2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Combining Image Features and Patient Metadata to Enhance Transfer Learning

Spencer A. Thomas[1]

Abstract— **In this work, we compare the performance of six state-of-the-art deep neural networks in classification tasks when using only image features, to when these are combined with patient metadata. We utilise transfer learning from networks pretrained on ImageNet to extract image features from the ISIC HAM10000 dataset prior to classification. Using several classification performance metrics, we evaluate the effects of including metadata with the image features. Furthermore, we repeat our experiments with data augmentation. Our results show an overall enhancement in performance of each network as assessed by all metrics, only noting degradation in a vgg16 architecture. Our results indicate that this performance enhancement may be a general property of deep networks and should be explored in other areas. Moreover, these improvements come at a negligible additional cost in computation time, and therefore are a practical method for other applications.**

I. Introduction

Deep learning has emerged as a powerful suite of tools for image classification [1], and has a huge potential to solve challenges in healthcare settings. The use of deep neural networks is successful at tasks such as classification of medical images [2], analysis of electronic health records [3]–[5] and segmenting data from emerging medical technologies [6], [7]. This enormous potential comes with the caveat that very large amounts of data are required to train robust models that generalise beyond the training set. This requirement is unfortunately difficult to satisfy in the majority of biological and medical studies due to barriers to data availability.

Transfer learning has emerged as a promising method for circumventing the need for vast amounts of data to train deep networks [8]. For domains with limited data, transfer learning utilises networks pre-trained on similar tasks with large amounts of data [9]. Transfer learning is often used in medical imaging [2], [10], [11] due to the limited availability of data that require expert labeling [12]. Transferring the image features from one domain to another can at least match the performance of models trained directly on the new domain [13]. However the configuration of the transfer can be performed in a number of ways [12], [14] and more research is needed in this area.

Medical imaging data often has associated metadata used by clinicians in patient assessments. These metadata are multi type (numeric, categorical, etc) and are essential for maintaining the value of archived data [15]. The information may be content related, e.g. scanner parameters, or relevant extracts from computerised medical records (CMR). These resources contain rich information relating to diseases [16], [17], and data driven methods can identify patterns of patients [3], [4].

Classification tasks based on the combination of imaging with genomics data has been shown to surpass clinical experts in digital pathology [18]. Combining relevant information about the sample, e.g. patient demographics, with imaging data has also yielded high accuracy scores in binary classification tasks [19]. However, the effect of combining these data is unknown and an assessment of any improvements or degradation to the networks in these frameworks is needed.

Clinicians will typically base diagnosis on several information sources either implicitly or explicitly. Demographic factors such as age can influence the likelihood of disease prevalence. In this work we investigate the combination of imaging data with related metadata to enhance classification performance evaluated by several metrics. We utilise transfer learning due to the limited volumes of data available, comparing the performance with and without metadata. Additionally we repeat the experiments with and without data augmentation during the training of the model.

II. Methods

A large collection of digital skin images from the International Skin Imaging Collaboration (ISIC) Melanoma Project [20] have been collated, processed and classified by expert dermatologists. The HAM1000 dataset from the ISIC database contains 10,015 digital images of skin lesions, each belonging to one of eight classes of skin conditions. Additionally the images have associated metadata containing clinical and acquisition information. The clinical fields contain a small amount of patient information including diagnosis of the images, an example is shown in Table I. Specifically, these are, age (numerical), sex (categorical) and anatomical site of the lesion (text).

TABLE I
Image Metadata with ISIC Images

Clinical Field	Example Entry
age approx	55
sex	female
anatom site general	lower extremity
melanocytic	true
benign malignant	malignant
diagnosis	melanoma
diagnosis confirm type	histopathology

[1]Spencer A. Thomas is with the Data Science group, National Physical Laboratory, Teddington, UK spencer.thomas@npl.co.uk

