Transfer Representation-Learning for Anomaly Detection

Jerone T. A. Andrews[†,‡] JERONE.ANDREWS@CS.UCL.AC.UK
Thomas Tanay[†] THOMAS.TANAY.13@UCL.AC.UK
Edward J. Morton[⋆] EMORTON@RAPISCANSYSTEMS.COM
Lewis D. Griffin[†] L.GRIFFIN@CS.UCL.AC.UK
[†]Department of Computer Science, University College London, UK
[‡]Department of Statistical Science, University College London, UK
[⋆]Rapiscan Systems Ltd, USA

Abstract

We evaluate transfer representation-learning for anomaly detection using convolutional neural networks by: (i) transfer learning from pre-trained networks, and (ii) transfer learning from an auxiliary task by defining sub-categories of the normal class. We empirically show that both approaches offer viable representations for the task of anomaly detection, without explicitly imposing a prior on the data.

1. Introduction

Anomaly detection is a general term used to describe the task of discovering data items that are historically atypical with respect to expected behaviour patterns. For example, an anomaly may indicate the presence of breast cancer in mammograms (Dheeba et al., 2014), surface metal landmines in hyperspectral images (Ranney & Soumekh, 2006), or hazardous algal blooms in satellite imagery of water surface colour (Stumpf et al., 2003). Ergo, the concept of what defines an anomaly will vary between application domains. Due to this variation, it is often not a simple matter of transferring methodologies developed for a disparate task.

Many of the successes in pattern recognition and classification are dependent on hand-crafted visual representations. However, when only normal data is available, how do we go about developing discriminative features for what is in essence a binary-classification problem when we have only seen one of the classes?

Deep representation-learning, in particular convolutional neural networks (CNNs) (LeCun et al., 1998), offer a versatile method for automatically discovering multiple levels of representation, within data, that are extremely adept when utilised in tasks such as classification. Many of the layers non-linearly transform their input, creating more abstract, task-specific representations that are insensitive to large unimportant variations, yet highly-tuned to the important particularities (LeCun et al., 2015). Furthermore, these representations are often generic enough that they can be transferred to dissimilar vision tasks and still achieve highly competitive results (Donahue et al., 2013; Razavian et al., 2014; Oquab et al., 2014; Zeiler & Fergus, 2014) when compared with their more elaborate hand-engineered counterparts.

In this paper, we investigate two representation-learning frameworks for anomaly detection: (i) transfer learning of pre-trained deep convolutional representations, and (ii) transfer learning of deep convolutional representations from an *auxiliary* task. In the former, we transfer learn representations from a related (sharing the same input space) supervised domain to our tasks. In the latter, we learn representations from scratch, on a moderately sized dataset, by training a CNN to make distinctions within the normal class. In both, our anomaly detection systems are strictly constructed on normal data only. Moreover, we empirically show that both approaches offer viable representations for the task of anomaly detection, without explicitly imposing a prior on the data.

We start by reviewing related work on representation-learning, then move on to describe our anomaly detection tasks in Section 3. In Section 4, we detail our representation-learning frameworks, and analyse the empirical results in Section 5.

2. Related work

Our approach is related to transfer representation-learning, where rich representations are learnt in a *source* task, using convolutional neural networks, with the aim of transferring

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

them to a different *target* task. Recent work has been highly successful in leveraging CNN representations learnt on large-scale, fully-supervised, computer vision datasets to other visual tasks with insufficient training data, e.g. scene classification (Donahue et al., 2013), object classification (Zeiler & Fergus, 2014; Oquab et al., 2014), object localisation (Sermanet et al., 2013; Girshick et al., 2014), attribute-detection and fine-grained recognition (Razavian et al., 2014). Whilst many of the successes have primarily been in supervised representation-learning, there is a growing corpus of work on unsupervised representation-learning, which focus on a discriminative approach, (Ahmed et al., 2008; Collobert et al., 2011; Dosovitskiy et al., 2014), based on the idea of creating auxiliary tasks in order to learn robust, generic data representations.

