Activities report from March 2014 to March 2015

Chetak Kandaswamy

11th March 2015

Project "Reusable Deep Neural Networks: Applications to Biomedical Data"
(PDTC/EIA-EIA/119004/2010)

Instituto de Engenharia Biomédica (INEB)
Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Contents

1 Introduction **3**
1.1 Contributions published and under review 3

2 Biomedical Journal overview: High-Content Screening **4**
2.1 Conclusion 4

3 Published International Conferences: **5**
3.1 Paper overview: IEEE SMC 5
3.2 Paper overview: ICANN 6
3.3 Paper overview: ICMLA 6
3.4 Paper overview: ICIAR 6
3.5 Paper overview: RecPAD 7
3.6 Paper overview: IbPRIA 7

4 Works under review at conference: **8**

5 Creation and maintenance of the group's webpage **8**

A Slides: ICANN conference presentation **10**

B Slides: SMC conference presentation **15**

1 Introduction

This report presents an overview of the work done of the research assistant grant from March 15th, 2014 to March 11th, 2014 within the project “Reusable Deep Neural Networks: applications to biomedical data” with reference PDTC/EIA-EIA/119004/2010. In the following sections I describe the work done, including an overview of the international Journal and Conferences published and awaiting for review. Inclusive of this document, I also attach the published, accepted, under review papers. Also the biweekly group meeting discussions captured as presentation slides. Finally, the maintains of the developed code the groups webpage are described.

1.1 Contributions published and under review.

1. Kandaswamy, Chetak, Marques de Sá, Luís M. Silva, Luís A. Alexandre & Jorge M. Santos. "Improving transfer learning accuracy by reusing Stacked Denoising Autoencoders." Systems, Man and Cybernetics (SMC), IEEE International Conference on. IEEE, 2014.
2. Kandaswamy, Chetak, Luís M. Silva, Luís A. Alexandre, Jorge M. Santos et al, Marques de Sá. "Improving deep neural network performance by reusing features trained with transductive transference." Artificial Neural Networks and Machine Learning (ICANN). Springer International Publishing, 265-272, 2014.
3. Amaral, Telmo, Kandaswamy, Chetak, Luís M. Silva, Luís A. Alexandre, Joaquim Marques de Sá, Jorge M. Santos. "Improving performance on problems with few labelled data by reusing stacked auto-encoders.", International Conference on Image Analysis and Recognition (ICIAR) Vilamoura, Algarve, Portugal 2014.
4. Amaral, Telmo, Luís M. Silva, Luís A. Alexandre, Kandaswamy, Chetak., Marques de Sá, & Jorge M. Santos. “Transfer learning using rotated image data to improve deep neural network performance. In Image Analysis and Recognition” Springer International Publishing, pp. 290-300, 2014.
5. Kandaswamy, Chetak, Luís M. Silva, Jaime S Cardoso. "Improving Classification Accuracy of Deep Neural Networks by Transferring Features from a Different Distribution", 20th edition of the Portuguese Conference on Pattern Recognition, University of Beira Interior, Covilhã, 2014.
6. Amaral, Telmo, Luís M. Silva, Luís A. Alexandre, Kandaswamy, Chetak., Marques de Sá, & Jorge M. Santos. "Using different cost functions to train stacked auto-encoders." Artificial Intelligence (MICAI), 12th Mexican International Conference on. IEEE, 2013.
7. Kandaswamy, Chetak, Luís M. Silva, Jaime S Cardoso. "Source-Target-Source Classification using Stacked Denoising Autoencoders”, 7th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA) at Santiago de Compostela, Spain, 2014 (Accepted)

8. Kandaswamy, Chetak, Luís M. Silva, Luís A. Alexandre, & Jorge M. Santos. “High-Content Screening of MCF7 Breast Cancer Cells Using Deep Transfer Learning" Journal of Biomolecular screening. (under review)
9. Kandaswamy, Chetak, Luís M. Silva, Luís A. Alexandre, & Jorge M. Santos. “Ensemble of Deep Transfer Learning Classification”. Special session on Transfer Learning at IWANN, Spain. (under review)

2 Biomedical Journal overview: High-Content Screening

The current work is under review and planned to submit to Journal of Biomolecular screening. I am attaching the final version of the "*High-Content Screening of MCF7 Breast Cancer Cells Using Deep Transfer Learning*" for SAGE journal.

High-content screening is increasingly used in several drug-discovery applications with small molecules and biological probes. We describe a cell classifier for automated analysis of multiparametric data from immunofluorescence microscopy and prediction of a compound’s mechanism of action (MOA) on MCF7 breast cancer cell line. We build the classifier using all the features obtained using open-source implementations of cell profiling. We obtained substantial improvements over published results with the same data while correctly predicting MOA for 98% of the treatments.

2.1 Conclusion

To stimulate the development of new drugs effective against a wide spectrum of cancers, we propose DTL classifying method that uses high-content as high as possible from super- resolution phenotypic images without overfitting the model, which is often reduced for the sake of algorithmic limitation and computation complexity. We compare our approach with the state- of-the-art machine learning algorithms that are commonly used for large volumes of high-content images. We categorize the machine learning algorithms into three main types of architectures and compare the accuracy for the task of classifying moa as listed in Table 5. First, machine learning algorithms based on shallow architecture like, random forest, SVM and principal component analysis. Second, algorithms based on deep architecture like SDA, and finally based on DTL framework.