A. Deep Image Features

Consider the input data as $X \in \mathbb{R}^{N \times D}$ where X_i is a D dimensional data point with N instances of the data. For imaging analysis X is the imaging data with D pixels and N images. Deep learning takes X as an input and applies a series of transformations through hidden layers typically in the form of convolutions. Following the notation of [21], a matrix $W^k \in \mathbb{R}^{d_{k-1} \times d_k}$ is used to linearly transform the output of the $(k-1)$th layer, $X_{k-1} \in \mathbb{R}^{N \times d_{k-1}}$, into a d_k - dimensional space, $X_{k-1}W^k \in \mathbb{R}^{N \times d_k}$, at the kth layer. The linear transformations are followed by a non-linear function, $\sigma_k(z)$, at each layer. The output of a network with K layers is given by

$$\mathcal{F}(X) = \sigma_K\left(\ldots \sigma_2\left(\sigma_1\left(XW^1\right)W^2\right)\ldots W^K\right). \quad (1)$$

$\mathcal{F}(X) \in \mathbb{R}^{N \times d_K}$, where d_K is the dimensionality of $\mathcal{F}(X)$. For each network we select K such that $\mathcal{F}(X)$ corresponds to the deepest set of image features, typically with the lowest dimensionality.

We compare several state-of-the-art deep convolutional neural network architectures for obtaining $\mathcal{F}$. All the networks used here have been pretrained using the ImageNet [22] dataset, and the network weights transferred to the ISIC image dataset. In this configuration, we are using the networks as feature extractors. Specifically we evaluate alexnet [23], densenet201 [24], resnet50 [25], inception-resnetv2 [26], vgg16 [27] and googlenet [28] each with and without augmentation added to the input images. To account for the difference in input size to each network, all images are resized to the required dimensions using bi-linear interpolation.

For the augmentation experiments, we introduce a subset of image manipulations, $X' = \Omega(X)$, where Ω represents the augmentation to the image prior to passing it to the network. The augmentation function introduces a random shift in the image of up to 30 pixels from its origin along the X axis and separately along the Y axis, random reflections in X and/or Y, and random rotations up to 90 degrees. This transformation is applied to the training and testing data.

B. Integrating Images and Metadata

The metadata for the images, M are mapped such that they contain only numerical values to be compatible with standard neural networks. The mapping function $\mathcal{G}(M)$ converts the data to ASCII decimal introduced in [3]. The conversion is performed element wise for an input string to allow maximum flexibility, for example, distinguishing upper and lower case letters, and mixed numerical and text inputs. When the input data differ in length, all instances are padded with trailing white space to the same size as the largest input string prior to conversion. Any missing entries in the fields are recorded as not a number (zero in ASCII decimal).

The metadata fields are integrated with the image data by concatenating the image features obtained by the CNN at its deepest layer prior to classification, $\mathcal{F}(X)$ (blue vector in Fig. 1), with the encoded metadata inputs, $\mathcal{G}(M)$ (red vector in Fig. 1).

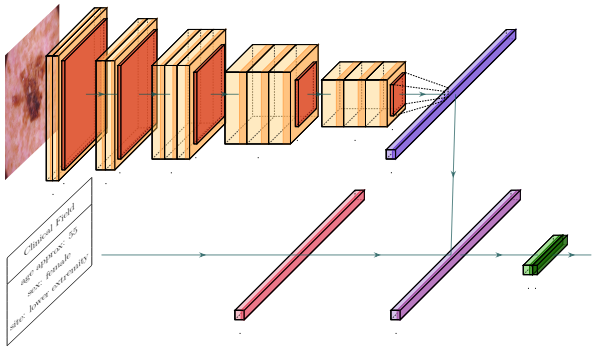

Fig. 1. Combination of imaging and non-imaging data in deep networks. A series of convolution and pooling operations (orange) yield a lower dimensional feature vector (blue) for image data. The non-imaging data are encoded numerically by mapping to ASCII decimal [3] providing a metadata feature vector (red). The imaging and non-imaging feature vectors are concatenated (purple) and used as input for a softmax classifier (green).

$$\mathcal{H} = \begin{pmatrix}\mathcal{F}(X) & \mathcal{G}(M)\end{pmatrix} = \begin{pmatrix}\mathcal{F}_{1,1} & \cdots & \mathcal{F}_{1,d_K} & \mathcal{G}_{1,1} & \cdots & \mathcal{G}_{1,d_{K'}} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{F}_{N,1} & \cdots & \mathcal{F}_{N,d_K} & \mathcal{G}_{N,1} & \cdots & \mathcal{G}_{N,d_{K'}}\end{pmatrix}, \quad (2)$$

where N is the number of images, d_K is the dimensionality of the output of the neural network, $\mathcal{F}(X)$, and $d_{K'}$ is the dimensionality of the converted metadata, $\mathcal{G}(M)$.