3. Datasets

In order to evaluate the usefulness of transfer learnt representations, we conceive a range of anomaly detection tasks, listed in Table 1, with combinations of tight and diverse, normal and anomaly classes, by employing the following three datasets:

X-ray transmission images of freight containers (Andrews et al., 2016) consists of 5,120 greyscale images of freight containers containing cargo (non-empty) and containers containing no cargo (empty) sized 9×32. All images vary due to small differences in freight containers and their furniture, while cargo images also vary in the cargo.

MNIST handwritten digits (LeCun et al., 1998) contains a total of 70,000 greyscale handwritten digits 0 through 9 of size 28×28.

Augmented CASIA-WebFace (Yi et al., 2014) consists of 988,828 greyscale celebrity face images of 10,575 subjects sized 100×100.

Table 1. Anomaly detection tasks.

#	TASK	NORMAL CLASS	ANOMALY CLASS
1	MNIST-1	5	2
2	MNIST-2	5	0, 2, 4, 6, 8
3	MNIST-3	1, 3, 5, 7, 9	2
4	MNIST-4	1, 3, 5, 7, 9	0, 2, 4, 6, 8
5	X-RAY-1	EMPTY	NON-EMPTY
6	X-RAY-2	NON-EMPTY	EMPTY
7	CASIA	MALE	FEMALE

4. Representation-learning frameworks

Below we outline our approaches, which follow one of two routes: (i) transfer learning of deep convolutional representations from a pre-trained CNN, or (ii) transfer learning of deep convolutional representations from an auxiliary task. For all extracted representations, we will use as a baseline for comparison the original greyscale intensity image representations (prior to any network processing, such as resizing).

4.1. Transfer learning pre-trained representations

We utilise two publicly available[1] pre-trained CNN models: ImageNet-MatConvNet-VGG-F (VGG-F) and ImageNet-MatConvNet-VGG-M (VGG-M), to transfer learn representations, based on the networks outlined in (Chatfield et al., 2014). The networks were trained for the image classification task ILSVRC 2012 (Russakovsky et al., 2015) using 1.2 million colour images of 1,000 diverse object categories. A shorthand notation of the architectures can be found in Table 2.

Each image input to one of the networks is first preprocessed: (i) it is resized to $224 \times 224 \times 3$, and (ii) has the mean image from the trained network subtracted. We extract representations learnt from layers: P_5, FC_6, FC_7 and FC_8, which we rescale to unit length. Using tasks 1-6, we assess the performance of these representations by randomly selecting 2,048 normal samples for training a one-class classifier and 1,024 (512 normal and 512 anomaly) unseen samples for testing. This procedure is repeated three times, with the test set kept fixed.

4.2. Transfer learning auxiliary representations

We formulate an auxiliary, fine-grained classification, task so as to transfer learn representations. Our aim is to learn a mapping from normal samples to their sub-category. To assess this strategy, we use task 7, with sub-categories defined as the identity of a male.

We select 5,045 male subjects consisting of 442,362 face image samples. Next, we randomly sample without replacement 432,272 face images from the 5,045 subjects to train a CNN on inputs of size $100 \times 100 \times 1$ using gradient descent with momentum, batch normalisation (Ioffe & Szegedy, 2015), and softmax loss. A shorthand notation of the CNN (Male-5045) architecture can be found in Table 2. The hyper-parameters used are: momentum 0.9; weight decay $5 \cdot 10^{-4}$; initial learning rate 10^{-2} for the first 150 epochs, 10^{-3} for the next 75 epochs and 10^{-4} for the final 75 epochs. The layers are initialised from a standard Normal distribution, and the activation function for all weight layers, except for the fully-connected layer, is the rectified linear unit (ReLU).

For evaluation, we randomly select 20,180 male face im-

[1]Convolutional Neural Networks for MATLAB (MatConvNet): http://www.vlfeat.org/matconvnet.

ages used to learn the CNN, for training a one-class classifier and 5,000 (2,500 male and 2,500 female) unseen samples for testing. We extract representations learnt from layers: P_8 and FC_9, which we rescale to unit length. This procedure is repeated three times, with the test set kept fixed.