The classification accuracy of shallow architecture like, random forest which uses ensemble of decision trees and SVM which non-linearly maps input vectors to high-dimensional decision surface relatively use lower level representation than deep architecture. To compare the result of dimensionality reduction approach using principal component analysis 4 is not straight forward, as not all the data used in the experiments are available for academic usage. The classification accuracy of factor analysis+means 6 approach for MFC7 dataset is 94% by using cell profiling feature extraction approach with leave one out cross validation. However, the results used in factor analysis+means 6 approach cannot be compared directly because the difference in model selection technique. We use 2-fold cross validation which is more robust and difficult to generalize than the leave-one-out approach, and due to the possibility of having a completely separate test set for evaluating the selected model. This way, we do not use the same sample while training and testing, thus the resulted model generalizes well for real world data. Also, the factor analysis needs to heuristically determine the number of factors required for particular problem. Deep learning approach uses hierarchical representation which enables the model to discriminates the data well. We use baseline SDA approach to test the deep representation approach which significantly improves over the shallow architecture algorithms. We studied the DTL performance of supervised layer based feature transference (SSDA) approach for retraining different layers between source and target problems. The results showed significant improvement in the average accuracy and computation time from the baseline for transfer problems. We achieved the best classification accuracy of 98.4% for TL [0011] approach.

3 Published International Conferences:

3.1 Paper overview: IEEE SMC

This work is published in IEEE Systems, Man and Cybernetic conference at California, USA. I am attaching the published version of the "*Improving Transfer Learning Accuracy by Reusing Stacked Denoising Autoencoders*" for IEEE proceedings.

Transfer learning is a process that allows reusing a learning machine trained on a problem to solve a new problem. Transfer learning studies on shallow architectures show low performance as they are generally based on hand-crafted features obtained from experts. It is therefore interesting to study transference on deep architectures, known to directly extract the features from the input data. A Stacked Denoising Autoencoder (SDA) is a deep model able to represent the hierarchical features needed for solving classification problems. In this paper we study the performance of SDAs trained on one problem and reused to solve a different problem not only with different distribution but also with a different tasks. We propose two different approaches: 1) unsupervised feature transference, and 2) supervised feature transference using deep transfer learning. We show that SDAs using the unsupervised feature transference outperform randomly initialized machines on a new problem. We achieved 7% relative improvement on average error rate and 41% on average computation time to classify typed uppercase letters. In the case of supervised feature transference, we achieved 5.7% relative improvement in the average error rate, by reusing the first and second hidden layer, and 8.5% relative improvement for the average error rate and 54% speed up w.r.t the baseline by reusing all three hidden layers for the same data. We also explore transfer learning between geometrical shapes and canonical shapes, we achieved 7.4% relative improvement on average error rate in case of supervised feature transference approach.

3.2 Paper overview: ICANN

This work is published in International conference on artificial neural network at Hamburg, Germany. I am attaching the published version of the "*Improving Deep Neural Network Performance by Reusing Features Trained with Transductive Transference*" for ICANN proceedings.

We proposed a layer based feature transference approach that supports standard neural networks like CNN and SDA for solving transductive transfer learning problems. By transferring either low or high layer features on machines trained either unsupervised or supervised way. Using this approach we achieved performance improvement with significant reduction in computation time and also de- creased classification error rate. We achieved significant performance by transfer- ring learning from source to target problem, by using lower-layer features trained in supervised fashion in case of CNN's and unsupervised features trained in case of SDA's.

3.3 Paper overview: ICMLA

This work is published in 13th International Conference on Machine Learning and Applications (ICMLA'14) held in Detroit, MI USA on December 3-5, 2014. I am attaching the published version of the "*Improving Performance on Problems with Few Labelled Data by Reusing Stacked Auto-Encoders*" for ICMLA proceedings.

Deep architectures have been used in transfer learning applications, with the aim of improving the performance of networks designed for a given problem by reusing knowledge from another problem. In this work we addressed the transfer of knowledge between deep networks used as classifiers of digit and shape images, considering cases where only the set of class labels, or only the data distribution, changed from source to target problem. Our main goal was to study how the performance of knowledge transfer between such problems would be affected by varying the number of layers being retrained and the amount of data used in that retraining. Generally, reusing networks trained for a different label set led to better results than reusing networks trained for a different data distribution. In particular, reusing for less classes a network trained for more classes was beneficial for virtually any amount of training data. In all cases, retraining only one layer to save time consistently led to poorer performance. The results obtained when retraining for upright digits a network trained for rotated digits raise the hypothesis that transfer learning could be used to better deal with image classification problems in which only a small amount of labelled data is available for training.

3.4 Paper overview: ICIAR

This work is published in International Conference on Image Analysis and Recognition (ICIAR) 2014 held in Vilamoura, Algarve, Portugal. I am attaching the published version of the "*Transfer Learning Using Rotated Image Data to Improve Deep Neural Network Performance*" for ICIAR proceedings.

In this work we explore the idea that, in the presence of a small training set of images, it could be beneficial to use that set itself to obtain a transformed training set (by performing a random rotation on each sample), train a source network using the transformed data, then retrain the source network using the original data. Applying this transfer learning technique to three different types of character data, we achieve average relative improvements between 6% and 16% in the clas- sification test error. Furthermore, we show that it is possible to achieve relative improvements between 8% and 42% in cases where the amount of original training samples is very limited (30 samples per class), by in- troducing not just one rotation but several random rotations per sample.

3.5 Paper overview: RecPAD

This work is published and held at 20th edition of the Portuguese Conference on Pattern Recognition, organized by the APRP (Associação Portuguesa de Reconhecimento de Padrões – Portuguese Association for Pattern Recognition), University of Beira Interior, Covilhã, 2014. I am attaching the published version of the "*Improving Classification Accuracy of Deep Neural Networks by Transferring Features from a Different Distribution*".

