THE EFFECTIVENESS OF TRANSFER LEARNING IN ELECTRONIC HEALTH RECORDS DATA

Sebastien Dubois, Nathanael Romano, Kenneth Jung, & Nigam Shah
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA 90035, USA
{sdubois,naromano,kjung,nigam}@stanford.edu

David C. Kale*
USC Information Sciences Institute
University of Southern California
Marina Del Rey, CA 90292, USA
kale@isi.edu

ABSTRACT

The application of machine learning to clinical data from Electronic Health Records is limited by the scarcity of meaningful labels. Here we present initial results on the application of transfer learning to this problem. We explore the transfer of knowledge from source tasks in which training labels are plentiful but of limited clinical value to more meaningful target tasks that have few labels.

1 INTRODUCTION

Computer vision has seen a remarkable recent breakthrough in the widespread success of *transfer learning*, in which knowledge from one task is used to aid learning of another task, often one with far less labeled data. In particular, large scale convolutional neural nets (ConvNets) trained to on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012) have been repeatedly repurposed to help solve other sometimes very different problems, often with parameters adapted by additional training on new data or used as a fixed feature extractors for an entirely separate classifier (Simonyan & Zisserman, 2014; Long et al., 2014; Karpathy & Fei-Fei, 2015).

In this workshop paper, we present an initial investigation into whether and how similar ideas can be applied to clinical problems and electronic health records (EHR) data. We focus on the problem of *electronic phenotyping*, where our goal is to train statistical classifiers to answer questions like "Does this patient have diabetes?" from digital health records data (Oellrich et al., 2015). Phenotyping has a variety of applications in cohort construction for genomic studies (Crawford et al., 2014), quality improvement (Weiner & Embi, 2009), risk adjustment (Elixhauser et al., 1998), and detection undiagnosed diseases (Lindbeg et al., 1968).

Ground truth phenotype labels are not recorded during delivery of care and so are typically unavailable in large numbers during training.[1] Obtaining high quality labels after the fact is time-consuming and expensive, even in combination with active learning, because the domain experts are highly trained physicians (Chen et al., 2013).[2] Further, this problem is common to many of the questions that biomedical informatics researchers seek to address; manual labeling simply does not scale.

We aim to overcome this challenge through the application of transfer learning: we first train a neural network to predict a source task with ubiquitous labels that, while not directly related to our phenotype targets, contain enough information to provide a training signal for learning a useful

*Mr. Kale performed this research as an affiliated researcher with the Stanford Center for Biomedical Informatics Research.

[1] Billing and diagnostic codes are known to be unreliable indicators of actual disease.(O'Malley et al., 2005).

[2] Medical students, while trainees, are also highly trained and expensive!

representation of raw EHR data. Specifically, we apply a variation of the *split brain autoencoder* architecture described in Zhang et al. (2016) to predict, e.g., prescriptions from diagnostic codes. We then use this dense, lower dimensional representation as features in a classifier trained to predict the target task using a much smaller number of reliable labels.

2 DATA AND TASK SETUP

We extracted patient data spanning 2009 through 2014 from the Stanford Translational Research Integrated Database Environment (STRIDE) (Lowe et al., 2009), a repository of de-identified patient data from the Electronic Medical Record system used at Stanford Hospital. The patients were split into training, validation and test sets (122860, 10000, and 50000 patients, respectively). In the reported experiments, all models are trained on the training set, while hyperparameters are tuned using 5-fold cross-validation. We report held-out performance on the validation set, leaving the test set untouched for future work.

For each patient, the data consisted of timestamped occurrences of diagnosis (ICD-9), procedure (CPT) and drug (RXCUI) codes. This data was split into input and target eras such that for each patient the last year of data was reserved to form prediction targets, while the previous years of data were used to form inputs for that patient. Input data were aggregated into counts across the entire input era. There were 8927 raw input features, and the input space is very sparse, with less than 0.5% non-zero.

