Neuro-symbolic learning for dealing with sparsity in cultural heritage image archives: an empirical journey

Agnese Chiatti[1], Enrico Daga[1]

[1]*The Open University, Milton Keynes, United Kingdom*

Abstract
Deep Learning (DL) methods have proved to be very successful for many image classification tasks. In the SPICE project, we are researching on an intelligent system that classifies artworks to support several tasks such as metadata curation and linking across image collections. However, applying DL methods to real-world cultural heritage collections for the task of artwork subject classification is problematic. Objects in this domain are characterised by different levels of heterogeneity: of media and techniques, of categories, of time-periods, just to mention a few. This heterogeneity makes the related training features sparsely distributed. In this paper, we report on an empirical investigation where we apply neuro-symbolic, Deep Learning techniques to a paradigmatic case of cultural heritage archive: the Tate Gallery collection open data. We pose the question of what type of feature engineering could help in reducing the impact of data sparsity in this domain. Crucially, we explore how neuro-symbolic learning, combining image features, textual metadata, and Knowledge Graph embeddings, could help in mitigating the problems derived from data sparsity in cultural heritage image archives.

Keywords
Artwork image classification, Neuro-Symbolic Learning, Knowledge Graph Embeddings

1. Introduction

Deep Learning (DL) methods have expedited the advancement on image classification tasks [1]. However, image classification through DL is still an open challenge in domains characterised by a high variance, for example, of data samples and labels [2, 3]. The negative impact of noisy labels, in particular, has been sufficiently acknowledged in the literature as one unavoidable problem in many real-world settings [4].

In the SPICE project, we are researching on an intelligent system based on DL that classifies artworks to support several tasks in the domain such as metadata curation or knowledge linking and discovery across cultural heritage archives. Crucially, in this domain, datasets are characterised by different types of heterogeneity - e.g., diversity of media and techniques, and of time-periods. Due to this variance, training features are sparsely distributed across the categories of interest.

In this paper, we explore the application of Deep Learning (DL) techniques to a paradigmatic case of cultural heritage archive: the Tate Gallery collection open data [5]. We interrogate on (a) what type of data preparation strategies could be applicable to these data and on (b) how neuro-symbolic learning, combining image features, tex-

tual metadata, and Knowledge Graph (KG) embeddings, could help in mitigating the problem of data sparsity.

To answer these questions, we devise a layered set of experiments. First, we aim at evidencing the negative impact of data sparsity on standard DL approaches and start by only considering visual features. Secondly, we look into how metadata could help in partitioning the training space and mitigating some effects of the sparsity of image features. Third, we incrementally introduce new features from textual metadata and background knowledge, including Knowledge Graph embeddings, and explore how they improve the classification performance.

The paper is structured as follows. After introducing the related work (Section 2), we provide the background context of this research (Section 3). Concurrently, we characterise the problem of data sparsity in artwork subject classification and present our research questions. In Section 4, we illustrate the approach and system architecture, which is based on current state of the art methods applied to this domain. Section 5 reports on the implementation of the experiments and results. Findings from these experiments are instrumental in deriving the lessons learnt and future directions of this work, as further discussed in Section 6.

2. Related work

Deep Learning (DL) is applied to a wide variety of problems in the context of cultural heritage applications [2, 3]. The tasks can range from the identification of artworks from noisy Web pictures (NoisyArt [6]), to the classification of artistic media [2], stylistic, and genre-specific artwork traits [7]. In this work, we focus on the problem

✉ agnese.chiatti@open.ac.uk (A. Chiatti); enrico.daga@open.ac.uk (E. Daga)

🆔 0000-0003-3594-731X (A. Chiatti); 0000-0002-3184-5407 (E. Daga)

of learning subject classifications from an heterogeneous cultural heritage archive. The problem of *classifying artwork subjects* is unique in its own respects compared to the types of image classification tasks that are typically tackled in the Cultural Heritage literature, which are thoroughly reviewed in [8]. The most relevant state of the art approach which we have identified to model the case of artwork subject classification is ContextNet [3], which focuses on learning a set of tasks such as Genre, Period, and School from an homogeneous set of paintings. A limitation of this approach is that only attributes in the target dataset are considered in the learning, disregarding other potential sources of artistic knowledge. To harness this potential, Castellano and Vessio proposed an extension of ContextNet, where properties gathered from Wikidata and DBpedia are used to construct a dedicated ArtGraph [9]. Inspired by the work in [3, 9] we propose to reuse the knowledge which has been previously distilled from DBpedia in the form of KG embeddings, through the RDF2Vec model [10]. Differently from prior works, we intend to explore the integration of off-the-shelf KG embeddings, as an alternative method to curating ad-hoc artistic Knowledge Graphs.

Moreover, we propose to adopt different types of embeddings to qualify different artistic features. Specifically, we test the integration of KG embeddings produced on the artist metadata with a linguistic model (distilBERT [11]), as a feature preparation from artwork titles. Combining different types of features and embeddings to leverage the strengths of the various approaches is common to recent research in neuro-symbolic learning [12]. However, no work so far explored the data sparsity problems that emerge when DL methods are applied to cultural heritage image collections.