Deep Neural Networks (DNN) are feed-forward, artificial neural networks that allow learning of multiple levels of abstraction that help to make sense of data such as images, sound, and text. We study the performance of DNNs using transfer learning approaches. Transfer learning is a process where a network trained on a source problem is reused to solve a new target problem by applying minor modifications to the network. Generally, transfer learning is done when the distributions between the source and target are similar and the task is equal. In this paper, we hypothesis that, if the distance between the distributions is different, and the tasks are different transfer learning will help to improve classification performance and speed up the process.

For this purpose we propose unsupervised feature transference using Stacked Denoising Autoencoder (SDA). In unsupervised feature transference approach we explore: 1) transfer learning between completely different tasks drawn from different distributions and, 2) transfer learning between equal tasks drawn from different distributions. We achieved significant improvement on average error rate and on average computation time using SDA on two types of transfer learning approaches to test our hypothesis.

3.6 Paper overview: IbPRIA

This work is accepted and will be held in 7th Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA) at Santiago de Compostela, Spain in the month of June 17-19. I am attaching the camera ready version of the "*Source-Target-Source Classification using Stacked Denoising Autoencoders*".

Deep Transfer Learning (DTL) emerged as a new paradigm in machine learning in which a deep model is trained on a source task and the knowledge acquired is then totally or partially transferred to help in solving a target task. Even though

DTL offers a greater flexibility in extracting high-level features and enabling feature transference from a source to a target task, the DTL solution might get stuck at local minima leading to performance degradation-negative transference-, similar to what happens in the classical machine learning approach. In this paper, we propose the Source-Target-Source (STS) methodology to reduce the impact of negative transference, by iteratively switching between source and target tasks in the training process. The results show the effectiveness of such approach.

4 Works under review at conference:

This work is submitted to Special session on Transfer Learning at IWANN, Spain. I am attaching the submitted version of the "*Ensemble of Deep Transfer Learning Classification*".

Transfer learning algorithms typically assume that the train- ing data and the test data come from different distribution. It is better at adapting to learn new tasks and concepts more quickly and accurately by exploiting previously gained knowledge. Deep Transfer Learning (DTL) emerged as a new paradigm in transfer learning in which a deep model offer greater flexibility in extracting high-level features. DTL offers selec- tive layer based transference, and it is problem specific. In this paper, we propose the Ensemble of Deep Transfer Learning (ETL) methodology to reduce the impact of selective layer based transference and provide optimized framework to work for three major transfer learning cases. Empirical results on character, object and biomedical image recognition tasks achieves that the proposed method indicate statistically significant classification accuracy over the other established transfer learning method.

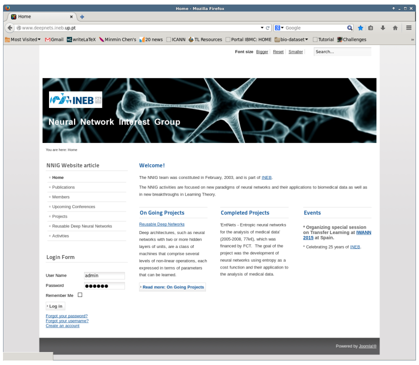
5 Creation and maintenance of the group's webpage

The movement of the group from FEUP to INEB building also initiated series of changes. We transferred the webpage from the FEUP server, gnome.fe.up.pt to the INEB server. The new website is as shown in Fig.1 along with its link as given below.

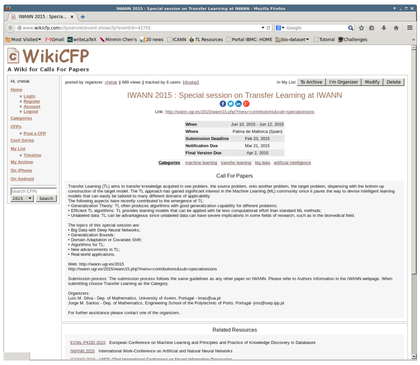
http://www.deepnets.ineb.up.pt/

The webpage is created in Joomla version 2.5 open source content management system and hosted at INEB server. The code's are managed in a open source centralized online repository Bitbucket with version control system. The dissemination of the Transfer learning workshop organised by the group at the news for "The International Work-Conference on Artificial Neural Networks (IWANN), 2015" is also published at the Wiki call for papers as shown in Fig 2.

http://www.wikicfp.com/cfp/servlet/event.showcfp?eventid=42755



(a) Group Homepage



(b) Wiki call for papers

Figure 1: Group Homepage at INEB and Dissemination of IWANN 2015 at Wiki call for papers

A Slides: ICANN conference presentation

Filename: DTL_ICANN_chetak.pdf

Improving Deep Neural Network Performance
by Reusing Features Trained with Transductive Transference

Chetak Kandaswamy,[1,2] Luís M. Silva,[1,3] Luís A. Alexandre,[4]
Jorge M. Santos,[1] Joaquim Marques de Sá[1]

[1]Instituto de Engenharia Biomédica (INEB), Portugal
[2]Universidade de Porto, Portugal
[3]Universidade de Aveiro, Portugal
[4]Universidade da Beira Interior, Portugal

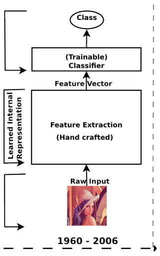
24th International Conference on Artificial Neural Networks,
15-19 September 2014, Hamburg, Germany

Outline

1. Motivation
2. Deep Learning
   - CNN and SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset and Model
   - Reuse: digits vs alphabets.
   - Reuse: Latin digits vs Arabic digits

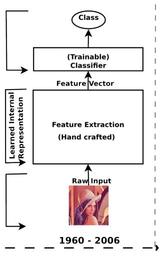
Why deep learning?

- Handcrafted features
- Shallow architectures outperform Deep
- **Failing to train deep Architectures**

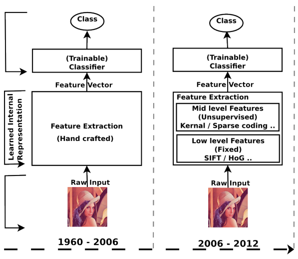


Why deep learning?