C. Classification

In all cases we use a softmax function to build a classification model for K classes,

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_j^K e^{z_j}}. \quad (3)$$

This classification model is trained using gradient descent for a maximum of 2000 epochs or when the gradient falls below 10^{-6}. In this work we compare the performance of the transfer learning based classification of the ISIC image data, to the performance of transfer learning when images are combined with their associated metadata. In the former case, we extract the image features, $\mathcal{F}(X)$, from each network pretrained on ImageNet, which are then passed to the softmax function to classify the images. In the latter case, we combine $\mathcal{F}(X)$ and $\mathcal{G}(M)$ as in Eq. (2), and pass $\mathcal{H}$ to the softmax classifier. In all experiments the data are split into 70:30 training:testing sets that are fixed for all networks for comparability of results.

We evaluate the performance of our classification models via several metrics. Specifically we evaluate the accuracy, specificity, sensitivity, precision, F-measure, informedness, markendness and Matthews correlation coefficient (MCC). The definitions of these are taken from [6] and omitted here for brevity.

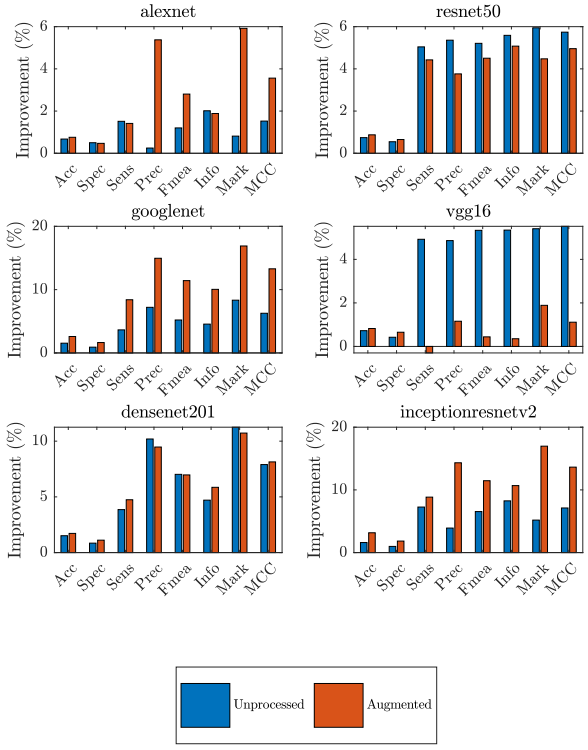

Fig. 2. Improvement in macro average performance of transfer learning in deep neural networks when using image metadata. Values are the difference in performance scores with positive values demonstrating improved performance when using metadata with image features. For example scores of 70% (image only) and 80% (combined image and metadata) would be plotted as 10%.

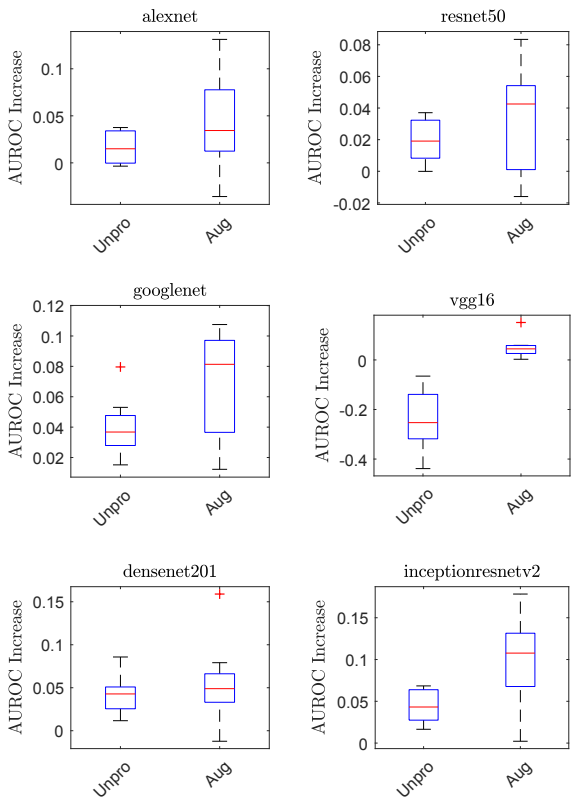

Fig. 3. Box plots of the class-wise AUROC improvement due to the inclusion of metadata. Values are the difference in AUROC between models that combine image features with metadata and those based only on images features. Positive vales represent enhanced performance and negative values indicate model degradation. Results from both unprocessed (Unpro) and augmented images (Aug) are presented. Note the AUROC ranges from 0 to 1.

