ADDRESSING CLASS IMBALANCE FOR TRAINING A MULTI-TASK CLASSIFIER IN THE CONTEXT OF SILK HERITAGE

M. Dorozynski

Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany
dorozynski@ipi.uni-hannover.de

KEY WORDS: Deep learning, image classification, multi-task learning, class imbalance, incomplete labelling, silk heritage.

ABSTRACT:

Collecting knowledge in the form of databases consisting of images and descriptive texts that represent objects from past centuries is a fundamental part of preserving cultural heritage. In this context, images with known information about depicted artifacts can serve as a source of information for automated methods to complete existing collections. For instance, image classifiers can provide predictions for different object properties (tasks) to semantically enrich collections. A challenge in this context is to train such classifiers given the nature of existing data: Many images do not come along with a class label for all tasks (incomplete samples) and class distributions are commonly imbalanced. In this paper, these challenges are addressed by a multi-task training strategy for a classifier based on a convolutional neural network (*SilkNet*) that requires images with class labels for the tasks to be learned. The proposed approach can deal with incomplete training examples, while implicitly taking interdependencies between tasks into account. Extensions of the training approach with a focus on hard examples during training as well as the use of an auxiliary feature clustering are developed to counteract problems with class imbalance. Evaluation is conducted based on a dataset consisting of images of historical silk fabrics with labels for five tasks, i.e. silk properties. A comparison of different variants of the classifier shows that the extensions of the training approach significantly improve the classifier's performance; the average F1-score is up to 5.0% larger, where the largest improvements occur with underrepresented classes of a task (up to +14.3%).

1. INTRODUCTION

Preserving our cultural heritage for future generations and making it available to both historians and the wider public are important tasks. In this context, a key strategy is the *digitization* of collections of historical objects in the form of searchable databases with standardized annotations and, potentially, images. It was the goal of the EU H2020 project SILKNOW (`http://silknow.eu/`) to take one step in the direction of searchable databases for the preservation and better understanding of European cultural heritage related to silk. To make silk-related knowledge from the past accessible for future generations, a database related to silk fabrics was built by harvesting existing online collections (Alba Pagán et al., 2020). However, the information that is relevant for art historians or other users is not always readily available in digital online collections. Given the fact that a digital collection may contain tens or even hundreds of thousands of records representing artifacts, a manual input of this information is tedious, expensive and, consequently, often impossible. Thus, automated procedures have to be developed. For artifacts, such as silk fabrics, for which one or several images are available, relevant properties, such as the time or place of production, the material a fabric is made of, or the technique that was used for its production, can be predicted automatically from images of the artifacts. From a user's perspective, the present work is motivated by the need for a database containing historically relevant objects with standardized metadata as complete as possible.

For the automatic derivation of complete and standardized properties of artifacts, such as silk fabrics, images are exploited as an information source. Machine learning techniques allow training of an image classifier per property of interest (*semantic variable*) using labelled training images. After training, the classifier is able to predict missing class labels of unseen samples. Early works in the context of cultural heritage differentiate different painters of artworks utilizing a Support Vector Machine, e.g. (Blessing and Wen, 2010). Inspired by the huge successes of deep learning-based classification methods, supervised learning based on deep *convolutional neural networks (CNNs)* (Krizhevsky et al., 2012) are used in more recent works aiming to learn historically relevant information from images of artistic pictures, e.g. (Tan et al., 2016; Sur and Blaine, 2017). Instead of independently training one classifier for each variable in the context of *single-task learning (STL)*, interdependencies between the variables are exploited in *multi-task learning (MTL)* by combining several related (classification) tasks in the training procedure with the goal of an improved generalisation (Caruana, 1993). This is why MTL was also investigated in the domain of image classification with applications in cultural heritage preservation, e.g. (Strezoski and Worring, 2017; Garcia et al., 2020; Yang et al., 2022). However, standard multi-task classification frameworks require one reference label for every task to be learned during training for every training sample. The challenge that has to be faced in real-world data, such as cultural heritage collections, is that there may be many training samples for which annotations are unavailable for some of the target variables to be predicted. Such samples are referred to as *incomplete samples* in this paper. Additionally, the distribution of the available class labels of a variable is often imbalanced for real-world datasets, which constitutes a further challenge to supervised learning. It is a well-known problem that training using data with *imbalanced class distributions* results in a classifier that tends to predict classes that were represented in the training data rather well, whereas classes with only a few examples in the training data often cannot be distinguished from other classes (Krawczyk, 2016; Johnson and Khoshgoftaar, 2019; Sridhar and Kalaivani, 2021). It is of special interest to apply a classifier that can

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-1-W1-2023-175-2023 | © Author(s) 2023. CC BY 4.0 License.

distinguish the classes of all silk properties well such that added value is delivered for the user of a silk database thanks to the predictions. In general, there are works addressing class imbalance, e.g. (Chawla et al., 2002), and in particular, there are works addressing class imbalance in the context of training CNNs, e.g. (Pouyanfar et al., 2018). Nevertheless, only one work could be identified addressing class imbalance in the context of MTL and, in particular, allowing for training data that is only partly labelled: In (Dorozysnki et al., 2021), a focal multi-task training strategy for incomplete samples is proposed This seems to be the only work in the context of cultural heritage preservation except for (Dorozynski and Rottensteiner, 2022a); in (Dorozynski and Rottensteiner, 2022a), class imbalance is investigated in an STL scenario by exploiting an auxiliary feature clustering in training.

In this paper, class imbalance in the context of multi-task classification for cultural heritage applications is investigated. For this purpose, a CNN-based multi-task classifier will be trained, where the training strategy allows for both complete as well as incomplete training samples. In contrast to existing works, different strategies addressing class imbalance are combined resulting in a superior classification performance. Thus, the scientific contributions are the following:

- We propose a combination of feature space clustering and focal training to tackle problems with class imbalance.
- An existing feature space clustering approach in the context of STL is adapted to an MTL scenario allowing for incomplete training samples.
- Finally, comprehensive experiments are conducted to investigate the performance of the proposed combined approach compared to the performance of the individual approaches.

2. RELATED WORK

In general, image classification aims to assign a class label to an input image. In recent years, CNNs (Krizhevsky et al., 2012) have been to be superior in solving this task compared to classical machine learning techniques in case a sufficient number of training samples is available. However, in case the task to be learned is represented by a rather small dataset consisting of some ten thousand images, such as in the context of cultural heritage applications, determining all weights of a CNN by means of training on such a dataset might be challenging. This is, fine-tuning (Yosinski et al., 2014; Tajbakhsh et al., 2016) of networks trained on a larger dataset, e.g. ImageNet (Russakovsky et al., 2015), became a common strategy to overcome such limitations and particular in the context of predicting properties of artifacts, e.g. (Tan et al., 2016; Sur and Blaine, 2017). Instead of training individual classifiers for a set of classification tasks to be solved, i.e. one classifier per task, a single multi-task network can be trained to simultaneously learn all of the tasks. The fact that the joint training of related tasks can be beneficial in comparison to a separate training of the individual tasks was already stated in (Caruana, 1993), who introduced MTL for artificial neural networks and decision trees. The idea behind MTL is to take advantage of dependencies between the tasks to be learned with the goal of improved generalisation. Against this background, the joint training of classifiers for different tasks is addressed in different contexts, e.g. remote sensing, (Leiva-Murillo et al., 2013), human pose estimation, e.g. (Li et al., 2014), depth estimation and semantic segmentation, e.g. (Zhang et al., 2019), as well as cultural heritage preservation, e.g. (Strezoski and Worring, 2017; Dorozynski et al., 2019; Garcia et al., 2020). In this context, the approach in (Dorozynski et al., 2019) is the only one that allows for missing information in the training strategy, which is a requirement for MTL in the context of cultural heritage preservation. Moreover, no work could be identified that addresses class imbalance in the context of multi-task learning. Nevertheless, class imbalance occurs in almost all heritage-related classification tasks.