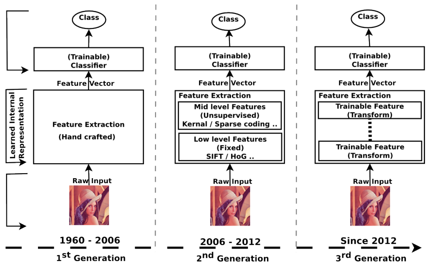
- Handcrafted features
- Shallow architectures outperform ANN
- Failing to train deep Architectures
- 2006: Breakthrough
- **Unsupervised Pre-training is must for Backpropogation**



2nd Generation of Feature Extraction



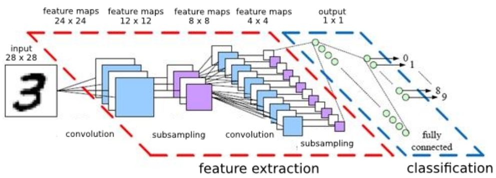
3rd Generation of Feature Extraction



Outline

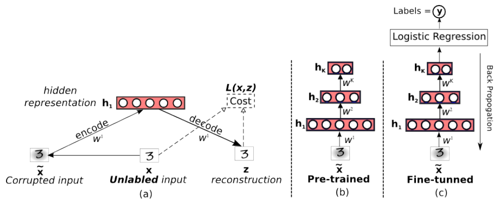
1. Motivation
2. Deep Learning
   - CNN and SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset and Model
   - Reuse: digits vs alphabets.
   - Reuse: Latin digits vs Arabic digits

CNN: Simple and Complex cells



Y. LeCun, et al., Gradient-based learning applied to document recognition. In: proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324 (1998)

Stacked Denoising Autoencoder



$$finetune(pretrain(\mathbf{w}), c)$$

P. Vincent, et al., Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res., (2010).

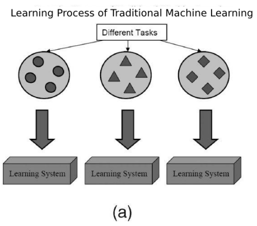
Outline

1 Motivation

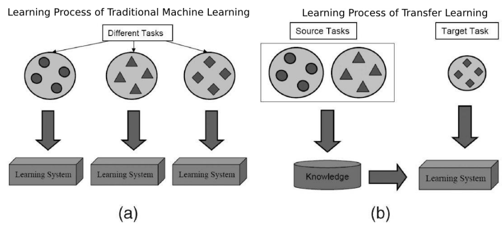
2 Deep Learning
- CNN and SDA
- Transfer Learning
- Reuse approaches:

3 Results
- Dataset and Model
- Reuse: digits vs alphabets.
- Reuse: Latin digits vs Arabic digits

Traditional Machine Learning



Transfer Learning



S. J. Pan and Q. Yang.: A survey on transfer learning. Knowledge and Data Engineering, IEEE

Distributions

Comparing distributions

We use Jensen-shannon divergence (JSD) as a measure to compute the difference between the distributions of two datasets. Given two probability functions $P_S(X)$ and $P_T(X)$,
when $JSD = 0$, we can consider that two distributions are identical,
when $JSD < 0.5$, the distributions are similar,
when $JSD \geq 0.5$, the distributions are different.

Classify alphabets reusing digits

Transductive approach

A machine is trained on a specific problem to solve another specific problem, where the target problem distributions are not necessarily related to the source problem.

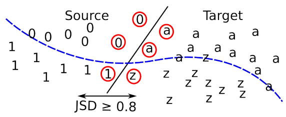


Figure: $P_S(X) \neq P_T(X)$ and $Y_S \neq Y_T$

Settings

distributions are equal, i.e., $P_S(X) = P_T(X)$

1. labels are equal $Y_S = Y_T$
2. labels are not equal $Y_S \neq Y_T$

We consider the transductive case when,

distributions are different, i.e., $P_S(X) \neq P_T(X)$

1. labels are equal $Y_S = Y_T$
2. labels are not equal $Y_S \neq Y_T$

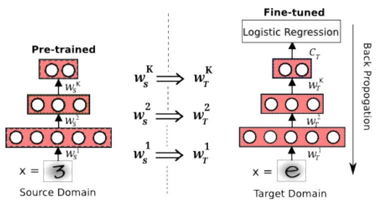
Outline

1 Motivation

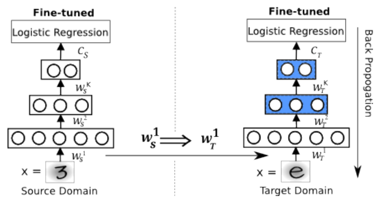
2 Deep Learning
- CNN and SDA
- Transfer Learning
- Reuse approaches:

3 Results
- Dataset and Model
- Reuse: digits vs alphabets.
- Reuse: Latin digits vs Arabic digits

Unsupervised Feature Transfer (UFT)



Supervised Layer-based Feature Transfer (SLFT)



