Applying deep learning approaches to mixed quantitative-qualitative analyses

Applying deep learning approaches to mixed quantitative-qualitative analyses

LORENZO STACCHIO, University of Bologna, Department for Life Quality Studies, Italy
ALESSIA ANGELI, University of Bologna, Department of Computer Science and Engineering, Italy
GIUSEPPE LISANTI, University of Bologna, Department of Computer Science and Engineering, Italy
GUSTAVO MARFIA, University of Bologna, Department of the Arts, Italy

We here verify whether a quantitative approach, i.e., a deep learning-based one, may be used to synthesize a model apt to perform specific qualitative analyses. To this aim, we leverage a previous contribution, where we approached the concrete problem of implementing a socio-historical classification toolchain for a collection of vernacular photos. In such a work, after individuating a corpus of vernacular photographs we devised the process that follows. First, we resorted to existing socio-historical categories derived from previous qualitative studies. Secondly, we involved the people included in the photos in the annotation process of a subset of the corpus of data. We then fine-tuned and deployed existing deep learning models to classify the entire corpus of data. Finally, we compared the results obtained with our approach to the ones obtained by a socio-historian. We hence here focus on the relationship between quantitative and qualitative methods considering the specific case of socio-historical analyses.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; **Ubiquitous and mobile computing systems and tools**; *Empirical studies in collaborative and social computing*;

Additional Key Words and Phrases: multimedia document analysis, mixed methods, deep learning, vernacular photos

ACM Reference Format:
Lorenzo Stacchio, Alessia Angeli, Giuseppe Lisanti, and Gustavo Marfia. 2022. Applying deep learning approaches to mixed quantitative-qualitative analyses. In *Conference on Information Technology for Social Good (GoodIT'22), September 7–9, 2022, Limassol, Cyprus.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3524458.3547265

1 INTRODUCTION

The relations between quantitative methods and qualitative analyses, their potentials, and limits, represent open questions within different research communities [3, 4, 18, 34]. Due to the growth of digital and digitized data, qualitative analyses are becoming more and more expensive and difficult to apply to massive datasets, and quantitative methods seem to be the only way to deal with them [3]. Eventually, the results obtained adopting quantitative methods could converge to those returned by qualitative ones: the authors of [3] compared a qualitative approach, from interpretive social science, and a quantitative one, from natural language processing, on textual data showing that these methods produced similar results.

However, some criticism emerges also for quantitative methods, which may improperly apply the definition of measurement, simply matching tasks, objects, and events to numbers, according to specific rules [18]. In addition, they may be misleading due to insufficient care in data collection, feature definition and processing, and adherence to the domain of interest [4]. For example, in social studies, Choy [8] pointed out that throughout quantitative methodologies,

Manuscript submitted to ACM

different people and communities characteristics, such as identities, perceptions, and beliefs, cannot be meaningfully converted to numbers or adequately explained without references to the proper context in which people live.

Despite such critics, recent research in computer science has been focusing on how quantitative methods may be able to support qualitative analyses [3, 12, 25, 29]. Along this line, the authors of [25] highlighted that, although a variety of issues have emerged with the use of machine learning models in social data analyses, the intersection of machine learning and the social sciences has provided critical new insights into social behavior. Indeed, they stated that *"machine learning can and should become a critical piece of social science"* and *"similarly, social science should become an increasingly important part of machine learning"*. In support of this aspect, in [12], the authors provided a critical analysis of a corpus of research that resorted to human annotation to produce datasets for supervised learning, considering data from Twitter. Doing so, they observed the similarities between creating human-labeled datasets and content analysis, underlining the importance of utilizing high-quality training data to produce high-quality classifiers. Another work that highlights the importance of data quality comes from [29], which criticizes how computer vision datasets are built, emphasizing the importance of categories structuring methods.

This work wants to contribute to such debate, showing how quantitative and qualitative methods can coexist to carry out integrated analyses to evaluate more data than those usually examined in a qualitative process while adopting a well-defined theoretical foundation. To do this, we build on the contribution presented in [32], concentrating on the relationship between qualitative and quantitative approaches. In [32], we analyzed a dataset built on top of analog family album photos using a well-defined socio-historical protocol for its digital cataloging.