Learning from imbalanced data is a well-known problem in the fields of Photogrammetry and Computer Vision (Johnson and Khoshgoftaar, 2019; Sridhar and Kalaivani, 2021). Different strategies have been developed to address learning using data with imbalanced class distributions. While the training class distribution is artificially balanced in (Chawla et al., 2002; Ando and Huang, 2017), in (Lin et al., 2017; Khan et al., 2017) the training objectives are adapted such that classes with few training examples have a higher impact on the classifier's parameters. Dong et al. (2018) combine aspects of both strategies. According to (Krawczyk, 2016), class imbalance may be irrelevant if there are sufficiently good representations for both, frequent as well as less frequent classes. Using CNNs, both representations of images as well as the mapping to class scores, are learned. Thus, one way of achieving such a sufficient representation is to guide the CNN to learn that the feature vectors belonging to the same class should be close in feature space and that clusters corresponding to different classes should be further away from each other, e.g. (Huang et al., 2016; Cao et al., 2019; Dorozynski and Rottensteiner, 2022a). All clustering approaches in the works just mentioned force the distances between features belonging to the same class to be small and those belonging to different classes to be large. Thus, problems with class imbalance are partly mitigated (Huang et al., 2016; Cao et al., 2019; Dorozynski and Rottensteiner, 2022a). Nevertheless, class imbalance is nearly exclusively investigated in the context of STL. To the best of the knowledge of the author, multi-task multi-class image classification is exclusively addressed in (Dorozysnki et al., 2021), where a variant of the focal loss (Lin et al., 2017) for multi-task learning is proposed.

Applying image-based classification techniques to derive information about depicted artifacts is not new. Early works rely on support vector machines to predict properties of ancient paintings, e.g. (Blessing and Wen, 2010), whereas more recent approaches use CNNs to solve such classification tasks, e.g. (Tan et al., 2016; Sur and Blaine, 2017). Instead of training one CNN per task, e.g. predicting a painting's artist, genre or style (Tan et al., 2016), several related tasks can be learned by one CNN-based classifier in the context of multi-task learning (Strezoski and Worring, 2017; Garcia et al., 2020; Yang et al., 2022). The network architecture generally consists of a shared feature extraction network, e.g. a residual network (ResNet) (He et al., 2016) such as in (Sur and Blaine, 2017; Strezoski and Worring, 2017; Garcia et al., 2020; Yang et al., 2022), and a classification head with one branch per task. Similarly, CNNs have been applied in the context of image classification related to historic silk fabrics (Dorozysnki et al., 2021; Dorozynski and Rottensteiner, 2022a). In contrast to the other works dealing with cultural heritage classification, these two works seem to be the only ones that investigate class imbalance in this context; while in (Dorozynski and Rottensteiner, 2022a)

an auxiliary clustering is exploited to mitigate such problems in an STL scenario, focal multi-class multi-task learning is proposed in (Dorozysnki et al., 2021). However, problems with class imbalance are still only partly mitigated.

Accordingly, the approach in this work aims to take a further step to reduce problems with class imbalance in the context of cultural heritage-related classification. For that purpose, focal training (Lin et al., 2017; Dorozysnki et al., 2021) is combined with an auxiliary feature clustering, relying on constraints for the distance between features (Huang et al., 2016; Dorozynski and Rottensteiner, 2022a). In contrast to (Huang et al., 2016; Dorozynski and Rottensteiner, 2022a), being in the context of STL, the approach in the current work allows for MTL. This requires an expansion of the formulation of the auxiliary clustering loss. Just as the multi-task classification approach in (Dorozysnki et al., 2021), the proposed method can deal with both complete as well as incomplete samples. Thus, the classification method proposed in this work is aimed to be applicable to complex data in terms of incompleteness and class imbalance, while it is aimed to generalize well at the same time.

3. METHODOLOGY

The goal of the proposed MTL classification method is to automatically predict a class label per classification task based on images utilizing a single classifier. For that purpose, a CNN architecture based on a ResNet (He et al., 2016) is proposed that takes an RGB image of the size 224 x 224 pixels as an input and provides normalized class scores for each task. In the context of this work, a classification task is related to a property of an object depicted in the image, e.g. the production *time*, the manufacturing *technique*, the *material*, the *place* of origin and the subject depicted type, denoted as *depiction* of a silk fabric, but it could also be another property of another object type, such as the artist of a depicted ancient painting. The proposed MTL CNN architecture, denoted as *SilkNet*, is presented in section 3.1 and the proposed training strategy for determining optimal values for the network parameters is presented in section 3.2.

3.1 Network Architecture (*SilkNet*)

The main objective of the proposed CNN denoted as *SilkNet* is to allow for learning a classifier providing normalized class scores $y_{mk}(x)$ for the M classification tasks. For that purpose, the network architecture shown in Figure 1 is proposed. At training time, it consists of three main parts; a feature extraction part delivering features $f_{jfc}(x)$, a clustering head delivering features $f(x)$ for the auxiliary clustering, and a classification head. The latter one consists of M classification branches delivering normalized class scores $y_{mk}(x)$ that can be interpreted as posterior probabilities $P(C_{mk}|x)$ for the k^{th} class of the m^{th} semantic variable C_{mk}. At test time varies, only the classification head is active for image classification, while the clustering head is exclusively required in training to learn the auxiliary clustering.

First of all, the image x is mapped to a 2048-dimensional feature vector $f_{RN}(x)$ by means of a ResNet-152 backbone (He et al., 2016) with parameters $\mathbf{w}_{RN}$, followed by a Rectified Linear Unit (ReLU) activation (Nair and Hinton, 2010) and a dropout layer (Srivastava et al., 2014) with a dropout rate of 30%. Dropout is introduced to enable the network to learn a more general application-specific representation out of the features $f_{RN}(x)$

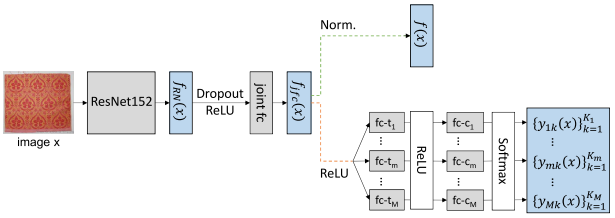

Figure 1. CNN architecture of *SilkNet*. An input image x is presented to the network. After a feature extraction network, a feature vector $f_{jfc}(x)$ is presented to both, a clustering head (connected with a green broken line) and a classification head (connected with an orange broken line). During training, both network heads are active, whereas at test time, exclusively the classification head is active. For details see section 3.1.