3. Background and research questions

In the SPICE project [13], a team of researchers and museum professionals are developing novel methods for citizen participation and engagement, focused on the method of *slow looking*. This approach is based on designing scripts made of a set of prompts or questions about selected artworks [14]. Prompts are designed based on properties of the artworks that are *factual* (e.g. abstract, landscape, people, objects) rather then *contextual* (e.g. genre, period, author). A Deep Learning system should classify artworks according to a given subject list (e.g. abstract, landscape, people, objects) and use these metadata to link images across collections, thus helping curators in reusing scripts across similar artworks. However, cultural heritage collections are characterised by a heterogeneity of images and subject metadata. This characteristic hinders the successful application of state of the art DL approaches, which are typically developed on homogeneous samples (e.g. only on paintings), and optimised for categories that are not related to the actual content of the artwork (e.g. "Genre", "Century" [3].

The Tate Gallery archive is a paradigmatic case of cultural heritage image collection, characterised by artworks with a significant heterogeneity of *factual* properties. The dataset provides metadata and image urls of collection items summarised in two CSV files with general metadata (artworks and artists), and detailed metadata distributed in approximately 100k JSON files[1]. The collection includes artworks from more than three thousand artists spanning 142 genres over a period of approximately 500 years. Metadata was manually annotated by expert curators, including a taxonomy of more than 16 thousands distinct *subjects*, organised in 11 top-level subjects covering key concepts relevant to the slow looking application scenario: *abstraction*, *architecture*, *nature*, *people*, etc.. The Tate Gallery collection demonstrates dimensions of heterogeneity that are typical of cultural heritage image archives: • **Image heterogeneity**: images represent artworks produced with different mediums and techniques • **Sample heterogeneity**: the data distribution is very unbalanced • **Semantic heterogeneity**: subject annotations are based on the content of the assets but are produced incrementally over a large time-span by an unspecified number of annotators. This makes the labelled data incomplete, messy, and sparsely distributed.

In this paper, we explore how to address data sparsity from different perspectives, that we illustrate.

Tackling data sparsity by configuring the learning space. On the one hand, real-world subject taxonomies are overly heterogeneous both semantically and in terms of class population. This characteristic of real-world collections significantly complicates the learning of robust classification models. For example, learning to differentiate a collie from a spitz, is, in principle, more difficult than learning to tell dogs and cats apart, especially in the lack of sufficient examples representing different dog breeds. On the other hand, the few high-level subjects that have sufficient population (e.g., macro-classes such as people, nature, society, etc.), are also too generic, and, therefore, difficult to abstract from heterogeneous training samples. Thanks to the fact that subjects are organised taxonomically, we entertain the idea that we could use such sparsity at our advantage. Specifically, we make the hypothesis that learning low-level subjects can help the classification of high level subjects. That is, we ask: *[RQ1] Can we use the knowledge of the subject taxonomy structure to help with the categorization?*

---

[1]The following summaries are produced with SPARQL Anything [15] on the original data sources from the Tate Gallery Collection open data project on GitHub. Data and queries can be reviewed and reproduced [16]

Tackling data sparsity by partitioning data by the means of key features. We observe how metadata could provide a useful input in understanding the reasons of the visual heterogeneity of artwork collections. Specifically, we partition the learning task dividing the data by technique/medium used. For example, we compare the task of learning subjects on artworks of any medium with the results obtained when subjects are learned only on photography. Thus, we pose the following research question: *[RQ2] Does splitting the set by artwork medium (e.g., photography, graphite, sculpture,...) help?*

Tackling data sparsity with neuro-symbolic learning. Finally, we explore the impact of different combinations of features on the learning performance, including both visual and non-visual features, such as the textual embeddings from the artwork title and the knowledge graph embedding from the artist entity on DBpedia. In other words: *[RQ3] Does combining text embeddings, KG embeddings, and visual embeddings improve the categorization performance?*

4. Methodology

To devise an architecture for subject classification from artwork images, we take as a reference ContextNet [3], a state of the art architecture for neuro-symbolic learning on artwork classification tasks. Namely, we treat the different artwork subjects - e.g., nature, people, architecture, as tasks to learn jointly, in a Multi Task Learning (MTL) fashion. To generate visual embeddings from the input images, we maintain the same Convolutional Neural Network (CNN) backbone as ContextNet, i.e., a ResNet50 [17] from which the last fully-connected layer is removed.

However, differently from [3], our aim is to classify artwork subjects which can express a variety of factual elements about the artwork - e.g, physical, social, or abstract concepts. Therefore, we introduce a few modifications to the ContextNet framework, to accommodate the task of subject artwork classification. The resulting pipeline is illustrated in Figure 1.

First, different annotations which concern the same task (or subject) can co-exist in an artwork. For instance, J.M.W. Turner's "Nant Peris, looking towards Snowdon" in Figure 1 depicts a mountain peak and a river view, both overlooked by a cloudy sky. These elements all fall under the *nature* subject. Thus, we configure the Network for multi-label classification. Specifically, in the last layer of the Network, Softmax activation is replaced by a Sigmoid activation. Indeed, while Softmax activations, which are typically interpreted as classification probabilities, are distributed across neurons, sigmoid outputs are computed independently on each class node. As a result, Sigmoids can predict more than one class with high

probability. Similarly, we rely on a binary cross-entropy loss function, so that multiple subject predictions can be generated for an artwork.