III. Results

We report the macro average (mean class) performance in order to concisely summarise the findings of our experiments. In all of our experiments, we find that combining metadata with the image features improves classification performance for all networks compared to classifying using only image features. This enhancement is observed in all metrics indicating this may be a general characteristic of deep networks. This is clearly illustrated in Fig. 2 where positive values indicate an improvement when including metadata. The only degradation observed was in the sensitivity of a vgg16 network when using data augmentations. However, this decrease is small and this network exhibits relatively low improvements with augmented data compared to the other networks in this work. Improvements in accuracy and specificity are relatively small in all cases, though substantial improvements in the other metrics are seen in all networks. Specifically, googlenet, densenet201 and inceptionresnetv2 show improvements of more than 10 percentage points, meaning a score of 0.7 when using only image data increases to ≥ 0.8, a significant improvement.

To further evaluate the effects of combining the image metadata with the image features we also consider the area under the receiver operator characteristic curve (AUROC). For each network and experimental set up we perform a class wise ROC analysis, yielding eight receiver operator curves and corresponding AUROCs for each network. We subtract the AUROC for the image data alone from the AUROC when combining the image features and metadata. This class-level measure of improvement or degradation is represented as boxplots presented in Fig. 3. There is an overall increase in AUROC in all cases except the unprocessed images when using a vgg16 network which shows a considerable degradation. When using augmented images vgg16 shows an enhancement in line with the other networks.

It is worth noting that these improvements come at a negligible cost as seen in Table II. The training time for the softmax classifier when using the combined data is comparable to when using the image features alone. Moreover, this is insignificant compared to the feature extraction time in all networks, smaller by up to two orders of magnitude. The low time cost makes this a practical extension of current methods where metadata are available.

IV. Conclusion

Adding metadata to image features enhances classification overall. These improvements are noted in six different deep convolutional neural networks, as assessed by several perfor-

TABLE II

NETWORK SUMMARY OF RUNTIMES.

Network	d_K	Extraction (s)	$\mathcal{F}$ (s)	$\mathcal{H}$ (s)
alexnet	4096	217; 238	24; 21	95; 95
resnet50	2048	2160; 2174	22; 22	66; 63
googlenet	1024	780; 800	52; 53	52; 50
vgg16	4096	2365; 2323	17; 27	90; 92
densenet201	1920	6435; 6362	65; 69	67; 63
inceptionresnetv2	1536	5750; 5728	34; 38	66; 64

Extraction is the time to obtain the d_K dimensional features from the network processing over 48 CPU cores. Training times refer to time to train the softmax classifier based on either input features from $\mathcal{F}$ or $\mathcal{H}$. Times for the unprocessed (left) and augmented (right) data are provided respectively for each case.

mance metrics. Moderate to large enhancements are observed in all networks, with degradation only noted in a vgg16 architecture. Our results indicate that this may be a general property in classification of images with deep neural networks, though more work is required. These improvements come at a negligible additional cost in computation time, and therefore are a practical method for other applications.

ACKNOWLEDGMENT

The author would like to thank Nadia Smith and Peter Harris (NPL) for valuable feedback on this work. This work was funded by the Department of Business, Engineering and Industrial Strategy through the cross-theme national measurement strategy (Digital Health, 121572).

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," vol. 42, 2017.

[3] S. A. Thomas, N. Smith, V. Livina, I. Yonova, R. Webb, and S. de Lusignan, "Analysis of primary care computerized medical records (cmr) data with deep autoencoders (dae)," *Front. Appl. Math. Stat.*, vol. 5, pp. 1–12, 2019.