provided by the potentially fully pre-trained ResNet. Afterwards, $f_{RN}(x)$ is presented to a sub-network *joint fc* consisting of NL_{jfc} fully connected layers with $[NN^1_{jfc}, ..., NN^{NL_{jfc}}_{jfc}]$ nodes, respectively, resulting in a feature vector $f_{jfc}(x)$. The sub-network joint fc is parameterized by the weight vector $\mathbf{w}_{jfc}$. The layers *joint fc* are at the core of the proposed method, because the resulting feature vectors $f_{jfc}(x, \mathbf{w}_{RN}, \mathbf{w}_{jfc})$ are the input to both the clustering and classification heads. Thus, the weights $\mathbf{w}_{jfc}$ of the *joint fc* layers are both influenced by the multi-task classification loss as well as by the auxiliary clustering. The clustering head consists of a simple normalization of the feature vector $f_{jfc}(x, \mathbf{w}_{RN}, \mathbf{w}_{jfc})$ to unit length, leading to the feature vector $f(x, \mathbf{w}_{RN}, \mathbf{w}_{jfc})$. This vector will exclusively be used during training in the auxiliary clustering loss. The classification head consists of M separate branches, each corresponding to one classification task to be learned. Each branch is connected to the sub-network *joint fc* via $f_{jfc}(x)$, being processed by a ReLU activation function, and consists of NL_{tfc} task-specific fully connected layers of $[NN^1_{tfc}, ..., NN^{NL_{tfc}}_{tfc}]$ nodes, respectively, and each layer is followed by a ReLU activation. This network part is denoted by $fc\text{-}t_m$, where each task-specific branch is defined to have an identical number of layers and a corresponding number of nodes. Finally, each branch has a classification layer $fc\text{-}c_m$ with K_m nodes, where K_m is the number of classes to be distinguished for the m^{th} task, delivering unnormalized class scores $a_{mk}(x)$. It is parameterized by the weights $\mathbf{w}_{class}$ and delivers the normalized class scores $y_{mk}\left(x, \mathbf{w}_{RN}, \mathbf{w}_{jfc}, \mathbf{w}_{class}\right) = y_{mk}\left(x, \mathbf{w}\right)$.

3.2 Network Training

Training is based on a set of training samples $\mathbf{x}$ that consist of images with semantic annotations (class labels) for at least one of the M variables; the proposed training strategy does not require any further information. The CNN *SilkNet* depicted in Figure 1 is trained by minimizing a loss function $\mathcal{L}\left(\mathbf{x}, \mathbf{w}\right)$ based on such a set $\mathbf{x}$. *SilkNet* has two sets of parameters from the perspective of training: the weights $\mathbf{w}_{RN}$ of the ResNet-152 and the remaining weights $\mathbf{w}_{head} := [\mathbf{w}^T_{jfc}, \mathbf{w}^T_{class}]^T$ of the additional layers. The weights $\mathbf{w}_{RN}$ are initialized by pre-trained weights obtained on the ILSVRC-2012-CLS dataset (Russakovsky et al., 2015) (ImageNet), whereas the weights $\mathbf{w}_{head}$ of the additional layers of the CNN are initialized randomly using variance scaling (He et al., 2015). As it is expected that silk fabrics or other objects in the context of cultural heritage belong to another domain than objects depicted

in the ImageNet dataset, the last NB_{RN} residual blocks are potentially fine-tuned (Yosinski et al., 2014). Denoting the parameters of the frozen ResNet layers by $\mathbf{w}_{RN_{fr}}$ and those of the fine-tuned ResNet layers by $\mathbf{w}_{RN_{ft}}$, the parameters to be determined in training are $\mathbf{w}_{tr} = [\mathbf{w}_{RN_{ft}}^T, \mathbf{w}_{head}^T]^T$. Note that the entire parameter vector $\mathbf{w}$ can thus also be represented by $\mathbf{w} = [\mathbf{w}_{RN_{fr}}^T, \mathbf{w}_{tr}^T]^T = [\mathbf{w}_{RN_{fr}}^T, \mathbf{w}_{RN_{ft}}^T, \mathbf{w}_{jfc}^T, \mathbf{w}_{class}^T]^T$. During training, the respective loss function is minimized using mini-batch stochastic gradient descent with adaptive moments, i.e. Adam (Kingma and Ba, 2015). In each training iteration, only a mini-batch $\mathbf{x}^{MB} \subset \mathbf{x}$ consisting of N^{MB} training samples is considered, and only the loss $\mathcal{L}\left(\mathbf{x}^{MB}, \mathbf{w}\right)$ achieved for the current mini-batch is used to update the parameters $\mathbf{w}_{tr}$.

The proposed loss function $\mathcal{L}$ consists of a classification loss $\mathcal{L}_C$ weighted by $\lambda_C \in [0, 1]$, an auxiliary clustering loss $\mathcal{L}_{aux}$ weighted by $\lambda_{aux} \in [0, 1]$:

$$\begin{aligned}\mathcal{L}\left(\mathbf{x}^{MB}, \mathbf{w}\right) =& \lambda_C \cdot \mathcal{L}_C\left(\mathbf{x}^{MB}, \mathbf{w}\right) \\ &+ \lambda_{aux} \cdot \mathcal{L}_{aux}\left(\mathbf{x}^{MB}, \mathbf{w}_{RN}, \mathbf{w}_{jfc}\right).\end{aligned} \tag{1}$$

In order to mitigate problems with underrepresented classes, the classification loss is decided to be a variant of the focal loss (Lin et al., 2017). To avoid restricting the training method to complete samples, which would drastically reduce the set of training samples and might also reduce the set of tasks to be learned in an MTL scenario (the more tasks are considered the smaller tends to be the set of complete samples), the proposed method should handle both complete and incomplete samples. Thus, the focal multi-task multi-class classification loss proposed in (Dorozysnki et al., 2021) is selected, i.e.

$$\mathcal{L}_C(\mathbf{x}^{MB}, \mathbf{w}) = -\frac{1}{N_M} \sum_{i=1}^{N^{MB}} \sum_{m \in \mathcal{M}_i^{av}} \sum_{k=1}^{K_m} \omega_{mk}^{fo} \cdot t_{imk} \cdot l_{imk} \tag{2}$$

with

$$\omega_{mk}^{fo} = \left(1 - y_{mk}\left(x_i, \mathbf{w}\right)\right)^{\gamma} \tag{3}$$

and

$$l_{imk} = ln\left(y_{mk}\left(x_i, \mathbf{w}\right)\right). \tag{4}$$

In equation 2, the loss is calculated for all N^{MB} samples in the mini-batch, where for each sample x_i the loss is calculated for the K_m classes of a task m for which the i^{th} sample has a known class label, i.e. m is in the set of tasks with an available label $\mathcal{M}_i^{av}$. The the focal weight ω_{mk}^{fo} (eq. 3) with the focusing parameter γ controls the impact of the actual loss term l_{imk} (eq. 4) on the total loss $\mathcal{L}_C$. Thus, the network weights $\mathbf{w}_{tr}$ are influenced more strongly by "hard" training examples that are assumed to be related to samples of underrepresented classes. Hard samples are indicated by smaller values of the softmax activation $y_{mk}\left(x_i, \mathbf{w}\right)$ for the correct class ($t_{imk} = 1$). For all other classes, the binary indicator variable t_{imk} is zero. The total loss $\mathcal{L}_C$ is normalized by the total number of available annotations for all M variables $N_M := \sum_{i=1}^{N^{MB}} \sum_{m \in \mathcal{M}_i^{av}} \sum_{k=1}^{K_m} t_{imk}$, i.e. the number of non-zero loss terms constituting $\mathcal{L}_C$.