Analog photographs stand, indeed, a unique opportunity to learn about the recent past. No other visual media captured the human habits in the 20th century as vernacular photography (e.g., travel photos, family snapshots, photos of friends and classes) [19, 20]. Among vernacular photography, family photo albums have gained attention in academia due to their ability to reveal sociological and historical insights regarding specific cultures and times [5]. As reported in [10], *"as people struggled with this broadening of their family album, other narratives began to emerge within those already established of colonialism, imperialism, migration, and dispossession"*. Indeed, family album photos represent a reference point for the conservation, transmission, and development of a community Social Heritage [7]. In addition, their importance stands not only in their value to study the habits of humankind but also for education and psychology [16, 21]. To summarize, scholars from different fields describe such pictures as capable of capturing peculiar characteristics regarding local communities in space and time.

Nevertheless, despite their importance, there is a lack of digital archives accompanied by a shortage of high-quality cataloging and quantitative methods [12, 29]. Indeed, in studying this kind of phenomenon, scholars usually resort to a small corpus of photos often gathered and verified adopting custom protocols [7, 28] and draw their conclusions embracing qualitative analyses approaches [24]. The adoption of qualitative methods has been so far justified by the small number of items socio-historians have at their disposal and by a general skepticism around the adoption of quantitative methods.

We show how mixing qualitative and quantitative methods may overcome such difficulties. Therefore, we verify whether quantitative techniques, built resorting to results obtained from qualitative processes, may be employed to perform specific qualitative analyses, yielding a mixed qualitative-quantitative one. More in detail, we exhibit the process that follows, focusing on the concrete problem of implementing a socio-historical classification toolchain for a collection of vernacular photos. To start, we focused on existing socio-historical categories derived from previous sociological and historical qualitative studies [6]. These are the same categories that were also adopted, following a rigorous protocol, to label the photos included in the dataset. The annotation process amounted to the step which let us

build quantitative methods to perform an analysis that typically demands qualitative ones: in brief, classify a photo according to the given socio-historical categories. The generated dataset allowed us to build deep learning models capable of recognizing salient features of socio-historical interest, opening to the possibility of exploiting quantitative methods with qualitative ones to improve cataloging processes. To this aim, we fine-tuned and deployed existing deep learning architectures for the socio-historical context classification. After this, we compared the results obtained with our approach to the ones obtained by a socio-historian who manually assessed the photos. The results acquired with this test confirmed that quantitative approaches may integrate qualitative ones to benefit the overall performance and speed of socio-historical analyses.

In summary, the contributions of this work amount to:

- The discussion of how a "Family Album" collection obtained through qualitative approaches has been exploited to perform a quantitative analysis that mimics the qualitative one performed by socio-historical scholars;
- A quantitative analysis of the collection, resorting to Convolutional Neural Network (CNN) [15, 17, 33] and Transformer-based [11] models, both trained to provide socio-historical context predictions;
- A comparison of the performance obtained by a socio-historical scholar, employing a qualitative analysis of the photos, with the performance of the CNN and Transformer-based deep learning models. This provides an exemplar case integrating quantitative and qualitative approaches to speed and increase the amount of processed and cataloged data.

The rest of the paper is organized as follows: Section 2 delivers a description and the main characteristics of the IMAGO dataset. Section 3 and Section 4 present, discuss and validate several deep learning models applied to the proposed dataset to highlight the difficulties and define an evaluation baseline. Follows Section 5 with a comparison between quantitative methods (i.e., based on deep learning models) and a qualitative analysis (i.e., performed by a socio-historical scholar). Finally, Section 6 concludes this work.

2 FAMILY ALBUM DATASET

The IMAGO project is a digital collection of analog family album photos gathered and maintained by the Department of the Arts of the University of Bologna (a presentation of the project that has led to the creation of the collection may be found at `imago.unibo.it`). Including more than 80,000 photos shot between 1845 and 2009, it contains approximately 1,500 family albums. A total of 16,642 images received a label by their owners under the supervision of a socio-historical faculty. This label describes the socio-historical category the photo belongs to [7]. These 16,642 images, from now on, will be referred to as the IMAGO dataset[1], the dataset analyzed in this work.