Our target task is to predict phenotypes. Because of the aforementioned unavailability of ground truth phenotype labels, in this proof of concept work we approximate ground truth phenotype labels using diagnostic code categories based on the HCUP Clinical Classification Software (Cost et al., 2014). We use ten phenotype targets with prevalence ranging from 11% to 28%. To reduce target leakage (our phenotype labels are derived from diagnosis codes, which are included in our input), we perform a temporal variant of phenotyping: we predict the occurrence of phenotype labels in the future (target era) based on data from the input era.

3 METHODS AND EXPERIMENTS

The source task for our initial experiments is cross channel prediction similar to that described in Zhang et al. (2016) in which auto-encoders are trained to predict drug codes from diagnosis and procedure codes, and vice versa. The source task thus uses data only from the input era of the training patients. Auto-encoders were trained independently of each other (med2rx and rx2med). For both models, we found using 2 layers with tanh activations, dropout after each hidden layer (dropout probility 0.06) and l2 regularization on the weights worked well, with mean AUROCs of 0.88 and 0.83 for the med2rx and rx2med tasks respectively. The med2rx and rx2med models used 936 and 746 hidden units respectively, resulting in a final concatenated representation of dimensionality 1682.

Transfer learning was carried out by using med2rx and rx2med models as fixed feature extractors. The concatenated last hidden layer activations from the cross channel auto-encoders were used as inputs into L1 regularized logistic regression models, with the regularization hyper-parameter set by five fold cross validation to optimize AUROC, and fit separately for each target code group. Because label sparsity is a ubiquitous problem in the medical setting, we focused our experiments on the effectiveness of this transfer learning scheme using small sample sizes. 20 subsamples of N=500 patients from the training set were used for each of the targets, and models fit using 5-fold cross validation on each subsample to select the regularization hyper-parameter. The resulting models were then evaluated on the validation set patients (the test set is reserved for future work).

Our baseline for comparison is L1 regularized logistic regression applied to the raw inputs, tuned via 5-fold cross validation on the same training subsamples and evaluated on the validation patients.

4 RESULTS

We compared the performance of transfer learning versus a baseline of regularized logistic regression on the sparse, high dimensional raw inputs on ten diagnosis code groups selected for high

prevalence. Table 1 shows the mean over the subsamples of the area under the ROC for each target, along with target prevalence. Overall, the majority of the targets show a modest benefit from transfer learning.

Table 1: Results

	Mean AUC (SE)		
Code Group Name	**Baseline**	**Auto-encoder**	**Delta**
Cancer of breast	0.842 (0.013)	0.845 (0.009)	0.0033
Thyroid disorders	0.753 (0.015)	0.777 (0.016)	0.024
Diabetes w/o complication	0.768 (0.008)	0.793 (0.006)	0.024
Disorders of lipid metabolism	0.794 (0.005)	0.814 (0.005)	0.021
Deficiency and other anemia	0.645 (0.016)	0.677 (0.015)	0.032
Retinal detachments	0.746 (0.014)	0.745 (0.025)	-0.0017
Essential hypertension	0.798 (0.008)	0.847 (0.004)	0.049
Coronary atherosclerosis and other heart disease	0.734 (0.018)	0.812 (0.010)	0.078
Cardian disrhythmias	0.662 (0.010)	0.722 (0.013)	0.059
Chronic kidney disease	0.745 (0.016)	0.782 (0.028)	0.037
Osteoarthritis	0.661 (0.010)	0.730 (0.014)	0.069
Spondylosis and intervertebral disc. disorders	0.639 (0.009)	0.688 (0.011)	0.049
Anxiety disorders	0.625 (0.016)	0.665 (0.020)	0.04
Mood disorders	0.687 (0.013)	0.753 (0.009)	0.066
		Mean	0.039

5 Related Work

One classic solution to learning without labels is active learning, which has been applied successfully to phenotyping (Chen et al., 2013). However, in many practical settings, it suffers from the *cold start* dilemma: when starting with zero labeled examples, it can perform no better than random sampling until it acquires enough labeled examples to train a reasonably good classifier Kale & Liu (2013). This is compounded when positive examples are relatively rare, as is the case for many clinical phenotypes Cost et al. (2014).