The auxiliary clustering loss $\mathcal{L}_{aux}$ consists of three loss terms, each referring to a concept of similarity. Learning these concepts of similarity is supposed to improve learning image features to form clusters such that each cluster belongs to a distinct class or class combination in the regarded case of MTL. For that purpose, similar to the classification approach in (Dorozynski and Rottensteiner, 2022a), the concepts of similarity and the corresponding loss terms proposed in (Dorozynski and Rottensteiner, 2022b) in the context of image retrieval are adapted as auxiliary losses for training a classifier. In contrast to (Dorozynski and Rottensteiner, 2022a), being in the context of STL, the approach in this work is in the context of MTL. $\mathcal{L}_{aux}$ considers semantic similarity in $\mathcal{L}_{sem}$ weighted by $\alpha_{sem} \in [0, 1]$, colour similarity in $\mathcal{L}_{co}$ weighted by $\alpha_{co} \in [0, 1]$ and self-similarity in $\mathcal{L}_{slf}$ weighted by $\alpha_{slf} \in [0, 1]$, leading to the loss

$$\begin{aligned}\mathcal{L}_{aux}\left(\mathbf{x}^{MB}, \mathbf{w}\right) =& \alpha_{sem} \cdot \mathcal{L}_{sem}\left(\mathbf{t}^{MB}, \mathbf{w}\right) \\ &+ \alpha_{co} \cdot \mathcal{L}_{co}\left(\mathbf{p}_{co}^{MB}, \mathbf{w}\right) \\ &+ \alpha_{slf} \cdot \mathcal{L}_{slf}\left(\mathbf{p}_{slf}^{MB}, \mathbf{w}\right).\end{aligned} \tag{5}$$

The loss for learning semantic similarity is a variant of the triplet loss (Schroff et al., 2015), but it considers multiple semantic variables instead of a single binary similarity aspect, while allowing for missing labels for some of the variables:

$$\mathcal{L}_{sem}(\mathbf{t}^{MB}, \mathbf{w}) = \frac{1}{N_t^{MB}} \cdot \sum_{n_t=1}^{N_t^{MB}} max\left(M_{i,p,n}^{n_t} + \Delta_{i,p,\mathbf{w}}^{n_t} - \Delta_{i,n,\mathbf{w}}^{n_t}, 0\right),$$

$$M_{i,p,n}^{n_t} = Y_{sem}^{i,p,n^t} - \left(Y_{sem}^{i,n,n^t} + u^{i,n,n^t}\right). \tag{6}$$

The semantic similarity loss is calculated for all triplets $t := (x_i, x_p, x_n) \in \mathbf{t}^{MB}$ for which the positive sample x_p is semantically more similar to the anchor sample x_i than the negative sample x_n, which is fulfilled in case the margin $M_{i,p,n}^{n_t}$ is larger than zero. For such triplets, the feature distance $\Delta_{i,p,\mathbf{w}}^{n_t}$ between x_i, x_p is forced to be at least by $M_{i,p,n}$ larger than the feature distance $\Delta_{i,n,\mathbf{w}}^{n_t}$ between x_i, x_n as shown in Figure 2. In equation 6, $Y_{sem}^{i,o}$ denotes the relation of known identical class labels for x_i, x_o to the number of semantic variables M the class labels of which are compared. Thus, $M_{i,p,n}$ is the guaranteed difference in semantic similarity Y_{sem}^{i,p,n^t} between $x_i^{n_t}, x_p^{n_t}$ and the similarity Y_{sem}^{i,n,n^t} between $x_i^{n_t}, x_n^{n_t}$ under consideration of the percentage of tasks for which the class labels cannot be compared for $x_i^{n_t}, x_n^{n_t}$ due to missing information $u(x_i, x_n)$.

Colour similarity is determined based on the agreement between the colour distributions of two images x_i, x_o according to the normalized cross correlation coefficient $\rho(x_i, x_o)$ of colour feature vectors $h(x_i)$ and $h(x_o)$. The colour feature vector $h(x_q)$ of an image x_q describes the colour distribution of that image in the *HSV* (*H*: hue, *S*: saturation, *V*: value) colour space. The feature vector is built by calculating polar coordinates (x^c, y^c) out of the hue H and saturation S values of every pixel of the image x_q resized to 224 x 224 pixels, afterwards defining a discrete grid in (x^c, y^c) space and counting the number of points per grid cell; concatenating the rows of the grid delivers $h(x_q)$. The colour similarity loss aims to learn descriptors $f(x_i), f(x_o)$ whose Euclidean distance reflects the colour similarity $\rho(x_i, x_o)$ of the image pair (x_i, x_o) defined, but in an inverse way. This is achieved by the colour similarity loss

$$\mathcal{L}_{co}(\mathbf{p}_{co}^{MB}, \mathbf{w}_v) = \frac{1}{N_{co}} \cdot \sum_{n_{co}=1}^{N_{co}} max\left(0, |\Delta_{i,o,\mathbf{w}_v}^{n_{co}} - M_{i,o}^{n_{co}}|\right),$$

$$M_{i,o}^{n_{co}} = \left(1 - \rho(x_i^{n_{co}}, x_o^{n_{co}})\right). \tag{7}$$

The loss $\mathcal{L}_{co}$ forces the feature distance $\Delta_{i,o,\mathbf{w}}^{n_{co}}$ of the normalized features to reflect the degree of colour similarity of the

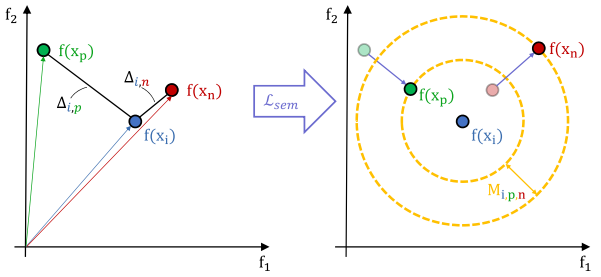

Figure 2. Principle of the semantic similarity loss $\mathcal{L}_{sem}(\mathbf{t}^{MB}, \mathbf{w})$ in eq. 3.2. In a two dimensional feature space with features f_1 and f_2, minimizing $\mathcal{L}_{sem}$ is equal to forcing (purple) *SilkNet*, parameterized by weights $\mathbf{w}$, to produce feature vectors $f(x)$ such that the difference in distance – Euclidean distance $\Delta_{i,p}$ between the feature vector $f(x_i)$ of an anchor sample x_i (blue) and the vector $f(x_p)$ of a positive sample x_p (green) and Euclidean distance $\Delta_{i,n}$ between the vector $f(x_i)$ and the feature vector $f(x_n)$ of a negative sample x_n (red) – is at least as large as a margin $M_{i,p,n}$ (yellow). $M_{i,p,n}$ is a function of the class labels available for a triplet $t := (x_i, x_p, x_n)$ and is equal to the guaranteed difference in semantic similarity between x_i, x_p ($Y^{i,p}_{sem}$) and x_i, x_n ($Y^{i,n}_{sem} + u^{i,n}$).

corresponding images $(x_i, x_o) =: p_{co} \in \mathbf{p}^{MB}_{co}$ considered in the colour margin $M^{n_{co}}_{i,o}$. Thus, $\Delta^{n_{co}}_{i,o,\mathbf{w}}$ is forced to be large for $\rho(x_i, x_o) \to -1$, i.e. negatively correlated colour distributions, and forced to be zero for $\rho(x_i, x_o) \to +1$, i.e. 100% correlated colour distributions.