In ContextNet, the visual embeddings are projected to the vector representation of the artwork context in the broader painting set [3]. This representation is derived from a KG where paintings are grouped by author, and also annotated with attributes such as timeframe and medium. In the pipeline of Figure 1, we adopt a similar approach to ContextNet and infuse the Network with background knowledge through an encoder module, optimised through a smooth $\ell 1$ loss function. However, we test a different combination of embeddings to represent non-visual artwork features. In particular, we embed the title and artist metadata in the visual representation of the artwork. For the artwork title, we capitalise on a pre-trained linguistic model which has been shown to provide compact textual embeddings: DistilBERT [11]. To model the authorship information, instead, we apply the off-the-shelf RDF2Vec model [10] to the DBpedia entities which represent each artist. We can further concatenate the linguistic and KG embeddings, to derive a unified representation for the injected background features. Nonetheless, because the individual features are maintained as separate modules (Figure 1), we can also test the effects of incrementally adding new features to the learning process.

Ultimately, different components contribute to the overall training loss of the Network. We use the same notation as [3] to characterise these contributing factors through a set of parameters. First, the visual classification is influenced by the different learning tasks. Formally, the contribution of the t-th task to the binary crossentropy loss (ℓ_c) is weighted with respect to a λ_t so that $\sum_{t=1}^{T} \lambda_t = 1$. Similarly, because the classification and encoder modules are optimised through different loss functions, the relative contribution of each function to the overall loss is weighted through different parameters. Let these weights be λ_c for the classifier loss, and λ_e, i.e., the complement to 1 of λ_c, for the encoder loss. By proxy, these two parameters allow us to leverage the degree to which the visual and non-visual components of the embeddings influence the training.

5. Experiments

5.1. Data preparation

In our experiments, we focus on the Tate Gallery Open Data as our reference cultural heritage archive. The published data includes two summaries CSVs and more then 10k JSON files with detailed metadata on artworks and artists. The SPARQL Anything framework [15] provides a means to filter and integrate the required features from
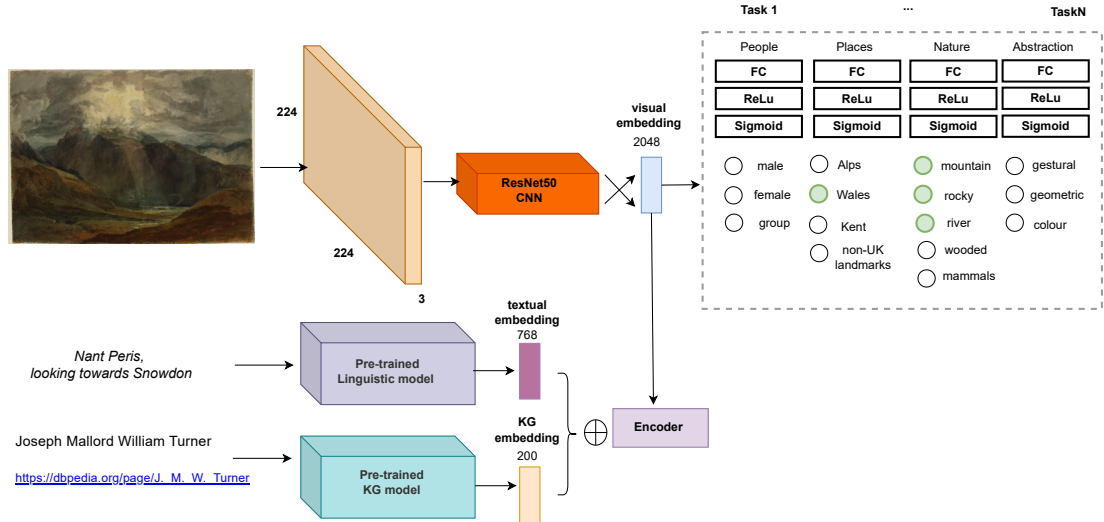
Figure 1: The proposed architecture for artist subject classification, where visual embeddings are combined with textual and KG embeddings, to jointly train the Network on multiple tasks - e.g., recognising people, places, nature, and so forth.

Table 1

Statistics of the top-level subjects in the Tate collection.

Top subject	Artworks	Subjects
Nature	36'477	24
Architecture	29'787	19
Places	23'842	12
People	20'798	14
Society	13'991	6
Objects	12'381	10
Abstraction	8'503	8
Emotions, Concepts, Ideas	8'248	4
Work and Occupations	5'133	3
Symbols and Personifications	5'022	3
Leisure and Pastimes	3'129	2

the broader CSVs and JSON documents provided in this collection [16]. In addition to retrieving the JPEG image files that are available from the Tate website, we extracted the following data fields: (i) the unique artwork identifier, (ii) the title, (iii) the subject annotations, as well as (iv) the medium, or material, of each artwork. We also queried DBpedia to retrieve the entities, marked through a Uniform Resource Identifier (URI), which match a certain artist name. Because a string can match multiple DBpedia entities and to account for homonyms, we manually validated the collected artist entities.

We start by considering the 11 top-level concepts that provide abundant training examples (i.e., at least 3'000 examples per class), for the purpose of supervised Deep Learning. These are listed in Table 1.