The annotation process of the photos followed a well-defined protocol. First, with a lecture the socio-historical background, the dataset construction goals, and the different classification categories were presented and explained to the owners of the photos. Important to note that the different categories, related to the socio-historical context, were produced by a qualitative coding process [27], based on sociological and historical criteria derived from different researches. Then, a second lecture covered the annotation problem in detail, focusing on the reliability and authenticity of sources of socio-historical materials. This process highlighted the importance of interviewing the original owner of the photo. In case such person(s) were not available (e.g., old photo), one could find a second-hand informed party (e.g., anyone informed of the context). Alternatively, an attempt to infer the socio-historical information could be made by

[1]The IMAGO resources are available upon request.

analyzing any written annotations behind the photo. Whenever none of these paths led to a solution, the photo would be discarded.

It is interesting to notice that, from a socio-historical perspective, the information provided by the owner of a photograph amounts to the ground truth, justifying the path followed during the annotation process. It is the owner that injects in the dataset the social component along with the historical one. Such an approach is not new to the computer vision community. Other works in literature have also considered as image metadata the information provided by their owners [2, 22]. The owner or a directly connected party (e.g., relatives, friends) holds the ground truth. For this reason, it is not possible to resort to just any automatic labeling services (e.g., Amazon SageMaker Ground Truth or the Google AI Platform Data Labeling Service [1, 13]) to obtain a high-quality annotation of given datasets. These elements emphasize the quality and the uniqueness of such datasets, including IMAGO.

Here follow the socio-historical context categories individuated in the IMAGO dataset [31] along with a brief explanation:

(a) *Affectivity*: these photographs show people (e.g., couples, friends, or families) bound by inter-personal relationships;
(b) *Work*: photos belonging to this class portray people sitting or standing in workplaces and wearing work clothes;
(c) *Fashion*: photos belonging to this class contain symbolic objects and clothes, such as suits, trousers, skirts, and coats;
(d) *Motorization*: this class includes symbolic objects such as cars and motorcycles, which represent a social and historical landmark;
(e) *Music*: this class may also include scenes from leisure time, characterized in this case by musical events or the presence of musical instruments;
(f) *Politics*: this class contains photos related to political gatherings and demonstrations;
(g) *Rites*: in the pictures comprised in this category, it is possible to find portraits of the sacred or celebratory events which characterize the life of a family;
(h) *School*: this class includes all the photos which represent schools, often characterized by symbolic objects (e.g., desk, blackboard) or a group of students;
(i) *Free-time*: such category investigates the forms and ways of experiencing leisure time, reconstructing, wherever possible, generational and gender differences. It also includes images representing people who make new experiences, visits far-off landmarks, expand social relationships, and interact with nature.

Fig. 1 reports the number of labelled images available per socio-historical category. Here it is possible to observe the unbalance among the different classes and, in particular, the dominant classes are *Affectivity*, *Fashion* and *Free-time*. Fig. 2, instead, shows a few exemplar images where it is possible to appreciate some of the characteristics that are present in the IMAGO dataset (e.g., number of people, clothing, colors, and location).

3 DEEP LEARNING MODELS

In this Section, we first describe the considered deep neural network architectures and then provide further information about their training settings.

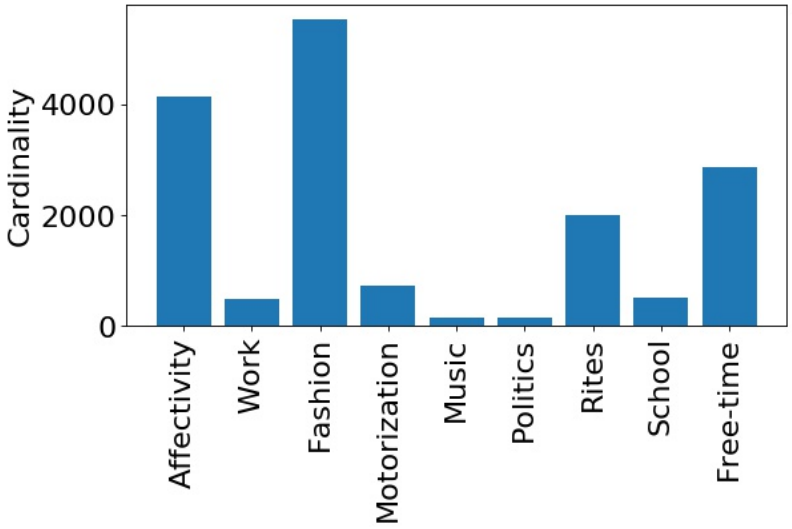


Fig. 1. Socio-historical classes distribution.