Finally, the self-similarity loss $\mathcal{L}_{slf}$ aims to learn that features of images showing the same object should be similar and thus, to learn features that are invariant to geometrical and radiometrical transformations to some degree. For that purpose, the feature distance $\Delta^{n_{slf}}_{i,i',\mathbf{w}}$ between two images $(x_i, x'_i) =: p_{slf} \in \mathbf{x}^{MB}_{slf}$ is forced to be zero, where the image x'_i is defined to be an augmentation of an image x_i in the mini-batch $\mathbf{x}^{MB}$. This leads to the following loss:

$$\mathcal{L}_{slf}(\mathbf{p}^{MB}_{slf}, \mathbf{w}) = \frac{1}{N^{MB}_{slf}} \cdot \sum_{n_{slf}=1}^{N^{MB}_{slf}} \Delta^{n_{slf}}_{i,i',\mathbf{w}}, \tag{8}$$

The realized augmentation of x_i to obtain x'_i consists of a rotation of 90°, horizontal and vertical flips, cropping by a random percentage $b_{crop} \in [0.7; 1]$, small random rotations $\omega \in [-5°; +5°]$ as well as a change of the hue $H \in [0; 1]$ by adding a random value delta $\Delta_H \in [-0.05; +0.05]$ and an adaptation of the saturation S by multiplying it by a random factor $\delta_S \in [0.9; 1.0]$. Finally, a random zero mean Gaussian noise with a standard deviation $\sigma_G = 0.1$ can be added.

4. DATASET

The use case considered in this paper is the automatic completion of missing information in a database of historic silk fabrics employing a classifier. Thus, the SILKNOW dataset[1] that was generated in the context of the EU-H2020 project SILKNOW with the goal to build and provide a platform containing information about the European silk heritage is selected for the experiments. The dataset consists of 38,873 silk records in a knowledge graph (Alba Pagán et al., 2020), each coming along with one or several images depicting a silk object as well as semantic annotations for the production *time*, the manufacturing *technique*, the *material*, the *place* of origin and the subject depicted type, denoted as *depiction*. Figure 3 shows an example of an image of a silk fabric with annotations. Restricting the dataset to images of plain silk textiles, i.e. excluding records representing for instance furniture or accessories, and requiring a known label for at least one of the $M = 5$ semantic variables just mentioned results in a dataset consisting of 48,912 images. The dataset is split into subsets of 60% of the samples to be used for training, 20% for validation and 20% for testing, where all images belonging to a single record, i.e. an identical silk object, are part of the same subset. A requirement on the subsets is that each subset contains at least one example for each class of each variable. Some classes are represented by too few examples to fulfil this requirement, i.e. they are represented by one or two examples only. Furthermore, the class labels of the five variables are (at least) partly dependent on each other so that it is not possible to split the data such that the requirement on the subsets is fulfilled for all variables simultaneously. Thus, images that come along with a class label exclusively for classes that, thus, have to be excluded are omitted also already omitted in the 48,912 images. In case an image among these 48,912 images belongs to one of the excluded classes, its label is set to *background* for that specific variable in order to differentiate between *unknown*, i.e. there is no information available, and a label that is different from the labels of interest even though it cannot be considered (background).

The class structures and class distributions of the five semantic variables, as well as the number of samples labelled as background per variable, are presented in Table 1. It can be seen that the number of available class labels for the foreground classes varies strongly between the individual variables; *Place* has the largest amount of known class labels with 73.1% of available semantic annotations, followed by *material* with 72.3%, *time* with 58.0% and *technique* with 32.7% of available class labels. *Depiction* has the lowest number of known labels of interest, with only 7.0% of the 48,912 images in the dataset coming along with a class label. Furthermore, in general, nearly all of the images (99.8%) have an unknown class label for at least one of the five variables, exemplifying the need for methods dealing with incomplete training samples.

In addition to the differences in the amount of labelled data available per variable, the class distributions of the individual variables in Table 1 have different characteristics. Besides the total number of classes K_m differentiated per variable, the number of underrepresented classes $|\mathcal{M}_m|$, defined to be the number of classes with a relative frequency $\zeta < 1/K_m$, varies. Fur-

Figure 3. Example for an image of a silk fabric. Annotations: *time*: 20^{th} c.; *technique*: damask; *material*: animal fibre; *place*: ES; *depiction*: unknown.

[1] https://doi.org/10.5281/zenodo.5743090

Variable	Class	NS	Class	NS
place	GB	7,998	RU	228
	FR	7,379	JM	191
	ES	4,708	CH	146
	IT	4,700	EG	117
	IN	2,353	AZ	115
	CN	1,399	MO	84
	IR	1,294	AT	81
	JP	1,097	PT	73
	BE	648	MA	63
	TR	593	BD	60
	DE	592	CA	52
	GR	479	AU	46
	NL	455	MM	46
	US	357	UZ	42
	PK	352	*bg place*	604
material	animal fibre	27,252	bg material	0
	metal thread	4,208		
	vegetal fibre	3,891		
time	19^{th} c.	9,975	16^{th} c.	1,829
	18^{th} c.	8,423	15^{th} c.	685
	20^{th} c.	4,012	13^{th} c.	43
	17^{th} c.	3,378	*bg time*	104
technique	embroidery	6,861	tabby	185
	velvet	3,051	printed	99
	damask	2,768	twill	67
	other techn.	2,526	cannele	65
	resist dyeing	355	*bg techn.*	44
depiction	flower	2352	text	129
	plant	336	animal	116
	geom. shape	202	fruit	95
	stripe	138	object	73
	bg depiction	56		

Table 1. Statistics of the distribution of samples for the SILKNOW dataset. *Variable*: name of the variable considered; *Class*: classes differentiated for each variables; *NS*: number of samples for a class.

