Combining Image Features and Patient Metadata to Enhance Transfer Learning

Spencer A. Thomas

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

<table>
<thead>
<tr>
<th>Clinical Field</th>
<th>Example Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>age approx</td>
<td>55</td>
</tr>
<tr>
<td>sex</td>
<td>female</td>
</tr>
<tr>
<td>anatom site general</td>
<td>lower extremity</td>
</tr>
<tr>
<td>melanocytic benign</td>
<td>malignant</td>
</tr>
<tr>
<td>malignant diagnosis</td>
<td>true</td>
</tr>
<tr>
<td>diagnosis confirm type</td>
<td>melanoma</td>
</tr>
<tr>
<td>histopathology</td>
<td></td>
</tr>
</tbody>
</table>

1Spencer 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 $k$th 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

$$F(X) = \sigma_K (\ldots \sigma_2 (\sigma_1 (XW^1) W^2) \ldots W^K). \quad (1)$$

$F(X) \in \mathbb{R}^{N \times d_K}$, where $d_K$ is the dimensionality of $F(X)$. For each network we select $K$ such that $F(X)$ 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 $F$. 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-v3 [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 $\Omega$ 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 $G(M)$ 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, $F(X)$ (blue vector in Fig. 1), with the encoded metadata inputs, $G(M)$ (red vector in Fig. 1).

\[
H = \begin{pmatrix}
F(X) \\
G(M)
\end{pmatrix}
= \begin{pmatrix}
F_{1,1} & \cdots & F_{1,d_K} \\
\vdots & \ddots & \vdots \\
F_{N,1} & \cdots & F_{N,d_K}
\end{pmatrix}
\begin{pmatrix}
G_{1,1} \\
\vdots \\
G_{N,1}
\end{pmatrix}, \quad (2)
\]

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

C. Classification

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

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

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, $F(X)$, from each network pretrained on ImageNet, which are then passed to the softmax function to classify the images. In the latter case, we combine $F(X)$ and $G(M)$ as in Eq. (2), and pass $H$ 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, markedness and Matthews correlation coefficient (MCC).

The definitions of these are taken from [6] and omitted here for brevity.
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**

<table>
<thead>
<tr>
<th>Network</th>
<th>$d_K$ (s)</th>
<th>$F$ (s)</th>
<th>$H$ (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>alexnet</td>
<td>4096</td>
<td>217; 238</td>
<td>24; 21</td>
</tr>
<tr>
<td>resnet50</td>
<td>2048</td>
<td>2160; 2174</td>
<td>22; 22</td>
</tr>
<tr>
<td>googlenet</td>
<td>1024</td>
<td>780; 800</td>
<td>52; 53</td>
</tr>
<tr>
<td>vgg16</td>
<td>4096</td>
<td>2365; 2323</td>
<td>17; 27</td>
</tr>
<tr>
<td>densenet201</td>
<td>1920</td>
<td>6435; 6362</td>
<td>65; 69</td>
</tr>
<tr>
<td>inceptionresnetv2</td>
<td>1536</td>
<td>7570; 5728</td>
<td>34; 38</td>
</tr>
</tbody>
</table>

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 $F$ or $H$. Times for the unprocessed (left) and augmented (right) data are provided respectively for each case.

**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**