4.3. Classification

Given a set of positive training samples, $\mathbf{x}_i \in R^n, i = 1, 2, \ldots, l$, we apply a linear one-class ν-support vector machine (OCSVM) (Schölkopf et al., 2001; Chang & Lin, 2011) to estimate the support of the high-dimensional distribution. The parameter $\nu \in (0, 1]$ is an upper bound on the fraction of training samples considered out-of-class and a lower bound on the fraction of training samples used as support vectors (SVs). The OCSVM decision function for a sample $\mathbf{x}$ is: $f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i \mathbf{x}^{\mathrm{T}} \mathbf{x}_i - \rho$, where $0 < \alpha_i < \frac{1}{\nu l}$ are the coefficients of the SVs ($\alpha_i = 0$ otherwise) and ρ is a bias term.

For each task, we obtain the area under the receiver operating characteristic (AUROC) averaged across 21 evenly spaced values of the OCSVM hyper-parameter, $\nu \in [0.01, 0.99]$. We report the mean and the standard deviation attained on the fixed test set over the three trials.

5. Results and analysis

Here we analyse the results of transfer learning representations on our anomaly detection tasks.

5.1. Transfer learning pre-trained representations

Table 3 displays the results of the VGG-F and the VGG-M network layers when applied to tasks 1-6. In both networks, the pooling layer P_5 is on average the top-performing representation, followed by the fully-connected representation FC_7, which is marginally better than representations FC_6 and FC_8. This is to be expected, as the learnt weights in the deeper layers become increasingly specific to the task the network was trained for, thus making them less generic to our semantically dissimilar datasets. Whereas, the greyscale intensity images, GS, offer unreliable representations, which is emphasised by their performance in tasks 5 and 6. In task 5, when the normal class are empty homogeneous freight containers it is commensurate with the CNN representations, however when we swap the class labels, as in task 6, and the normal class becomes the non-empty heterogeneous (ranging from seemingly empty to full) freight containers, the empties become inliers. Performing a Wilcoxon signed rank test gives evidence of a statistically significant difference (p-value of 0.03) between each transfer learnt representation, from both networks, and the greyscale intensity representation.

We also see from Table 3 that the VGG-F network outperforms VGG-M. Moreover, a Wilcoxon signed rank test gives a statistically significant difference (p-value of 0.00) between the two. This is contrary to what was reported in (Chatfield et al., 2014), where the VGG-M outperformed the VGG-F when applied to the task it was trained for, which was due to VGG-M utilising a decreased stride and smaller filters in the first convolutional layer, in addition to using more filters in convolutional layers 2-5. However, this *may* have made the VGG-M network more task-specific, and therefore less generic to inputs that originate from very different distributions to the network trained images.

5.2. Transfer learning auxiliary representations

Table 3 also displays the results of our trained Male-5045 CNN representations, P_8 and FC_9, with the greyscale intensity image representation, GS. The pooling layer, P_8, is by far the best performer, followed by FC_9, with the greyscale intensity images, GS, being the worst. By per-

Table 2. GS indicates the greyscale intensity image and I is the preprocessed image input. $C(d, f, s, p)$ indicates a convolutional layer with d filters sized $f \times f$, with stride s and zero-padding p. $P(f', s', p')$ indicates a max-pooling layer with spatial size $f' \times f'$, with stride s' and zero-padding p'. $FC(n)$ indicates a fully-connected layer with n neurons.