Variable	K_m	$\lvert\mathcal{M}_m\rvert$	IR (eq. 9)
place	29	22	190.4
time	7	5	232.0
technique	9	5	105.6
material	3	2	7.0
depiction	8	7	32.2

Table 2. Statistics of the characteristics of the class distributions of the classes of interest for the SILKNOW dataset. K_m: total number of classes; $|\mathcal{M}_m|$: number of underrepresented classes; IR: imbalance ratio.

thermore, an impression of the imbalance of the class distribution can be achieved by means of the imbalance ratio (Ortigosa-Hernández et al., 2017)

$$IR(\zeta_m) = \frac{max_i \zeta_i}{min_j \zeta_j}. \tag{9}$$

IR describes the ratio between the relative frequency ζ of the most frequent class i and the least frequent class j. Table 2 shows the quantities for the class distributions in Table 1. The class distributions of the variables vary strongly with respect to their IR (eq. 9) values; the variable *material* has the lowest IR of 7.0, indicating that the most dominant class has seven times as many examples as the class with the lowest number of examples, while the variable *time* has the highest IR (232.0). Furthermore, the total number of classes K_m varies between 3 (*material*) and 29 (*place*), where the amount of underrepresented classes $|\mathcal{M}_m|$ varies between 55.6% for *technique* and 87.5% for *depiction*.

Name	Loss weights		Similarity setting		
	λ_C	λ_{aux}	α_{sem}	α_{co}	α_{slf}
MTL	1.0	0.0	0.0	0.0	0.0
MTL^{fo}	1.0	0.0	0.0	0.0	0.0
$MTL + R^{sem}$	1.0	1.0	1.0	0.0	0.0
$MTL + R^{all}$	1.0	1.0	0.5	0.5	0.5
$MTL^{fo} + R^{sem}$	1.0	1.0	1.0	0.0	0.0
$MTL^{fo} + R^{all}$	1.0	1.0	0.5	0.5	0.5

Table 3. Overview of the conducted classification experiments. *Name*: name of the experiment, where all denotes that all concepts of similarity are used in the clustering. In MTL, the focal weight ω_{mk}^{fo} is 0; *Loss weights*: values for the weights λ_C and λ_{aux} in the loss in equation 1; *Similarity setting*: values for the weights in the auxiliary loss in equation 5.

5. EXPERIMENTS

The multi-task classification method for training with imbalanced data in section 3 is evaluated on the basis of experiments using the dataset described in section 4. The experimental setup as well as the evaluation protocol are described in section 5.1 and the results are presented and discussed in section 5.2.

5.1 Experimental Setup and Evaluation Strategy

Training *SilkNet* is conducted using early stopping, i.e. the training procedure is terminated when the validation loss, denoting the loss produced on an independent validation set using the current network parametrization, is saturated. This is realized using hyperparameters that were identified in preliminary experiments, where optimal parameter values are selected based on the average F1-score achieved on the validation set. Thus, training of *SilkNet* is realized using standard parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\hat{\epsilon} = 1 \cdot 10^{-8}$) for Adam, where preliminary experiments showed that a learning rate of $1 \cdot 10^{-4}$ is optimal. Furthermore, fine-tuning of $NB_{RN} = 3$ and a *SilkNet* configuration with $NL_{jfc} = 1$ shared fully connected layer with $NN^1_{jfc} = 1024$ nodes and $NL_{tfc} = 1$ further task-specific fully connected layer with $NN^1_{tfc} = 128$ per task leads to the best classification performance.

The conducted experiments presented in this paper are listed in Table 3. In each experiment, *SilkNet* is trained on the training set using the identified hyperparameters and the parametrization of the loss $\mathcal{L}\left(\mathbf{x}^{MB}, \mathbf{w}\right)$ (eq. 1). The goal of the experiments is to get an impression of the different training strategies to mitigate problems with class imbalance. Thus, the proposed training approach, i.e. combining focal training and an auxiliary clustering ($MTL^{fo} + R^{sem}$, $MTL^{fo} + R^{all}$), is compared to multi-task training without any auxiliary feature clustering and a focal weight ω_{mk}^{fo} of 0 (MTL), i.e. equally weights softmax loss terms. Furthermore, the approach is compared to focal training (MTL^{fo}) as well as different variants of an auxiliary clustering ($MTL + R^{sem}$, $MTL + R^{all}$). The comparison is based on the F1-scores and overall accuracies achieved by the respective classifiers on the independent test set. In order to get an impression of the impact of the random components during training of a specific multi-task classifier on the quality metrics, training and the respective evaluation will be conducted five times. The reported quality metrics for an experiment are the ones achieved on average in the five independent runs of that experiment.

5.2 Results and Discussion

The results of the experiments are presented in Table 4. The average F1-scores vary between 28.6% (MTL) and 33.6% ($MTL^{fo} + R^{all}$) and the respective Overall Accuracys (OAs) are in the range between 63.9% (MTL) and 66.2% ($MTL + R^{all}$). The remarkably higher OAs compared to the F1-scores observed for all experiments indicate that techniques addressing class imbalance are required in the context of classifying ancient silk fabrics. Both, the F1-scores as well as the OAs, are all higher for focal training (MTL^{fo}) than the respective metrics obtained for the baseline image classification experiments MTL. Furthermore, auxiliary feature space clustering ($MTL{+}R^{sem}$, $MTL{+}R^{all}$) can improve the ability of a multi-task classifier (MTL) to correctly predict the class for a given image in terms of $F1$ and OA. It can be observed that the combination of focal training and an auxiliary feature space clustering ($MTL^{fo} + R^{all}$) leads to the highest average F1-score of 33.6%. That is, the combined training strategy ($MTL^{fo}{+}R^{all}$) achieves on average a 5.0% higher F1-score and a 1.9% higher OA than the baseline training strategy (MTL). While $F1$ is 1.0% higher for $MTL^{fo} + R^{all}$ compared to focal training (MTL^{fo}), it is 1.7% higher for $MTL^{fo} + R^{all}$ compared to training with an auxiliary clustering ($MTL + R^{all}$). The average F1-scores achieved per experiment already indicate that combining focal training with an auxiliary clustering is to be preferred over applying the two training strategies independently from each other. A detailed analysis of the variable-specific F1-scores is provided in the following.

Analysing the **variable-specific F1-scores averaged over all classes of a variable** (Table 5), a similar behaviour of the F1-scores can be observed: Four of five variables achieve the highest score for the combined training strategy ($MTL^{fo} + R^{all}$), whereas the improvements caused by applying focal training (MTL^{fo}) and an auxiliary clustering ($MTL + R^{sem}$, $MTL + R^{all}$) individually results in much smaller improvements. Comparing the impact of an auxiliary clustering in training ($MTL + R^{all}$ compared to MTL) to the impact of focal weights (MTL^{fo} compared to MTL) on the F1-scores, the clustering is more beneficial than focal training for *technique* (+2.1%) and *depiction* (+4.9%), whereas training with focal weights is to be preferred over a clustering for *place* (+4.2%), *time* (6.8%) and *material* (+8.8%). This might be caused by the low proportion of available training samples for *depiction*, i.e. only 7.0% of the images come along with a known class label for *depiction*. Accordingly, the auxiliary clustering is assumed to support the classifier, particularly in case of a low number of available training samples. This assumption is supported by the scores obtained for *technique* ($MTL + R^{all}$ and MTL^{fo}), having the second lowest proportion of training data (32.8%): *technique* is the second variable besides *depiction* for which a higher F1-score can be obtained exploiting an auxiliary clustering ($MTL + R^{all}$) compared to focal training (MTL^{fo}). Combining the two strategies for training *SilkNet* ($MTL^{fo} + R^{all}$) further increases the F1-scores compared to applying the strategies independent from each other, i.e. compared to MTL^{fo} and $MTL + R^{all}$, respectively, for most of the variables. An exception in this regard is *depiction*; the F1-score for training with both strategies ($MTL^{fo} + R^{all}$) is only slightly higher (+0.6%) than the one achieved in the baseline experiment (MTL). This is reasonable, because focal training (MTL^{fo}) decreases the F1-score of *depiction* compared to MTL (-0.7%) and thus, counteracts the positive impact of the auxiliary clustering ($MTL + R^{all}$) in $MTL^{fo} + R^{all}$. All other variables obtain the highest F1-score in the experiment $MTL^{fo} + R^{all}$, where improvements of up to 8.9% (*material*) compared to MTL can be observed. In this context, a significant correlation between the differences in the F1-score $F1(MTL^{fo} + R^{all}) - F1(MTL)$ and the percentage of labelled examples for a variable (cf. section 4) of 93% (p-value: 0.02) can be determined, i.e. a larger positive effect is observed for variables with a larger percentage of labelled samples. As the differences in the F1-scores $F1(MTL^{fo}_{a-i}) - F1(MTL)$ also tend to be larger for variables with a larger percentage of labelled samples (87% correlation, p-value: 0.06), it is concluded that the magnitude of the improvements of $MTL^{fo}{+}R^{all}$ compared to MTL is limited by the need of focal training for a larger training set for a classification task. Nevertheless, the proposed combined training strategy is to be preferred over previously existing strategies addressing class imbalance for most of the variables.