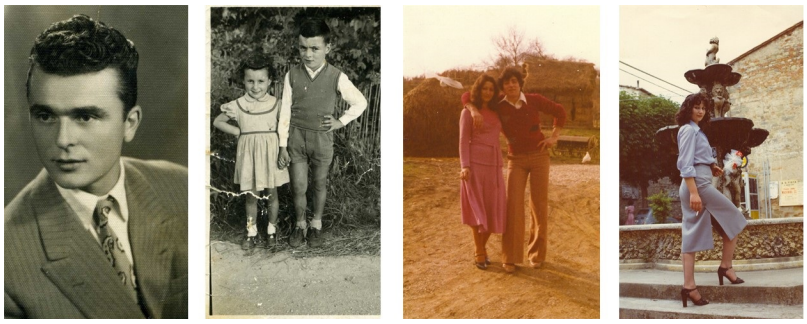


Fig. 2. Sample images.

3.1 Model architectures

In this work, we considered different deep learning models. Each model utilizes as backbone either a CNN or a Transformer, pre-trained on Imagenet [9], and a fully connected layer for the final classification. The Transformer is a deep learning architecture that relies entirely on the self-attention mechanism to draw global dependencies between input and output [36]. Recent works have shown that such an approach can achieve comparable or even superior performance compared to CNNs [11, 14, 35]. In particular, the Vision Transformer (ViT) architecture, proposed by Dosovitskiy et al. [11], achieved the state of the art performance on several computer vision benchmarks.

Each model was modified replacing the top-level classifier with a new classification layer, whose structure depends on the task (i.e., the number of output classes) and whose weights have been randomly initialized. The CNN models exploited in our experiments are based on three well-known architectures: DenseNet121 [15], ResNet-50 [17] and InceptionV3 [33]. The pre-trained convolutional layers have been specifically fine-tuned according to the architecture. Instead, for the ViT architecture, we proceeded to fine-tune the ViT-Tiny and the ViT-Small configurations, considering

Table 1. Socio-historical model accuracies for an increasing Top-k classification (k ranging from 1 to 5).

	Model				
	CNN			**Vision Transformer**	
Architecture	**DenseNet121**	**ResNet-50**	**InceptionV3**	**ViT-Tiny**	**ViT-Small**
input dim	256	256	299	224	224
#params (K)	6,963	23,526	25,130	5,526	22,669
Top-1	63.72	**64.35**	64.08	53.62	60.96
Top-2	83.38	**85.00**	83.83	76.22	81.70
Top-3	92.37	**92.85**	92.28	87.35	90.11
Top-4	96.54	96.66	**96.75**	93.15	95.07
Top-5	98.47	98.35	**98.53**	96.63	97.72

224×224 as input size and 16×16 pixels for the patches. We selected these configurations to fairly compare in terms of model complexity ViT to convolutional models. The number of parameters of ViT-Tiny is comparable with DenseNet121, while the number of parameters of ViT-Small is comparable to ResNet-50 and InceptionV3.

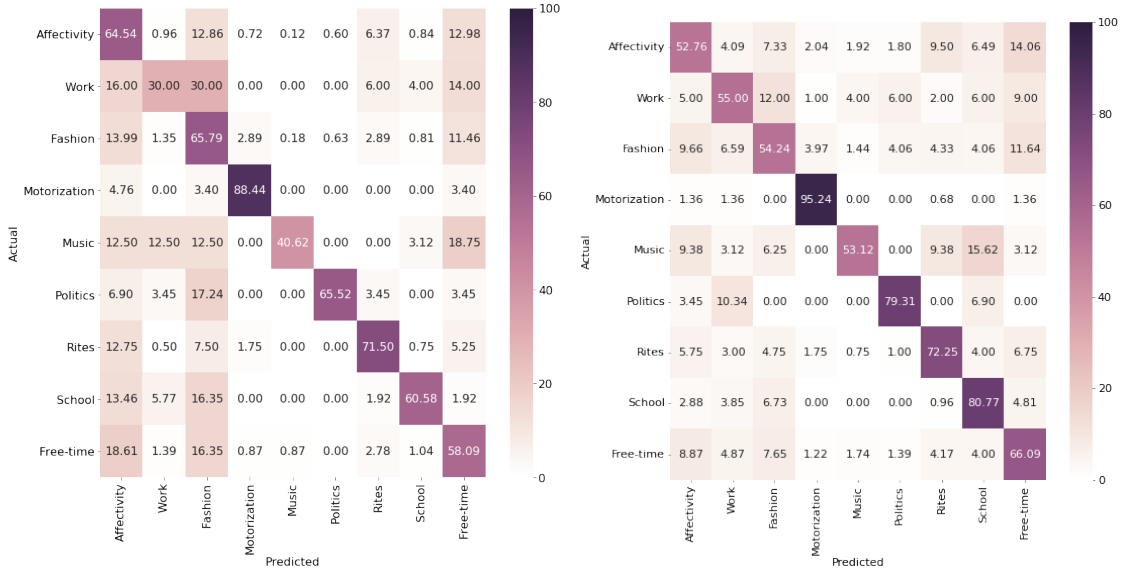
3.2 Dataset splitting and training settings