Moreover, we want to reduce the sparsity of the ex-

ample distribution across the sub-categories of a subject. With the term sub-category, we refer to the children of a subject, in the Tate taxonomy - e.g., "female figure" is a sub-category of "people". Thus, we focus on the six subjects which provide the highest number of sub-concepts. As highlighted in blue in Table 1, these are: nature, architecture, places, people, objects, and abstraction.

With this premise, we further prune the space of sub-concepts with a two-fold objective. First, because samples should ideally overlap across the top-level subjects that are learned jointly, we select sub-categories which are worth at least 700 examples. Second, for each subject, we want to select non-overlapping sub-categories which identify distinct concept groups. Thus, if two children categories of the same node are selected at the previous step, only the more specific one is retained. For instance, we prioritise annotations of male and female portraits over the more generic portrait label.

After retaining only records which are annotated with respect to the target categories and sub-categories, we are left with 54'494 records[2], 99.6% of which (54'293 records) have a non-corrupted image file associated.

Ultimately, we want to ensure that examples are balanced across categories when forming our training, validation, and test splits. Therefore, we resort to the multilabel stratified sampling strategy proposed in [18, 19], which is conveniently provided with the scikit-multilearn package[3]. At this stage, we apply a 80/10/10 ratio to split

---

[2]This number is lower than the sum of the figures in Table 1, as the same artwork can be annotated with more than one subject.
[3]http://scikit.ml/

the data into training, validation, and test sets.

Additionally, because we are also interested in grouping artworks by artistic medium (RQ2), we prepared a dedicated subset for each medium. The metadata which describe the different artistic media are sparsely annotated. Thus, we ought to apply a series of basic Natural Language Processing (NLP) steps to converge towards coherent groups. Specifically, we reduced the raw text to lowercase, and derived a set of tokens which excludes the standard English stopwords, and which is free from alphanumeric characters and spurious white spaces. After deriving word stems through the Snowball method, we also filtered out any duplicated tokens - e.g., "*paper graphite, on paper*".

Thanks to the availability of a reference glossary of art terms on the Tate website[4], we could derive a set of keywords to canonicalise heterogeneous medium annotations. Specifically, we merged semantically-related keywords (e.g., ink and pen) to gather a sufficient number of data points per medium. In sum, we converged towards ten subsets that are representative of different materials. These are: *painting, sculpture, graphite, etching, screenprint, watercolour & gouache, ink & pen, lithograph, engraving & intaglio*, and *photography*.

5.2. Experimental setup

To address our main research questions (Section 3), we configure three distinct experiments. Across all experiments, the performance on each task (i.e., subject) is evaluated in binary terms. Specifically, in addition to the overall classification accuracy on a task, we also track the Precision (P), Recall (R), and F1 achieved on the retrieval of positive examples of a task. For instance, in the context of the *people* task, the correct recognition of a person in a portrait increases the P, R, and F1 metrics, irrespective of the system's ability to recognise that a different artwork does not depict any people.

Experiment A. The objective of the first experiment is to assess whether or not configuring the learning space on the basis of the Tate subject taxonomy improves the classification performance (RQ1). Thus, we start by considering a simplified version of the architecture presented in Section 4, where only the visual embeddings extracted from a CNN are considered. In this context, we compare two training configurations. The first configuration only relies on the macro-categories which represent each task, whereas the second configuration considers finer-grained annotations for each task. In the former case, the Network is optimised to generically classify people, places, objects, natural, abstract elements, and architectural components. In the latter configuration, the goal

is to learn sub-classes of each subject - e.g., to recognise mountains, rivers, and beaches, as opposed to classifying "nature" generically - as in the example of Figure 1.

Experiment B. The second experiment is conceived to test the effects of splitting training examples by artwork medium - e.g., watercolour, sculpture, intaglio, to further reduce the sparsity of the learning space (RQ2). Therefore, in this setup, we start by evaluating the performance results obtained when the complete image collection is considered, without discriminating by artistic medium. We then repeat the performance assessment across the ten sub-samples which we have prepared for different artistic media, as described in Section 5.1.

Experiment C. The last experiment is a study of the impact of visual, textual, and KG embeddings when learning artwork subjects (RQ3). In particular, we explore the integration of numeric features representing the title and artist of an artwork. Therefore, at this stage, we contrast the performance of the following pipelines, or ablations:

(img) The first method considers only visual embeddings to classify artworks, thus following the same methodology of Experiments A and B.

(img + text) In this pipeline, the visual embeddings are also optimised with respect to the linguistic embeddings extracted from a pre-trained DistilBERT model. Specifically, the linguistic model is fed with the title of each artwork. To derive a single vector for each input sentence, we follow a series of transformations which are standard practice in NLP. First, the hidden states produced by the last four layers are summed together, to derive a word vector for each input token. Then, we average the second to last hidden layers of each token to form the sentence embedding.

(img + KG) We also test a neuro-symbolic variation of the "img" pipeline, where the visual embeddings are projected onto the 200-dimensional embeddings returned by a RDF2Vec model [10] which was pre-trained on DBpedia entities. Specifically, if the DBpedia URI associated with an author is found in the RDF2Vec feature space, the related KG embedding is retrieved to guide the optimisation of the visual embedding, through the encoder module (Section 4).

(img + text + KG) Lastly, we consider the scenario where the textual embedding and the KG embedding are concatenated, to contribute to the learning routine. In other words, this ablation models the methodology of Section 4 (Figure 1).