Experiment	$F1$ [%]	OA [%]
MTL	28.6 ± 0.46	63.9 ± 0.21
MTL^{fo}	32.6 ± 1.11	65.9 ± 0.66
$MTL + R^{sem}$	30.5 ± 2.53	65.5 ± 0.70
$MTL + R^{all}$	31.9 ± 2.45	**66.2** ± 0.44
$MTL^{fo} + R^{sem}$	31.4 ± 0.41	65.4 ± 0.32
$MTL^{fo} + R^{all}$	**33.6** ± 1.03	65.8 ± 0.70

Table 4. Average F1-scores $F1$ [%] and overall accuracies OA [%] of all experiments.

Experiment	**Variable** m				
	de	*pl*	*ma*	*ti*	*te*
MTL	27.2	15.4	38.0	31.6	30.6
MTL^{fo}	26.5	19.6	46.8	38.4	31.8
$MTL + R^{sem}$	30.0	14.7	40.9	36.0	31.0
$MTL + R^{all}$	**32.1**	16.4	42.0	36.5	32.7
$MTL^{fo} + R^{sem}$	25.2	18.6	44.1	37.7	31.2
$MTL^{fo} + R^{all}$	27.8	**21.3**	**46.9**	**38.7**	**33.3**

Table 5. Average variable-specific F1-scores $F1$ [%]. *de*: *depiction*, *pl*: *place*, *ma*: *material*, *ti*: *time*, *te*: *technique*.

Analysing the performance of the different classifiers in correctly **predicting underrepresented classes** based on the F1-scores exclusively considering those classes (Table 6), there is a trend that is similar to the one shown in Table 5: The highest F1-score for *depiction* is achieved for training with an auxiliary clustering ($MTL + R^{all}$), whereas focal training (MTL^{fo}) decreases the score compared to the baseline training (MTL). Accordingly, the score obtained by *SilkNet* trained with focal weights and clustering ($MTL^{fo} + R^{all}$) is not higher than the one in the experiment $MTL + R^{all}$. For the other four variables, focal training (MTL^{fo}) is more beneficial than training with an auxiliary clustering ($MTL + R^{sem}$, $MTL + R^{all}$), whereas the highest scores are obtained in the experiment $MTL^{fo} + R^{all}$. Differences in the F1-score for underrepresented classes between $MTL^{fo} + R^{all}$ and MTL of up to +14.3% (*material*) are achieved, where the differences in the F1-score for underrepresented classes are larger than the one considering all classes. Accordingly, it is concluded that the proposed training strategy predominantly helps to learn a classifier to correctly predict underrepresented classes.

To summarize, the individual approaches aiming to handle problems with class imbalance, i.e. focal training and exploiting an auxiliary feature clustering, improve the classifier's ability to distinguish the individual classes (higher F1-score). The

Experiment	Variable m				
	de	*pl*	*ma*	*ti*	*te*
MTL	22.6	7.0	14.0	26.2	8.3
MTL^{fo}	22.0	10.4	27.7	34.9	8.7
$MTL + R^{sem}$	25.4	5.6	18.5	31.7	7.5
$MTL + R^{all}$	**27.3**	7.1	20.2	32.3	8.6
$MTL^{fo} + R^{sem}$	20.5	9.8	23.6	34.4	8.3
$MTL^{fo} + R^{all}$	23.3	**12.3**	**28.3**	**35.7**	**11.7**

Table 6. Average variable-specific F1-scores $F1$ [%] of the underrepresented classes $\mathcal{M}_m$ of all variables (background class not considered). Notation according to Table 5.

proposed combination of these strategies leads to the best performance with respect to the variable-specific F1-scores for all classes as well as the ones for underrepresented classes, where the positive effect is larger for underrepresented classes. In this context, the focal training aspect is found to result in a larger positive impact on the F1-scores (for all classes and underrepresented classes) for variables with a larger number of labelled training samples. Accordingly, the proposed approach realizing feature space clustering combined with focal training helps to mitigate problems with class imbalance.

6. CONCLUSIONS AND OUTLOOK

In this paper, a classification approach allowing for training a multi-task classifier using data with imbalanced class distribution for all of the tasks was proposed. The training strategy can deal with images both with a class label for all tasks to be learned as well as with a label for a subset of the tasks, which is particularly important in the context of applications in the context of cultural heritage preservation. The proposed training strategy combines multi-task multi-class focal training and an auxiliary feature clustering with respect to visual and semantic aspects of similarity to help the classifier to distinguish individual classes in a better way. Comprehensive experiments on a dataset of images depicting ancient silk fabrics showed that the proposed training strategy indeed improves the F1-score compared to standard multi-task training (+5.0% on average). The task-specific F1-scores are improved by up to 8.9%, where predominantly underrepresented classes benefit from the proposed combined training strategy (up to +14.3%). Moreover, the proposed training strategy was shown to be preferred over both training with focal weights and using an auxiliary clustering.

There are several directions for potential future work. One option for future work could be the modification of the data basis for further experiments. It could be observed that lower accuracies tend to be obtained for classes with a low number of training samples. Thus, data augmentation strategies could be applied to synthetically increase the number of training samples for such classes, similar to (Chawla et al., 2002). In this context, generative adversarial networks might be exploited to obtain synthetic data, e.g. (Pérez and Cozman, 2021). Moreover, experiments on other datasets, e.g. OmniArt (Strezoski and Worring, 2017) or SemArt (Garcia and Vogiatzis, 2018), could be conducted to analyse the generality of the findings in this paper in the context of heritage-related applications. Future methodological work related to both of the proposed approaches could address modifications of the mapping from an input image to high-level image features by modifying the used neural network. A strategy to do so could be to apply another network as a generic feature extractor. Furthermore, further additional data available for the SILKNOW dataset, i.e. information about relations between different instances as well as textual descriptions, could be exploited for training requiring methodological modifications of the training strategy to do so: The relations in the graph could be exploited similarly to (Garcia et al., 2020), where additional features are derived from the graph per training samples. The textual descriptions could also be considered in the context of multi-modal classification, being a growing field of research. Preliminary experiments involving the combination of images, class labels and textual descriptions (Rei et al., 2022) have shown promising results for the classification of historic silk fabrics. Finally, the principle of task balancing was already shown in (Yang et al., 2022) to be beneficial for multi-task learning with complete samples in the context of cultural heritage applications. It would be interesting to develop and investigate an according approach for incomplete training samples and, in particular, to combine such an approach with techniques addressing class imbalance.