Approach

Table: Baseline and Feature Transference Approach

Approach	Transference	Target Problem
FT	$S(\mathbf{w}_S) \Rightarrow \mathbf{w}_T$	$finetune(\mathbf{w}_T, c_T)$
L1+L2+L3	$S(w_S^1, w_S^2, w_S^3) \Rightarrow w_T^1, w_T^2, w_T^3$	$finetune(c_T)$
L1+L3	$S(w_S^1, w_S^3) \Rightarrow w_T^1, w_T^3$	$finetune(w_T^2, c_T)$
L2+L3	$S(w_S^2, w_S^3) \Rightarrow w_T^2, w_T^3$	$finetune(w_T^1, c_T)$
L1+L2	$S(w_S^1, w_S^2) \Rightarrow w_T^1, w_T^2$	$finetune(w_T^3, c_T)$
L3	$S(w_S^3) \Rightarrow w_T^3$	$finetune(w_T^1, w_T^2, c_T)$
L2	$S(w_S^2) \Rightarrow w_T^2$	$finetune(w_T^1, w_T^3, c_T)$
L1	$S(w_S^1) \Rightarrow w_T^1$	$finetune(w_T^2, w_T^3, c_T)$
UFT	$S(\mathbf{w}_S) \Rightarrow \mathbf{w}_T$	$finetune(\mathbf{w}_T, c_T)$
Baseline	-	$finetune(pretrain(\mathbf{w}_T), c_T)$

Outline

1. Motivation
2. Deep Learning
   - CNN and SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset and Model
   - Reuse: digits vs alphabets.
   - Reuse: Latin digits vs Arabic digits

Dataset characteristics

Table: Average classification test error (%) ($\overline{\varepsilon}$), Average training times (seconds) ($\overline{t}$) for SDA and CNN baseline approach.

Data set		Labels			Instances			SDA		CNN	
Distribution		Y		c	Train	Valid	Test	$\overline{\varepsilon}$	$\overline{t}$	$\overline{\varepsilon}$	$\overline{t}$
Latin	P_L	0-to-9	Y_{09}	10	50,000	10,000	10,000	1.61±0.19	10698	0.93±0.06	1418
Arabic	P_A	•-to-9	$Y_{\bullet 9}$	10	50,000	10,000	10,000	1.37±0.07	8051	0.96±0.06	1209
Lowercase	P_{LC}	a-to-z	Y_{az}	26	13,208	6,604	6,604	4.95±0.16	2997	3.65±0.12	445
Uppercase	P_{UC}	A-to-Z	Y_{AZ}	26	13,208	6,604	6,604	5.01±0.27	2567	3.42±0.10	444

- Chars74k: In this experiment it has been reduced to 28 x 28 pixel to match the MNIST datasets.

SDA Architecture

SDA Architecture
=[576, 400, 256, c]
- PT epochs = 40
- PT learning rate =0.001
- max FT epochs = 1000
- FT learning rate =0.1
- batch size =1

CNN Architecture
=[20* 144, 50 *16, 500, c]
- Kernel size =[20, 50]
- Filter = [5, 5]
- max FT epochs= 200
- FT learning rate =0.1
- batch_size =500

Figure: Network architecture and hyper-parameters for SDA and CNN

Outline

1. Motivation
2. Deep Learning
   - CNN and SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset and Model
   - Reuse: digits vs alphabets.
   - Reuse: Latin digits vs Arabic digits

Classify alphabets reusing digits

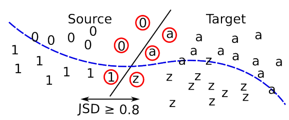


Figure: $P_S(X) \neq P_T(X)$ and $Y_S \neq Y_T$

Classify alphabets reusing digits

Table: Average classification test error (%) ($\bar{\varepsilon}$) obtained for different n_{ds}/c for SLFT approach on CNN model.

Approaches	$X_{ds.upper}$ reuse $X_{ds.latin}$		$X_{ds.lower}$ reuse $X_{ds.latin}$	
Source:	Latin		Latin	
Target:	Uppercase		Lowercase	
$n_{ds.source}/c$:	1320	5000	1320	5000
L1+L2+L3	5.96±0.13	5.32±0.18	6.13±0.13	5.63±0.15
L1+L3	4.49±0.14	4.24±0.10	4.75±0.13	4.57±0.09
L1+L2	3.61±0.12	**3.39±0.12**	3.83±0.06	**3.63±0.13**
L3	4.30±0.13	4.20±0.16	4.62±0.18	4.61±0.14
L2	3.54±0.14	3.43±0.06	3.72±0.11	3.58±0.15
L1	3.43±0.11	**3.35±0.09**	**3.64±0.06**	**3.56±0.11**
BL	**3.42±0.10**	3.42±0.10	3.65±0.12	3.65±0.12

Outline

1. Motivation
2. Deep Learning
   - CNN and SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset and Model
   - Reuse: digits vs alphabets.
   - Reuse: Latin digits vs Arabic digits

Classify Arabic digits reusing Latin digits

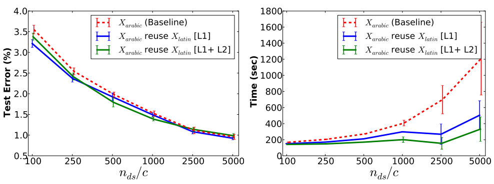


Figure: SLFT for different numbers n_{ds}/c of training samples per class. **Left:** Average classification test error rate. **Right:** Average time taken for classification.

Classify Latin digits reusing Arabic digits

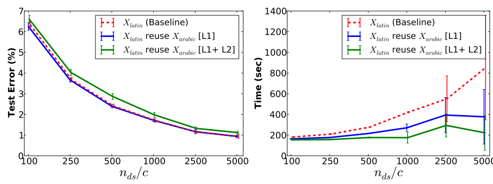


Figure: SLFT for different numbers n_{ds}/c of training samples per class. **Left:** Average classification test error rate. **Right:** Average time taken for classification.