---

[4]https://www.tate.org.uk/art/art-terms/

Table 2
Performance comparison when we change the hierarchical level (granularity) of the subject labels used for training.

(a) Training on subject level one

Task	Acc	Pre	Rec	F1
nature	0.64	0.64	1	0.78
people	0.65	0	0	0
architecture	0.47	0	0	0
objects	0.8	0	0	0
places	0.58	0	0	0
abstraction	0.86	0	0	0

(b) Training on subject level two

Task	Acc	Pre	Rec	F1
nature	0.39	0.45	0.01	0.02
people	0.65	0.29	0	0.01
architecture	0.52	0.37	0.01	0.01
objects	0.84	0.14	0	0
places	0.63	0.21	0	0
abstraction	0.86	0.75	0.01	0.02

5.3. Results

Experiment A. The results obtained on the test set when training the model only on macro-subject labels are reported in Table 3a. With the exception of the nature task, the model is incapable of detecting the presence of any artwork elements. The non-zero accuracies indicate that the model has learned to produce only negative predictions for the tasks (except for nature). In fact, for the *nature* class, which contributes the highest number of training examples, the model outputs mostly positive predictions. Hence, it has simply learned to replicate the imbalanced distributions of the training data.

However, when finer-grained subject annotations are introduced, the performance improves, particularly in terms of Precision (Table 3b). Nevertheless, the overall performance remains dramatically low across all tasks. Indeed, while introducing finer-grained categories may help discriminating different subjects, it also makes the distribution of training examples for each specialised category more sparse.

Experiment B. Based on findings from the previous experiment, here we consider finer-grained subject categories for training our models. However, in this experiment, data points are further sampled by artistic medium. As shown in Figure 2, in the majority of cases where the model is trained only on a specific medium, the F1 score is higher than in the baseline scenario, where all materials are considered. A marginal performance decay was instead recorded when classifying *places* from *sculptures*, where the already low F1 dropped to zero. In the scenario where only *graphites* were considered, all results are equivalent or only marginally higher than the baseline curve, except for the *abstraction* task, were the improvement was most pronounced. In the remaining scenarios, splitting examples by artistic medium significantly benefited the performance. In particular, the highest F1 scores were achieved by training only on: • photographies to classify *nature* and *people* • engraving & intaglio examples to classify *people* and *places* • ink & pen works on the *architecture* task • sculptures to classify

objects • screenprints to classify *abstraction*.

Experiment C. Figures 3 and 4 illustrate the results obtained with the top-performing methods of Experiment B (i.e., those trained solely on *photography*, *engraving & intaglio*, *sculpture*, *ink & pen*, and *screenprint*), through different combinations of features. Crucially, the integration of non-visual features enhanced the baseline DL performance across the majority of tasks and media. However, different performance trends can be observed that are medium-specific and task-specific.

To classify photography, linguistic embeddings were relatively more beneficial, in terms of performance increase, than KG embeddings (Figure 3a). However, combining different embedding types led to highest F1 on the classification of nature, people, and objects. Similarly, on the ink & pen sample, the integration of all tested features produced the highest performance when classifying nature, people, places, and abstraction (Figure 3d).

In the case of sculptures, the largest margin of improvement is associated with the introduction of linguistic embeddings, for the majority of tasks (Figure 3c). Screenprints, instead, exhibit an opposite trend: overall, integrating KG embeddings was preferable, in terms of performance, than only relying on linguistic features (Figure 3e). Interestingly, the top performance achieved through the baseline on the abstraction task was unmatched, even after integrating both the title and the artist features (Figure 4a). Indeed, the abstraction subjects explored in this evaluation mostly encode colour and geometric traits, which are best learned through visual features.

On the engraving & intaglio set, the effects of applying neuro-symbolic learning differs from task to task (Figure 3b). On the nature, architecture, places, and abstraction tasks, the introduction of text and KG embeddings ensured a significant performance increase. By contrast, the improvement is only marginal when classifying objects. The case of the people task is interesting because the integration of the linguistic embeddings led to a performance degrade, and the introduction of KG embeddings only matched the baseline F1. However, leveraging both types of embeddings improved the performance by 5%.
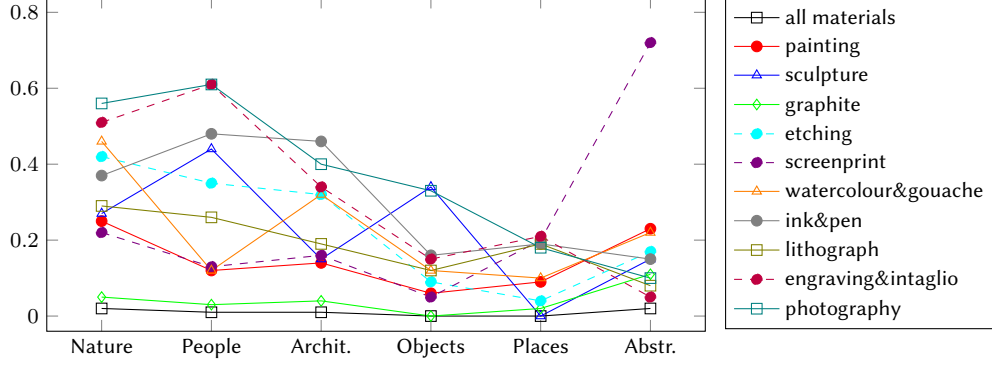
Figure 2: Comparison of F1 scores before (black curve) and after (coloured curves) splitting the dataset by artwork media.