NETWORK	ARCHITECTURE
VGG-F	GS-$I(224, 224, 3)$-$C_1(64, 11, 4, 0)$-$P_1(2, 2, 0)$-$C_2(256, 5, 1, 2)$-$P_2(2, 2, 0)$-$C_3(256, 3, 1, 1)$-$C_4(256, 3, 1, 1)$-$C_5(256, 3, 1, 1)$-$P_5(2, 2, 0)$-$FC_6(4096)$-$FC_7(4096)$-$FC_8(1000)$
VGG-M	GS-$I(224, 224, 3)$-$C_1(96, 7, 2, 0)$-$P_1(2, 2, 0)$-$C_2(256, 5, 2, 1)$-$P_2(2, 2, 0)$-$C_3(512, 3, 1, 1)$-$C_4(512, 3, 1, 1)$-$C_5(512, 3, 1, 1)$-$P_5(2, 2, 0)$-$FC_6(4096)$-$FC_7(4096)$-$FC_8(1000)$
MALE-5045	GS-$I(100, 100, 1)$-$C_1(32, 3, 1, 1)$-$C_2(64, 3, 1, 1)$-$P_2(2, 2, 0)$-$C_3(64, 3, 1, 1)$-$C_4(128, 3, 1, 1)$-$P_4(2, 2, 0)$-$C_5(96, 3, 1, 1)$-$C_6(256, 3, 1, 1)$-$P_6(2, 2, 1)$-$C_7(160, 3, 1, 1)$-$C_8(320, 3, 1, 1)$-$P_8(7, 1, 0)$-$FC_9(5045)$

forming fine-grained classification, the CNN representation space has been *stretched*, such that they have become sensitive to the important details in the face images. Consequently, this has created richer, and more abstract, representations of the male face images, ergo making anomalous female face images more perceptible. In addition, we analyse the effectiveness of transfer learnt pooling layers, P_5, from the networks VGG-F and VGG-M on this task. Again, we see that the pooling layer, P_5, of the VGG-F has superior performance over P_5 of VGG-M. Whilst neither offer groundbreaking AUROC scores, especially P_5 of VGG-M which is worse than random guessing, it is interesting to note the same occurrence as in tasks 1-6. That is, VGG-F appears to exhibit more generalisable representations for disparate image datasets. A detailed analysis on this phenomenon is beyond the scope of this paper and we leave it for future work.

6. Discussion

Our proposed methods for transferring representations to anomaly detection tasks: (i) transfer learning pre-trained representations, and (ii) transfer learning auxiliary representations by formulating a fine-grained classification task. Our results strongly demonstrate the utility of transfer learnt representations having made no prior assumptions on the generating distributions of either the normal or anomaly class. In the auxiliary task, we were able to better define the concept of normality by learning how to discriminate between the sub-categories of the normal class. These results clearly indicate that transfer learnt representations offer a good baseline in a diverse range of tasks, and we believe these results can be further improved with the use of handcrafted features, if domain expertise is available. Nevertheless, the results when transfer learning from pre-trained CNNs show that it is not a simple matter of choosing a pre-trained network that performed best at its original task. Further analysis into the reasons behind this phenomenon is required before one can draw any explicit conclusions.

Table 3. AUROC performance on the transfer learning representation tasks using CNNs VGG-F, VGG-M, and Male-5045, where **bold** indicates the best performing representation in a task.