Our research follows up on work by Agarwal et al. (2016); Halpern et al. (2016), who showed that robust phenotype classifiers could be trained via distant supervision: "noisy" labels were assigned using a semi-automated deterministic labeling function guided by domain knowledge. Models trained on such labeled data were found to be relatively robust to the label noise and even outperformed the noisy labeling function on held out ground truth labeled data.

Miotto et al. (2016) tackle temporal phenotyping with a similar task setup and general approach: they train use an autoencoder to patient history into dense features, which are then fed into a separate classifier that predicts future phenotypes. However, their use of a complex nonlienar classifier makes it difficult to assess the relative contributions of learned representation vs. classifier to prediction performance. Choi et al. (2015) use a long short-term memory network to model sequences of diagnostic codes, a proxy problem for disease progression, and show that this setup and architecture can be used to perform transfer learning to new data sets for the same task.

References

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, pp. ocw028, 2016.

Yukun Chen, Robert J Carroll, Eugenia R McPeek Hinz, Anushi Shah, Anne E Eyler, Joshua C Denny, and Hua Xu. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. *Journal of the American Medical Informatics Association*, 20(e2): e253–e259, 2013.

Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor ai: Predicting clinical events via recurrent neural networks. *arXiv preprint arXiv:1511.05942*, 2015.

Healthcare Cost, Utilization Project (HCUP), et al. Introduction to the hcup national inpatient sample (nis) 2012. *Agency for Healthcare Research and Quality, Rockville, MD*, 2014.

Dana C Crawford, David R Crosslin, Gerard Tromp, Iftikhar J Kullo, Helena Kuivaniemi, M Geoffrey Hayes, Joshua C Denny, William S Bush, Jonathan L Haines, Dan M Roden, et al. emergeing progress in genomics—the first seven years. *Frontiers in genetics*, 5:184, 2014.

Anne Elixhauser, Claudia Steiner, D Robert Harris, and Rosanna M Coffey. Comorbidity measures for use with administrative data. *Medical care*, 36(1):8–27, 1998.

Yoni Halpern, Steven Horng, Youngduck Choi, and David Sontag. Electronic medical record phenotyping using the anchor and learn framework. *Journal of the American Medical Informatics Association*, 23(4):731, 2016.

David Kale and Yan Liu. Accelerating active learning with transfer learning. In *2013 IEEE 13th International Conference on Data Mining*, pp. 1085–1090. IEEE, 2013.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, pp. 1106–1114, 2012.

D.A.B. Lindbeg, L.R. Rowland, C.R. Jr. Buch, W.F. Morse, and S.S. Morse. Consider: A computer program for medical instruction. In *Proceedings of the Ninth IBM Medical Symposium*, volume 69, pp. 54, 1968.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *CoRR*, abs/1411.4038, 2014. URL `http://arxiv.org/abs/1411.4038`.

H. J. Lowe, T. A. Ferris, P. M. Hernandez, and S. C. Weber. Stride–an integrated standards-based translational research informatics platform. *AMIA Annu Symp Proc*, 2009:391–5, 2009.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6, 2016.

Anika Oellrich, Nigel Collier, Tudor Groza, Dietrich Rebholz-Schuhmann, Nigam Shah, Olivier Bodenreider, Mary Regina Boland, Ivo Georgiev, Hongfang Liu, Kevin Livingston, Augustin Luna, Ann-Marie Mallon, Prashanti Manda, Peter N. Robinson, Gabriella Rustici, Michelle Simon, Liqin Wang, Rainer Winnenburg, and Michel Dumontier. The digital revolution in phenotyping. *Briefings in Bioinformatics*, 2015.

Kimberly J O'Malley, Karon F Cook, Matt D Price, Kimberly Raiford Wildes, John F Hurdle, and Carol M Ashton. Measuring diagnoses: Icd code accuracy. *Health services research*, 40(5p2): 1620–1639, 2005.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 568–576, 2014.

Mark G Weiner and Peter J Embi. Toward reuse of clinical data for research and quality improvement: the end of the beginning? *Annals of internal medicine*, 151(5):359–360, 2009.

Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. *arXiv preprint arXiv:1611.09842*, 2016.