Overall, different tasks are learned most efficiently through a different combination of features, on different mediums. Specifically, the highest F1 scores are observed: • **for nature**: *img+text* on *engraving* (Figure 4b) • **for people**: *img+text+KG* on *photography* (Figure 4d), • **for architecture**: *img+text* and *img+KG* on *engraving & intaglio* (Figures 4b,4c) • **for objects**: *img+text+KG* on *photography* (Figure 4d), *img+text* on *sculpture* (Figure 4b) • **for places**: *img+text* on *engraving & intaglio* (Figure 4b) • **for abstraction**: *img* on *screenprint* (Figure 4a).

5.4. Implementation details

The experiments discussed in this section were conducted on Google Colaboratory. In our training configuration, we initialised the ResNet50 module with weights pre-trained on ImageNet. Weights of the classification heads were instead initialised through the Xavier method [20]. Consistently with [3], parameters were updated via stochastic gradient descent. In particular, we set the learning rate to 0.0001, with a weight decay of 0.00001 and a momentum of 0.9. Through preliminary experiments to this paper, we found that updating parameters across the entire model is preferable than fine-tuning only the last classification layer, likely due to the marked differences between the ImageNet benchmark and the Tate collection. All tested models were trained for up to 150 epochs, with an early stopping condition whenever the validation loss did not decrease for 30 successive epochs.

Each task contributed equally to the classification loss, i.e., for the six tasks explored in this paper, λ_t was set to 0.165. After testing different weight configurations for the loss, we set $\lambda_c = 0.9$ and $\lambda_e = 0.1$ across all experiments. That is, we observed empirically that giving higher importance to the visual embeddings at training time ensures a higher performance, on average.

Input images were resized to 224×224 and normalised with respect to the ImageNet mean and standard devia-

tion. We relied on the Pytorch and transformers Python libraries to implement the proposed architecture. The RDF2Vec embeddings, which we downloaded locally to speed up the processing time, are conveniently exposed through the KGVec2Go resource [21].

The code, data, and pre-trained models which reproduce these experiments are available at: https://bit.ly/3p3WV3M.

6. Discussion and Conclusions

We conducted experiments with the purpose of exploring strategies for handling the data sparsity that characterises cultural heritage collections. We asked whether the taxonomical structure of the labels can help in learning top level categories (RQ1). Indeed, organising the training space according to the semantic hierarchy of objects helped improving the classification performance. However it is not sufficient, alone, to handle sparsity. Next, we explored the idea that image features may vary depending on the artistic medium (RQ2) and therefore learning should be performed separately for each medium. We demonstrated how splitting by artistic medium helped significantly improving the performance across the majority of tasks (with only one caveat: *places* on *sculptures*). Finally, we conducted extensive experiments to study the impact of different feature sets on the various subjects, learnt on the different media (RQ3). Here, we observe that different features are helpful for learning different tasks on different media, and that there is no feature set that performs systematically better on all media and subjects.

On the basis of these findings, we elaborate on possible future work. First, learning what combinations are performing well was a costly operation. This evidence poses the question of how to autonomously learn the feature set that is most representative for each subject. Hence,
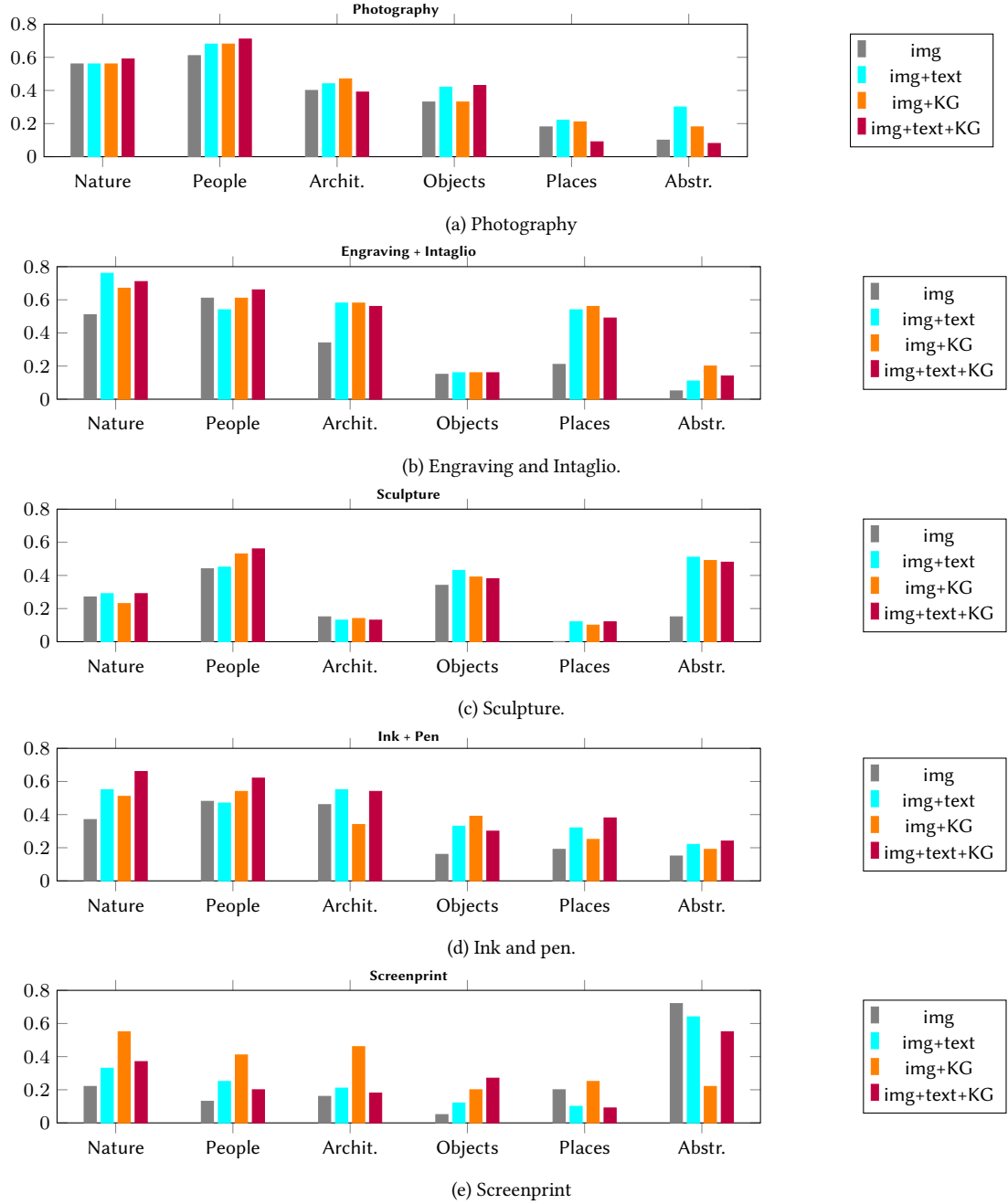
Figure 3: Experiment results by *medium*: comparison of F1 scores after incrementally combining visual embeddings (img) with textual embeddings on the artwork title (text) and Knowledge Graph embeddings on the artist feature (KG).