VGG-F	MEAN ± STD AUROC OVER 3 TRIALS				
TASK	GS	P_5	FC_6	FC_7	FC_8
#1: MNIST-1	0.5809 ± 0.0436	$\mathbf{0.9382 \pm 0.0066}$	0.8494 ± 0.0182	0.8395 ± 0.0187	0.8604 ± 0.0104
#2: MNIST-2	0.5089 ± 0.0298	0.8833 ± 0.0138	0.8948 ± 0.0150	0.9074 ± 0.0141	$\mathbf{0.9239 \pm 0.0051}$
#3: MNIST-3	0.4321 ± 0.0189	$\mathbf{0.8500 \pm 0.0072}$	0.7207 ± 0.0148	0.7000 ± 0.0150	0.6652 ± 0.0217
#4: MNIST-4	0.4315 ± 0.0089	$\mathbf{0.7662 \pm 0.0120}$	0.7151 ± 0.0061	0.7357 ± 0.0072	0.7240 ± 0.0033
#5: X-RAY-1	0.9983 ± 0.0000	0.9988 ± 0.0001	$\mathbf{0.9992 \pm 0.0000}$	$\mathbf{0.9992 \pm 0.0000}$	0.9989 ± 0.0000
#6: X-RAY-2	0.0358 ± 0.0128	$\mathbf{0.9874 \pm 0.0028}$	0.9847 ± 0.0034	0.9842 ± 0.0047	0.9744 ± 0.0050
AVERAGE	0.4979 ± 0.0190	$\mathbf{0.9040 \pm 0.0071}$	0.8607 ± 0.0096	0.8610 ± 0.0100	0.8578 ± 0.0076
VGG-M	MEAN ± STD AUROC OVER 3 TRIALS				
TASK	GS	P_5	FC_6	FC_7	FC_8
#1: MNIST-1	0.5809 ± 0.0436	$\mathbf{0.8933 \pm 0.0148}$	0.7519 ± 0.0131	0.7904 ± 0.0131	0.7925 ± 0.0194
#2: MNIST-2	0.5089 ± 0.0298	$\mathbf{0.8543 \pm 0.0144}$	0.8179 ± 0.0016	0.8461 ± 0.0015	0.8313 ± 0.0043
#3: MNIST-3	0.4321 ± 0.0189	$\mathbf{0.8196 \pm 0.0224}$	0.5604 ± 0.0828	0.5704 ± 0.0881	0.5679 ± 0.0741
#4: MNIST-4	0.4315 ± 0.0089	$\mathbf{0.7593 \pm 0.0174}$	0.5110 ± 0.0546	0.5201 ± 0.0486	0.4911 ± 0.0449
#5: X-RAY-1	$\mathbf{0.9983 \pm 0.0000}$	0.9977 ± 0.0001	0.9967 ± 0.0000	0.9978 ± 0.0001	0.9977 ± 0.0000
#6: X-RAY-2	0.0358 ± 0.0128	0.8099 ± 0.0557	0.9342 ± 0.0031	$\mathbf{0.9455 \pm 0.0027}$	0.9225 ± 0.0027
AVERAGE	0.4979 ± 0.0190	$\mathbf{0.8557 \pm 0.0208}$	0.7620 ± 0.0259	0.7784 ± 0.0257	0.7672 ± 0.0243
MALE-5045	MEAN ± STD AUROC OVER 3 TRIALS				
TASK	GS	P_5 (VGG-F)	P_5 (VGG-M)	P_8	FC_9
#7: CASIA	0.4404 ± 0.0033	0.5456 ± 0.0014	0.4961 ± 0.0067	$\mathbf{0.7849 \pm 0.0032}$	0.7166 ± 0.0964

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under CASE Award Grant 157760 and Rapiscan Systems Ltd. In addition, Jerone T. A. Andrews thanks Emma F. Shapiro for her invaluable support and assistance.

References

Ahmed, A., Yu, K., Xu, W., Gong, Y., and Xing, E. Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks. In *Computer Vision–ECCV 2008*, pp. 69–82. Springer, 2008.

Andrews, J. T. A., Morton, E. J., and Griffin, L. D. Detecting anomalous data using auto-encoders. *International Journal of Machine Learning and Computing*, 6(1):21, 2016.

Chang, C. and Lin, C. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. In *British Machine Vision Conference*, 2014.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Dheeba, J, Singh, N. A., and Selvi, S. T. Computer-aided detection of breast cancer on mammograms: A swarm intelligence optimized wavelet neural network approach. *Journal of biomedical informatics*, 49:45–52, 2014.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531*, 2013.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. In *Advances in Neural Information Processing Systems*, pp. 766–774, 2014.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1717–1724, 2014.

Ranney, K. I. and Soumekh, M. Hyperspectral anomaly detection within the signal subspace. *Geoscience and Remote Sensing Letters, IEEE*, 3(3):312–316, 2006.

Razavian, A., Azizpour, H., Sullivan, J., and Carlsson, S. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813, 2014.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Yann. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

Stumpf, R. P., Culver, M. E., Tester, P. A., Tomlinson, M., Kirkpatrick, G. J., Pederson, B. A., Truby, E., Ransibrahmanakul, V., and Soracco, M. Monitoring karenia brevis blooms in the gulf of mexico using satellite ocean color imagery and other data. *Harmful Algae*, 2(2):147–160, 2003.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *Computer vision–ECCV 2014*, pp. 818–833. Springer, 2014.