ACKNOWLEDGEMENTS

The research leading to these results is in the context of the "SILKNOW. Silk heritage in the Knowledge Society: from punched cards to big data, deep learning and visual/tangible simulations" project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 769504.

References

Alba Pagán, E., Gaitán Salvatella, M., Pitarch, M. D., León Muñoz, A., Moya Toledo, M., Marin Ruiz, J., Vitella, M., Lo Cicero, G., Rottensteiner, F., Clermont, D., Dorozynski, M., Wittich, D., Vernus, P., Puren, M., 2020. From silk to digital technologies: A gateway to new opportunities for creative industries, traditional crafts and designers. The SILKNOW case. *Sustainability*, 12(19)(19). 12(19), 8279.

Ando, S., Huang, C. Y., 2017. Deep over-sampling framework for classifying imbalanced data. *Machine Learning and Knowledge Discovery in Databases*, Springer, 770–785.

Blessing, A., Wen, K., 2010. Using machine learning for identification of art paintings. Technical Report CS 229, Stanford University, USA.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (eds), *Advances in Neural Information Processing Systems (NIPS)*, 32, Curran Associates, Inc.

Caruana, R. A., 1993. Multitask learning: A knowledge-based source of inductive bias. *International Conference on Machine Learning (ICML)*, 41–48.

Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1), 321–357.

Dong, Q., Gong, S., Zhu, X., 2018. Imbalanced deep learning by minority class incremental rectification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(6), 1367–1381.

Dorozynski, M., Clermont, D., Rottensteiner, F., 2019. Multi-task deep learning with incomplete training samples for the image-based prediction of variables describing silk fabrics. *ISPRS Annals of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, IV-2/W6, Göttingen: Copernicus GmbH, 47–54.

Dorozynski, M., Rottensteiner, F., 2022a. Addressing class imbalance in multi-class image classification by means of auxiliary feature space restrictions. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 43, 777–785.

Dorozynski, M., Rottensteiner, F., 2022b. Deep descriptor learning with auxiliary classification loss for retrieving images of silk fabrics in the context of preserving European silk heritage. *ISPRS International Journal of Geo-Information (IJGI)*, 11(2), 82.

Dorozysnki, M., Wittich, D., Rottensteiner, F., Clermont, D., 2021. Artificial intelligence meets cultural heritage: Image classification for the prediction of semantic properties of silk fabrics. *Weaving Europe Silk Heritage and digital technologies*, Tirant lo Blanch, 147–166.

Garcia, N., Renoust, B., Nakashima, Y., 2020. ContextNet: representation and exploration for painting classification and retrieval in context. *International Journal of Multimedia Information Retrieval*, 9(1), 17–30.

Garcia, N., Vogiatzis, G., 2018. How to read paintings: semantic art understanding with multi-modal retrieval. *Proceedings of the European Conference on Computer Vision (ECCV)*, 676–691.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1026–1034.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 630–645.

Huang, C., Li, Y., Loy, C. C., Tang, X., 2016. Learning deep representation for imbalanced classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5375–5384.

Johnson, J. M., Khoshgoftaar, T. M., 2019. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1–54.

Khan, S. H., Hayat, M., Bennamoun, M., Sohel, F. A., Togneri, R., 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8), 3573–3587.

Kingma, D. P., Ba, J., 2015. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations (ICLR), San Diego, CA, USA, Conference Track Proceedings*.

Krawczyk, B., 2016. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems (NIPS)*, 1, 1097–1105.

Leiva-Murillo, J. M., Gómez-Chova, L., Camps-Valls, G., 2013. Multitask remote sensing classification. *IEEE Transactions on Geoscience and Remote Sensing*, 51(1), 151–161.

Li, S., Liu, Z.-Q., Chan, A. B., 2014. Heterogeneous multi-task learning for human pose estimation with deep convolutional neural network. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 482–489.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal loss for dense object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2980–2988.

Nair, V., Hinton, G. E., 2010. Rectified linear units improve restricted boltzmann machines. *International Conference on Machine Learning (ICML)*, 807–814.

Ortigosa-Hernández, J., Inza, I., Lozano, J. A., 2017. Measuring the class-imbalance extent of multi-class problems. *Pattern Recognition Letters*, 98, 32–38.

Pérez, S. P., Cozman, F. G., 2021. How to generate synthetic paintings to improve art style classification. *Brazilian Conference on Intelligent Systems*, Springer, 238–253.

Pouyanfar, S., Tao, Y., Mohan, A., Tian, H., Kaseb, A. S., Gauen, K., Dailey, R., Aghajanzadeh, S., Lu, Y.-H., Chen, S.-C. et al., 2018. Dynamic sampling in convolutional neural networks for imbalanced data classification. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 112–117.

Rei, L., Mladenic, D., Dorozynski, M., Rottensteiner, F., Schleider, T., Troncy, R., Lozano, J. S., Salvatella, M. G., 2022. Multimodal metadata assignment for cultural heritage artifacts. *Multimedia Systems*, 29, 847–869.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al., 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3), 211–252.

Schroff, F., Kalenichenko, D., Philbin, J., 2015. A unified embedding for face recognition and clustering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823.

Sridhar, S., Kalaivani, A., 2021. A survey on methodologies for handling imbalance problem in multiclass classification. *Advances in Smart System Technologies*, 1163, Springer, Singapore, 775–790.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1), 1929–1958.

Strezoski, G., Worring, M., 2017. Omniart: multi-task deep learning for artistic data analysis. *arXiv preprint arXiv:1708.00684*.

Sur, D., Blaine, E., 2017. Cross-depiction transfer learning for art classification. Technical Report CS 231A and CS 231N, Stanford University, USA.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., Liang, J., 2016. Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, 35(5), 1299–1312.

Tan, W. R., Chan, C. S., Aguirre, H. E., Tanaka, K., 2016. Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. *International Conference on Image Processing (ICIP)*, IEEE, 3703–3707.

Yang, B., Xiang, X., Kong, W., Peng, Y., Yao, J., 2022. Adaptive Multi-Task Learning Using Lagrange Multiplier for Automatic Art Analysis. *Multimedia Tools and Applications*, 81(3), 3715–3733. https://doi.org/10.1007/s11042-021-11360-7.

Yosinski, J., Clune, J., Bengio, Y., Lipson, H., 2014. How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems (NIPS)*, 27, 3320–3328.

Zhang, Z., Cui, Z., Xu, C., Yan, Y., Sebe, N., Yang, J., 2019. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4101–4110.