these results spark an important meta-learning task: **To what extent can we automatically devise the appropriate learning strategy depending on the three dimensions of *feature modality*, artistic *medium*, and *subject*?**

Second, we found that certain tasks are better learned on certain media. Could we use this behaviour as an opportunity for transfer learning? In other words, could
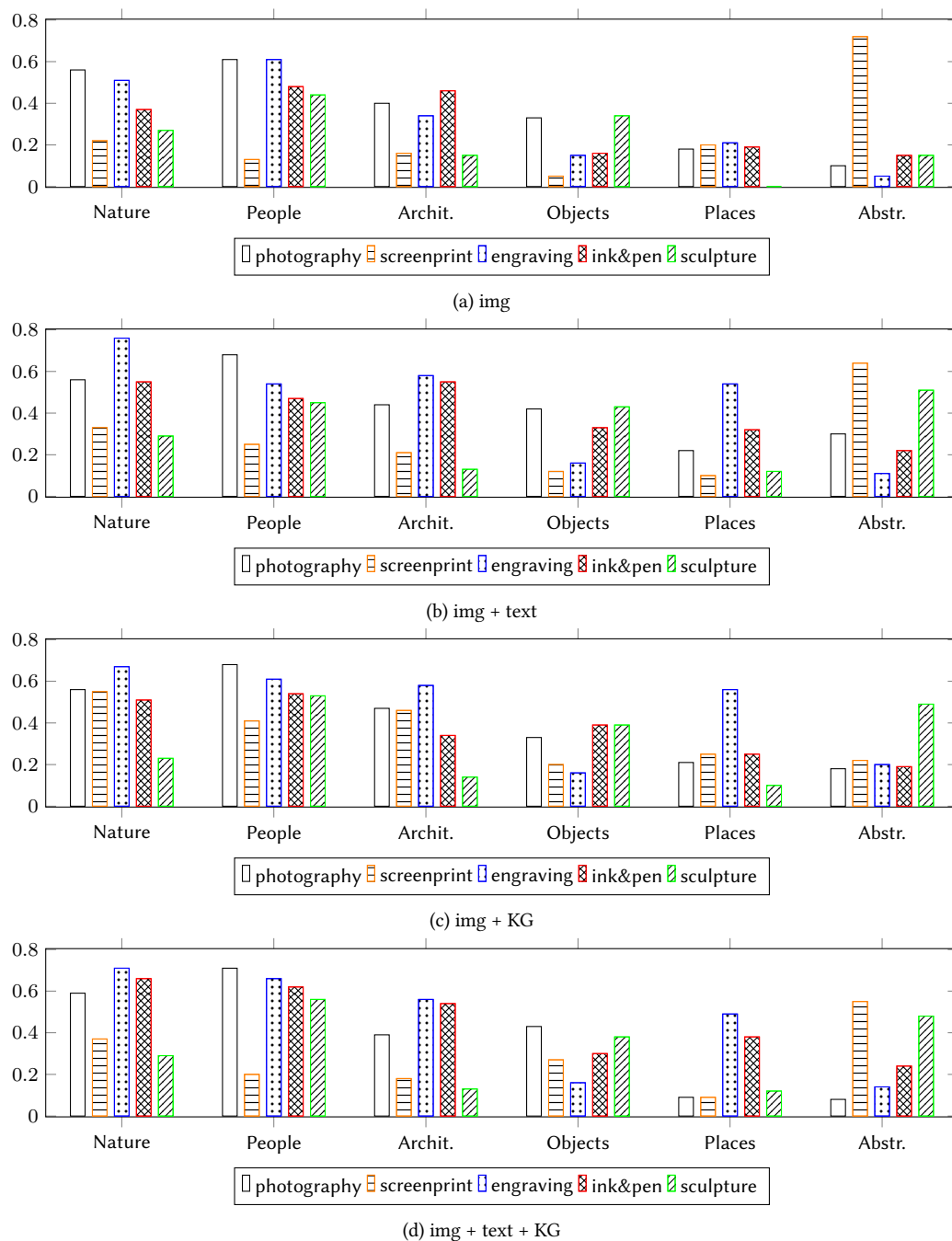
Figure 4: F1 results for the experiments with different features (img, text, KG), grouped by medium.