Summary

Table: Average classification test error (%) ($\bar{\varepsilon}$), Average training times (seconds) ($\bar{t}$) by reusing Latin at $n_{ds}/c = 1320$

Approaches		Lowercase		Uppercase	
		$\bar{\varepsilon}$	$\bar{t}$	$\bar{\varepsilon}$	$\bar{t}$
SDA	BL	4.95±0.16	2997	5.01±0.27	2567
SDA	SLFT: L1	4.72±0.17	2261	4.72±0.18	2515
SDA	UFT	4.67±0.38	1148	4.65±0.19	1498
SDA	SLFT: FT	**4.57±0.08**	1020	**4.58±0.19**	1180
CNN	SLFT: L1+L2	3.83±0.06	196	3.61±0.12	197
CNN	BL	3.65±0.12	445	**3.42±0.10**	444
CNN	SLFT: L1	**3.64±0.06**	292	3.43±0.11	293

Summary

- Deep neural network performance is improved using transductive approach
- Using UFT we achieved 7% relative improvement on average error rate and 41% on average computation time.
- Using SLFT,
  - reusing L1 and L2 we achieved 5.7% relative improvement in the average error rate,
  - reusing all three hidden layers we achieved 8.5% relative improvement for the average error rate and 54% speed up w.r.t the baseline
- JSD for
  - Arabic vs lowercase = 0.99,
  - Latin vs lowercase = 0.80.
  - As the distribution of Latin dataset is more related to lowercase dataset, we achieve better classification performance using Latin as source problem.

Conclusion and Discussion

- Deep neural network performance is improved using transductive approach
- Using Unsupervised Feature Transference and Supervised Layer based Feature Transference,
- Improves accuracy and speed up the computation time.
- Challenge to choose suitable source problem, can be reduced using Jensenn-Shannon divergence measure.

Acknowledgement

INEB Instituto de Engenharia Biomédica · FEUP · universidade de aveiro · COMPETE · QREN Quadro de Referência Estratégico Nacional · União Europeia

B Slides: SMC conference presentation

Improving Transfer Learning Accuracy by Reusing Stacked Denoising Autoencoders

Chetak Kandaswamy,[1,2] Luís M. Silva,[1,3] Luís A. Alexandre,[4]
Ricardo Sousa,[1] Jorge M. Santos,[1] Joaquim Marques de Sá[1]

[1]Instituto de Engenharia Biomédica (INEB), Portugal
[2]Universidade de Porto, Portugal
[3]Universidade de Aveiro, Portugal
[4]Universidade da Beira Interior, Portugal

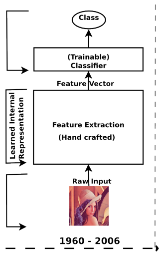
2014 IEEE International Conference on Systems, Man, and Cybernetics,
October 5-8, 2014, San Diego, California

Outline

1. Motivation
2. Deep Learning
   - SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset
   - Comparision of USDA Vs baseline.
   - Comparision of SSDA Vs baseline

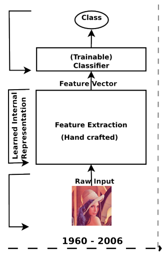
Why deep learning?

- Handcrafted features
- Shallow architectures outperform Deep
- **Failing to train deep Architectures**

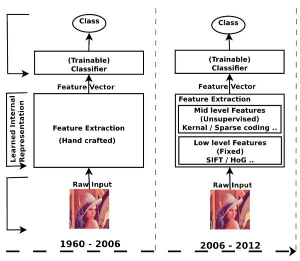


Why deep learning?

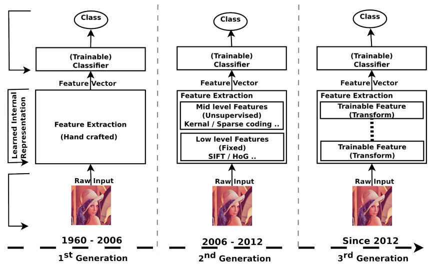
- Handcrafted features
- Shallow architectures outperform ANN
- Failing to train deep Architectures
- 2006: Breakthrough
- **Unsupervised Pre-training is must for Backpropogation**



2nd Generation of Feature Extraction



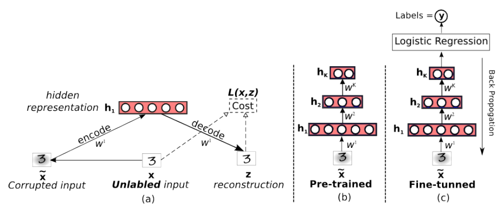
3rd Generation of Feature Extraction



Outline

1. Motivation
2. Deep Learning
   - SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset
   - Comparision of USDA Vs baseline.
   - Comparision of SSDA Vs baseline

Stacked Denoising Autoencoder



$$finetune(pretrain(\mathbf{w}), c)$$

P. Vincent, et al., Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. J. Mach. Learn. Res., (2010).

Outline

1. Motivation
2. Deep Learning
   - SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset
   - Comparision of USDA Vs baseline.
   - Comparision of SSDA Vs baseline

Transfer Learning

Transfer Learning

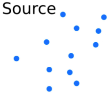
A process that allows reusing learning machines previously trained for a given Source problem in order to solve, with minor modifications, a different target problem.

Transfer Learning

Transfer Learning

A process that allows reusing learning machines previously trained for a given Source problem in order to solve, with minor modifications, a different target problem.

Source

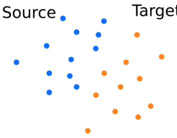


Transfer Learning

Transfer Learning

A process that allows reusing learning machines previously trained for a given Source problem in order to solve, with minor modifications, a different target problem.

Source Target

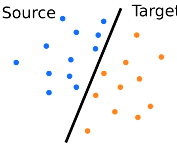


Transfer Learning

Transfer Learning

A process that allows reusing learning machines previously trained for a given Source problem in order to solve, with minor modifications, a different target problem.

Source Target

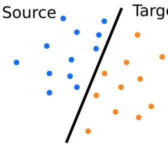


Distributions

Jensen-shannon divergence (JSD)

Given two probability functions $P_S(X)$ and $P_T(X)$: We measure the difference in disribution between the source and the target dataset.