[4] S. de Lusignan, N. Smith, V. Livina, I. Yonova, R. Webb, and S. A. Thomas, "Analysis of primary care computerised medical records with deep learning," in *Studies in Health Technology and Informatics: ICT for Health Science Research*, vol. 258, 2019, pp. 249–250.

[5] A. Avati, K. Jung, S. Harman, L. Downing, A. Ng, and N. H. Shah, "Improving palliative care with deep learning," vol. 18, p. 122, 2018.

[6] S. A. Thomas, Y. Jin, J. Bunch, and I. S. Gilmore, "Enhancing classification of mass spectrometry imaging data with deep neural networks," in *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–8.

[7] J. Behrmann, C. Etmann, T. Boskamp, R. Casadonte, J. Kriegsmann, and P. Maass, "Deep learning for tumor classification in imaging mass spectrometry," vol. 34, no. 7, pp. 1215–1223, 2018.

[8] P. Lakhani, D. L. Gray, C. R. Pett, P. Nagy, and G. Shih, "Hello world deep learning in medical imaging," *J Digit Imaging*, vol. 31, no. 283, pp. 283–289, 2018.

[9] S. J. Pan, Q. Yang, W. Fan, and S. J. Pan, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, 2010.

[10] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinformatics*, vol. 18, 2017.

[11] H. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, May 2016.

[12] T. Rai, A. Morisi, B. Bacci, N. J. Bacon, S. A. Thomas, R. M. L. Ragione, M. Bober, and K. Wells, "Can imagenet feature maps be applied to small histopathological datasets for the classification of breast cancer metastatic tissue in whole slide images?" in *Medical Imaging 2019: Digital Pathology*, vol. 10956, 2019.

[13] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?" *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, arXiv: 1706.00712.

[14] T. Rai, A. Morisi, B. Bacci, N. J. Bacon, S. A. Thomas, R. M. L. Ragione, M. Bober, and K. Wells, "An investigation of aggregated transfer learning for classification in digital pathology," in *Medical Imaging 2019: Digital Pathology*, vol. 10956, 2019.

[15] N. A. S. Smith, D. Sinden, S. A. Thomas, M. Romanchikova, J. E. Talbott, and M. Adeogun, "Building confidence in digital health through metrology," vol. 93, no. 1109, p. 20190574, 2020, publisher: The British Institute of Radiology. [Online]. Available: https://www.birpublications.org/doi/abs/10.1259/bjr.20190574

[16] S. de Lusignan, A. Correa, G. E. Smith, I. Yonova, R. Pebody, F. Ferreira, A. J. Elliot, and D. Fleming, "RCGP research and surveillance centre: 50 years' surveillance of influenza, infections, and respiratory conditions," vol. 67, no. 663, pp. 440–441, 2017.

[17] A. Correa, W. Hinton, A. McGovern, J. van Vlymen, I. Yonova, S. Jones, and S. de Lusignan, "Royal college of general practitioners research and surveillance centre (RCGP RSC) sentinel network: a cohort profile," vol. 6, no. 4, 2016, publisher: British Medical Journal Publishing Group.

[18] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. D. Cooper, "Predicting cancer outcomes from histology and genomics using convolutional networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, 2018. [Online]. Available: https://www.pnas.org/content/115/13/E2970

[19] E. Rocheteau and D. Kim, "Deep transfer learning for automated diagnosis of skin lesions from photographs," vol. 2011.04475 [cs, eess], 2020. [Online]. Available: http://arxiv.org/abs/2011.04475

[20] "International skin imaging collaboration: Melanoma project," https://www.isic-archive.com/#!/topWithHeader/tightContentTop/about/literature, accessed: 10/09/2019.

[21] R. Vidal, J. Bruna, R. Giryes, and S. Soatto, "Mathematics of deep learning," *ArXiv*, vol. abs/1712.04741, 2017.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," vol. 60, no. 6, pp. 84–90, 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3065386

[24] G. Huang, L. van der Maaten, Z. Liu, and K. Q. Weinberger, "Densely connected convolutional networks," *CoRR*, vol. abs/1608.06993, 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[26] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016. [Online]. Available: http://arxiv.org/abs/1602.07261

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, vol. abs/1409.1556, 2015. [Online]. Available: https://arxiv.org/pdf/1409.1556.pdf

[28] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 1–9. [Online]. Available: http://ieeexplore.ieee.org/document/7298594/