we use a model trained on a specific medium to recognise the same subject on another medium? Even further, could we use the learned model on a different image collection, to compensate for the imbalance (or even scarcity) of

training examples? In the context of the SPICE project, we are particularly interested in characterising subjects on citizen-curated collections such as the Irish Museum

of Modern Arts (IMMA) archive[5].

Finally, in this work, we have attacked the classification problem as a Multi Task Learning (MLT) setting, following relevant priors in the state of the art. Future work includes exploring the correlation between different tasks, i.e., studying which tasks are best learned jointly and which ones should be learned separately.

Acknowledgments

References

[1] L. Schmarje, M. Santarossa, S.-M. Schröder, R. Koch, A survey on semi-, self-and unsupervised learning for image classification, IEEE Access 9 (2021) 82146–82168.

[2] H. Yang, K. Min, Classification of basic artistic media based on a deep convolutional approach, The Visual Computer 36 (2020) 559–578.

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

[4] G. Algan, I. Ulusoy, Image classification with deep learning in the presence of noisy labels: A survey, Knowledge-Based Systems 215 (2021) 106771.

[5] The Tate Gallery, Tate Collection metadata, 2014. URL: https://github.com/tategallery/collection.

[6] R. Del Chiaro, A. D. Bagdanov, A. Del Bimbo, Noisyart: A dataset for webly-supervised artwork recognition., in: VISIGRAPP (4: VISAPP), 2019, pp. 467–475.

[7] S. Liu, J. Yang, S. S. Agaian, C. Yuan, Novel features for art movement classification of portrait paintings, Image and Vision Computing 108 (2021) 104121.

[8] G. Castellano, G. Vessio, Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview, Neural Computing and Applications 33 (2021) 12263–12282.

[9] G. Castellano, G. Sansaro, G. Vessio, Integrating contextual knowledge to visual features for fine art classification, in: DL4KG'21: Workshop on Deep Learning for Knowledge Graphs, CEUR, 2021.

[10] P. Ristoski, J. Rosati, T. Di Noia, R. De Leone, H. Paulheim, Rdf2vec: Rdf graph embeddings and their applications, Semantic Web 10 (2019) 721–752.

[11] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.

[12] M. K. Sarker, L. Zhou, A. Eberhart, P. Hitzler, Neurosymbolic artificial intelligence: Current trends, arXiv preprint arXiv:2105.05330 (2021).

[13] E. Daga, L. Asprino, R. Damiano, M. Daquino, B. D. Agudo, A. Gangemi, T. Kuflik, A. Lieto, M. Maguire, A. M. Marras, et al., Integrating citizen experiences in cultural heritage archives: requirements, state of the art, and challenges, ACM Journal on Computing and Cultural Heritage (JOCCH) 15 (2022) 1–35.

[14] P. Mulholland, E. Daga, M. Daquino, L. Díaz-Kommonen, A. Gangemi, T. Kulfik, A. J. Wecker, M. Maguire, S. Peroni, S. Pescarin, Enabling multiple voices in the museum: challenges and approaches, Digital Culture & Society 6 (2020).

[15] E. Daga, L. Asprino, P. Mulholland, A. Gangemi, Facade-X: an opinionated approach to SPARQL anything, in: Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, volume 53, IOS Press, 2021, pp. 58–73.

[16] E. Daga, SPARQL Anything showcase: open data from the Tate Gallery, 2022. URL: https://doi.org/10.5281/zenodo.6518424. doi:10.5281/zenodo.6518424.

[17] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[18] K. Sechidis, G. Tsoumakas, I. Vlahavas, On the stratification of multi-label data, Machine Learning and Knowledge Discovery in Databases (2011) 145–158.

[19] P. Szymański, T. Kajdanowicz, A network perspective on stratification of multi-label data, in: L. Torgo, B. Krawczyk, P. Branco, N. Moniz (Eds.), Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications, volume 74 of *Proceedings of Machine Learning Research*, PMLR, ECML-PKDD, Skopje, Macedonia, 2017, pp. 22–35.

[20] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the thirteenth international conference on artificial intelligence and statistics, 2010, pp. 249–256.

[21] J. Portisch, M. Hladik, H. Paulheim, Kgvec2go–knowledge graph embeddings as a service, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 5641–5647.

---

[5]http://imma.ie