore, the LM is typically adapted to a smaller held-out in-domain dataset that is drawn from the test distribution. We present three LM adaptation approaches for Deep NN and Long Short-Term Memory (LSTM): (1) Adapting the softmax layer in the NN; (2) Adding a non-linear adaptation layer before the softmax layer that is trained only in the adaptation phase; (3) Training the extra non-linear adaptation layer in pre-training and adaptation phases. Aiming to improve upon a hierarchical Maximum Entropy (MaxEnt) second-pass LM baseline, which factors the model into word-cluster and word models, we build an NN LM that predicts only word clusters. Adapting the LSTM LM by training the adaptation layer in both training and adaptation phases (Approach 3), we reduce the cluster perplexity by 30% compared to an unadapted LSTM model. Initial experiments using a state-of-the-art ASR system show a 2.3% relative reduction in WER on top of an adapted MaxEnt LM.

43

Avoiding Your Teacher's Mistakes: Training Neural Networks with Controlled Weak Supervision

Author: Severyn Aliaksei [1]

[1] *Google Research*

Corresponding Author: alia.seve@google.com

Making use of weak or noisy signals, like the output of heuristic methods or user click through data for training deep neural networks is increasing, in particular for the tasks where an adequate amount of data with true labels is not available. In a semi-supervised setting, we can use a large set of data with weak labels to pretrain a neural network and fine tune the parameters with a small amount of data with true labels. However, these two independent stages do not leverage the full capacity of clean information from true labels during pretraining. In this paper, we propose a semi-supervised learning method where we train two neural networks in a multi-task fashion: a target network and a confidence network. The target network is optimized to perform a given task and is trained using a large set of unlabeled data that are weakly annotated. We propose to weight the gradient

updates to the target network using the scores provided by the second confidence network, which is trained on a small amount of supervised data. Thus we avoid that the weight updates computed from noisy labels harm the quality of the target network model. We evaluate our learning strategy on two different tasks: document ranking and sentiment classification. The results demonstrate that our approach not only enhances the performance compared to the baselines but also speeds up the learning process from weak labels.

44

Better Text Understanding Through Image-To-Text Transfer

Author: Kurach Karol[1]

[1] *Google Brian*

Corresponding Author: karol@google.com

Generic text embeddings are successfully used in a variety of tasks. However, they are often learnt by capturing the co-occurrence structure from pure text corpora, resulting in limitations of their ability to generalize. In this paper, we explore models that incorporate visual information into the text representation. Based on comprehensive ablation studies, we propose a conceptually simple, yet well performing architecture. It outperforms previous multimodal approaches on a set of well established benchmarks. We also improve the state-of-the-art results for image-related text datasets, using orders of magnitude less data.

45

Depthwise Separable Convolutions for Neural Machine Translation

Author: Kaiser Łukasz [1]

[1] *Google Brain*

Corresponding Author: kaiser.luckas@google.com

Depthwise separable convolutions reduce the number of parameters and computation used in convolutional operations while increasing representational efficiency. They have been shown to be successful in image classification models, both in obtaining better models than previously possible for a given parameter count (the Xception architecture) and considerably reducing the number of parameters required to perform at a given level (the MobileNets family of architectures). Recently, convolutional sequence-to-sequence networks have been applied to machine translation tasks with good results. In this work, we study how depthwise separable convolutions can be applied to neural machine translation.

46

Efficient Natural Language Response Suggestion

Author: HENDERSON MATTHEW[1]

[1] *Google*

Corresponding Author: matt.hew@google.com

This paper presents a computationally efficient machine-learned method for natural language response suggestion. Feed-forward neural networks using n-gram embedding features encode messages into vectors which are optimized to give message-response pairs a high dot-product value. An optimized search finds response suggestions. The method is evaluated in a large-scale commercial e-mail application, Inbox by Gmail Compared to a sequence-to-sequence approach, the new system achieves the same quality at a small fraction of the computational requirements and latency.

47

Learning to Skim Text

Author: Wei Yu Adams[1]

[1] *Carnegie Mellon University*

Corresponding Author: weiyu_me@cs.cmu.edu

Recurrent Neural Networks are showing much promise in many sub-areas of natural language processing, ranging from document classification to machine translation to automatic question answering. Despite their promise, many recurrent models have to read the whole text word by word, making it slow to handle long documents. For example, it is difficult to use a recurrent network to read a book and answer questions about it. In this paper, we present an approach of reading text while skipping irrelevant information if needed. The underlying model is a recurrent network that learns how far to jump after reading a few words of the input text. We employ a standard policy gradient method to train the model to make discrete jumping decisions. In our benchmarks on four different tasks, including number prediction, sentiment analysis, news article classification and automatic Q\&A, our proposed model, a modified LSTM with jumping, is up to 6 times faster than the standard sequential LSTM, while maintaining the same or even better accuracy.

48

Natural Language Processing with Small Feed-Forward Networks

Authors: A. Botha Jan [1]; Bakalov Anton [2]

[1] *Google Inc*

[2] *Google Inc.*

Corresponding Authors: botha.salv@google.com, a_bakalov@google.com

We show that small and shallow feedforward neural networks can achieve near state-of-the-art results on a range of unstructured and structured language processing tasks while being considerably cheaper in memory and computational requirements than deep recurrent models. Motivated by resource-constrained environments like mobile phones, we showcase simple techniques for obtaining such small neural network models, and investigate different tradeoffs when deciding how to allocate a small memory budget.

49

Neural Symbolic Machines: Learning Semantic Parsers on Freebase with Weak Supervision

Author: Lao Ni [1]

[1] *Google Inc.*

Corresponding Author: lao_n@google.com

Modern semantic parsers, which map natural language utterances to executable logical forms, have been successfully trained over large knowledge bases from weak supervision, but require handcrafted rules and substantial feature engineering. Recent attempts to train an end-to-end neural network for semantic parsing have either used strong supervision (full logical forms), or have employed synthetic datasets and differentiable operations. In this work, we propose the Boss-Programmer-Computer framework to integrate neural network models with symbolic operations. Within this framework, we introduce Neural Symbolic Machines, in which a sequence-to-sequence neural network "programmer" controls a non-differentiable "computer" that executes Lisp programs (equivalent to logical forms) and provides code assistance. The interaction between the "programmer" and "computer" dramatically reduces the search space and effectively learns the semantic parser from weak supervision over a large knowledge base, such as Freebase. Our model obtained new state-of-the-art performance on \textsc{WebQuestionsSP}, a challenging semantic parsing dataset.

50

Towards better decoding and language model integration in sequence to sequence models

Author: Jaitly Navdeep [1]

[1] *Google Brain*

Corresponding Author: jai.nav@google.com

The recently proposed Sequence-to-Sequence (seq2seq) framework advocates replacing complex data processing pipelines, such as an entire automatic speech recognition system, with a single neural network trained in an end-to-end fashion. In this contribution, we analyse an attention-based seq2seq speech recognition system that directly transcribes recordings into characters. We observe two shortcomings: overconfidence in its predictions and a tendency to produce incomplete transcriptions when language models are used. We propose practical solutions to both problems achieving competitive speaker independent word error rates on the Wall Street Journal dataset: without separate language models we reach 10.6% WER, while together with a trigram language model, we reach 6.7% WER.