Source Target



JSD=0
the two distributions are **identical**,

JSD<0.5
the two distributions are **similar**,

JSD>0.5
the two distributions are **different**

Transfer Learning Settings

Domain Adaptation

A machine learns to perform a task on training instances drawn from the source problem, but then needs to perform the ***same task*** on the target problem instances drawn from a ***related distribution***.

Source Target

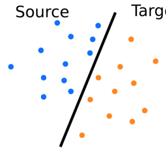


Figure : $P_S(X) \approx P_T(X)$ and $Y_S = Y_T$

Transfer Learning Settings

distributions are equal, i.e., $P_S(X) = P_T(X)$

1. labels are equal $Y_S = Y_T$
2. labels are not equal $Y_S \neq Y_T$

We consider the case when distributions are different,

distributions are different, i.e., $P_S(X) \neq P_T(X)$

1. labels are equal $Y_S = Y_T$
2. labels are not equal $Y_S \neq Y_T$

Classify alphabets reusing digits

Different Distribution

Transfer Learning between completely **different tasks** drawn from **different distributions**.

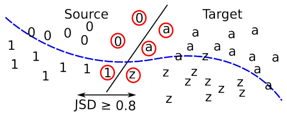


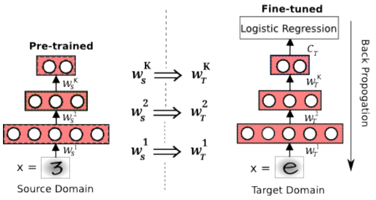
Figure : $P_S(X) \neq P_T(X)$ and $Y_S \neq Y_T$

Outline

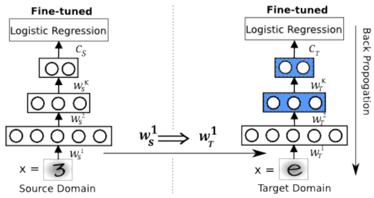
1 Motivation
2 Deep Learning
- SDA
- Transfer Learning
- Reuse approaches:

3 Results
- Dataset
- Comparision of USDA Vs baseline.
- Comparision of SSDA Vs baseline

Unsupervised Feature Transfer (USDA)



Supervised Layer-based Feature Transfer (SSDA)



Approach

Table : Baseline and Feature Transference Approach

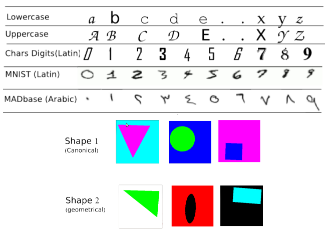
	Approach	Transference	Target Problem
SSDA	FT	$S(\mathbf{w}_S) \Rightarrow \mathbf{w}_T$	$finetune(\mathbf{w}_T, c_T)$
	L1+L2+L3	$S(w_S^1, w_S^2, w_S^3) \Rightarrow w_T^1, w_T^2, w_T^3$	$finetune(c_T)$
	L1+L3	$S(w_S^1, w_S^3) \Rightarrow w_T^1, w_T^3$	$finetune(w_T^2, c_T)$
	L2+L3	$S(w_S^2, w_S^3) \Rightarrow w_T^2, w_T^3$	$finetune(w_T^1, c_T)$
	L1+L2	$S(w_S^1, w_S^2) \Rightarrow w_T^1, w_T^2$	$finetune(w_T^3, c_T)$
	L3	$S(w_S^3) \Rightarrow w_T^3$	$finetune(w_T^1, w_T^2, c_T)$
	L2	$S(w_S^2) \Rightarrow w_T^2$	$finetune(w_T^1, w_T^3, c_T)$
	L1	$S(w_S^1) \Rightarrow w_T^1$	$finetune(w_T^2, w_T^3, c_T)$
USDA		$S(\mathbf{w}_S) \Rightarrow \mathbf{w}_T$	$finetune(\mathbf{w}_T, c_T)$
Baseline		-	$finetune(pretrain(\mathbf{w}_T), c_T)$

Outline

1 Motivation
2 Deep Learning
- SDA
- Transfer Learning
- Reuse approaches:

3 Results
- Dataset
- Comparision of USDA Vs baseline.
- Comparision of SSDA Vs baseline

Dataset characteristics



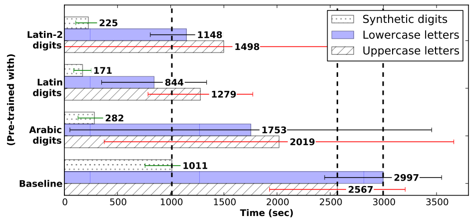
Dataset characteristics

Data set		Labels			Instances		
Distribution		Ω		c	Train	Valid	Test
Lowercase	P_{LC}	a-to-z	Ω_{az}	26	13,208	6,604	6,604
Uppercase	P_{UC}	A-to-Z	Ω_{AZ}	26	13,208	6,604	6,604
Digits	P_D	0-to-9	Ω_{09}	10	5,080	2,540	2,540
Arabic	P_A	•-to-9	$\Omega_{\bullet 9}$	10	50,000	10,000	10,000
Latin	P_L	0-to-9	Ω_{09}	10	50,000	10,000	10,000
Latin-2	P_{L2}	0-to-9	Ω_{09}	10	13,208	6,604	10,000
Shape1	P_{Sh1}	'eqt','cir','sqr'	Ω_{sh1}	3	10,000	5,000	5,000
Shape2	P_{Sh2}	'tri','ell','rec'	Ω_{sh2}	3	10,000	5,000	5,000

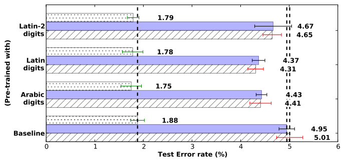
Outline

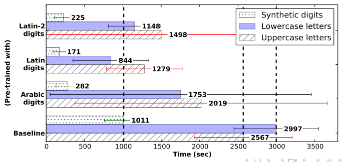
1. Motivation
2. Deep Learning
   - SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset
   - Comparision of USDA Vs baseline.
   - Comparision of SSDA Vs baseline

Hard transfer problems: Computational efficiency

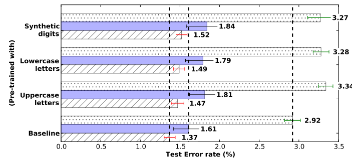


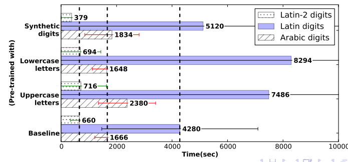
Hard transfer problems





Reverse transfer problems





Outline

1. Motivation
2. Deep Learning
   - SDA
   - Transfer Learning
   - Reuse approaches:
3. Results
   - Dataset
   - Comparision of USDA Vs baseline.
   - Comparision of SSDA Vs baseline

Average Test Error of SSDA approaches

Table : Average Test Error (%) of SSDA approaches for Harder case and Reverse case problems

	Harder Transfer			Reverse Transfer		
Target:	uppercase	lowercase	shape1	latin-2	latin-2	shape2
Source:	latin-2	latin-2	shape2	uppercase	lowercase	shape1
Labels:	$\mathscr{Y}_S \neq \mathscr{Y}_T$	$\mathscr{Y}_S \neq \mathscr{Y}_T$	$\mathscr{Y}_S = \mathscr{Y}_T$	$\mathscr{Y}_S \neq \mathscr{Y}_T$	$\mathscr{Y}_S \neq \mathscr{Y}_T$	$\mathscr{Y}_S = \mathscr{Y}_T$
Approaches	$\overline{\varepsilon}$	$\overline{\varepsilon}$	$\overline{\varepsilon}$	$\overline{\varepsilon}$	$\overline{\varepsilon}$	$\overline{\varepsilon}$
reuse FT	**4.58±0.19** ⇑	**4.57±0.08** ⇑	9.13±1.57 ⇓	3.49±0.19 ⇓	3.46±0.18 ⇓	25.52±14.71 ⇓
reuse L1+L2+L3	10.93±0.5 ⇓	10.70±0.3 ⇓	11.10±2.0 ⇓	9.29±0.54 ⇓	8.68±0.39 ⇓	39.81±09.88 ⇓
reuse L1+L3	5.28±0.16 ⇓	5.31±0.18 ⇓	**5.23±1.45** ⇑	4.14±0.24 ⇓	4.14±0.15 ⇓	20.34±16.42 ○
reuse L2+L3	5.41±0.25 ⇓	5.61±0.11 ⇓	9.94±2.54 ⇓	4.40±0.13 ⇓	4.36±0.12 ⇓	26.27±15.47 ⇓
reuse L1+L2	5.60±0.19 ⇓	5.68±0.10 ⇓	6.88±1.89 ⇑	4.22±0.13 ⇓	4.15±0.14 ⇓	22.69±15.43 ○
reuse L3	4.81±0.30 ○	5.17±0.15 ⇓	10.34±0.9 ⇓	3.86±0.11 ⇓	3.82±0.17 ⇓	26.89±13.53 ⇓
reuse L2	4.88±0.17 ○	4.95±0.13 ○	11.14±1.5 ⇓	3.78±0.13 ⇓	3.76±0.14 ⇓	29.71±13.79 ⇓
reuse L1	4.72±0.18 ⇑	4.72±0.17 ⇑	7.29±1.42 ○	3.59±0.14 ⇓	3.59±0.19 ⇓	23.96±15.45 ○
Baseline	5.01±0.27	4.95±0.16	7.88±0.93	**2.92±0.10**	**2.92±0.10**	**15.51±6.31**

⇑, ⇓, ○ statistically significant improvement or degradation or no change than baseline. The best test result obtained for a target dataset are marked in bold.

Summary

Table : Average classification test error (%) ($\overline{\varepsilon}$), Average training times (seconds) ($\overline{t}$) by reusing Latin at $n_{ds}/c = 1320$

Approaches		Lowercase		Uppercase	
		$\overline{\varepsilon}$	$\overline{t}$	$\overline{\varepsilon}$	$\overline{t}$
SDA	BL	4.95±0.16	2997	5.01±0.27	2567
SDA	SLFT: L1	4.72±0.17	2261	4.72±0.18	2515
SDA	UFT	4.67±0.38	1148	4.65±0.19	1498
SDA	SLFT: FT	**4.57±0.08**	1020	**4.58±0.19**	1180

Summary

- Deep neural network performance is improved using Hard transfer problems
- Using USDA we achieved 7% relative improvement on average error rate and 41% on average computation time.
- Using SSDA,
  - reusing L1 and L2 we achieved 5.7% relative improvement in the average error rate,
  - reusing all three hidden layers we achieved 8.5% relative improvement for the average error rate and 54% speed up w.r.t the baseline
- JSD for
  - Arabic vs lowercase = 0.99,
  - Latin vs lowercase = 0.80.
  - As the distribution of Latin dataset is more related to lowercase dataset, we achieve better classification performance using Latin as source problem.

Conclusion and Discussion

- Deep neural network performance is improved using hard transfer problems
- Using Unsupervised Feature Transference and Supervised Layer based Feature Transference,
- Improves accuracy and speed up the computation time.
- Challenge to choose suitable source problem, can be reduced using Jensenn-Shannon divergence measure.

Acknowledgement