The IMAGO dataset has been partitioned as follows: 80% for training and 20% for testing. In addition, 10% of the training images is used as the validation subset for the model hyperparameters tuning.
For the training of the CNN-based models, we applied random cropping and horizontal flipping to make the model less prone to overfitting. Each model has been fine-tuned using a weighted cross entropy loss [23] to counter the dataset unbalance. We employed the Adam optimizer, with a learning rate of 1e-4, a weight decay of 5e-4, and we set the batch size to 32. Instead, for the Transformer-based model training, we followed the procedure reported in [11] while adopting a weighted cross-entropy loss.

4 EXPERIMENTAL RESULTS

This Section reports on the results obtained for the socio-historical context classification using the different deep learning models described so far. The entire IMAGO dataset, 16,642 images, is used in our analysis. The results are reported in Table 1 in terms of top-k accuracy: if the correct class is not the one with the highest predicted probability but falls among the k with the highest predicted probabilities, it will be counted as correct.

It is evident from this Table that the CNN-based models exhibit a higher performance compared to those based on Transformers. It is interesting to highlight that: the model based on ResNet-50 achieves, in most cases, the best performance among all the other CNN backbones, and ViT-Tiny performed worse than Vit-Small. For these reasons, from now on, we will adopt the ResNet-50 and ViT-Small architectures for all the experiments that follow.



(a) Confusion matrix for ResNet-50 model

(b) Confusion matrix for ViT-Small model

Fig. 3. Confusion matrices for a CNN-based and a Trasformer-based deep learning model.

From the confusion matrix reported in Fig. 3a, it is possible to observe that, for the ResNet-50 model, three classes are responsible for the largest share of misclassifications. In fact, the model is often misled to choose one among *Fashion*, *Affectivity* and *Free-time*, instead of the correct class. On the one hand, this follows as some classes contain images that share visual elements with the three listed above. For example, photos from the *Free-time* category display people involved in different kinds of activities in heterogeneous environments. The *Work* class, instead, is mostly confused with *Fashion* and *Affectivity*. This could be explained by the fact that the images that fall within the *Work* class represent groups of people wearing work-suits or being in working environments (e.g., people in a cafeteria, market, or factory). In fact, for example, workers wearing a uniform or some particular cloth items could be classified in the *Fashion* class while belonging to the *Work* class. A similar phenomenon takes place also for the *Music* class. Again, people taking part in some musical events could be classified as *Free-time* while belonging to the *Music* class. On the other hand, as shown in Fig. 1 (Section 2), the different socio-historical classes are not balanced in the IMAGO dataset making it difficult to correctly classify samples from those classes which are underrepresented.

As mentioned before, ViT-Small exhibits a slightly lower performance but comparing the confusion matrices, it is worth noticing that it obtains, when compared to ResNet-50, a more balanced per-class accuracy, as shown in Fig. 3b. As also highlighted in [26], this could be explained by the fact that ViT incorporates more global information than ResNet-50 at lower layers, leading to different features while preserving spatial information.

This is also qualitatively shown in a few examples, reported in Fig. 4, where we compare the activations obtained by ViT-Small with the ones of ResNet-50.

The aim now is to understand which characteristics led the trained models to determine the socio-historical context of an image and, to do so, we exploited the Grad-Cam algorithm [30]. In brief, this algorithm delimits the areas of an image used by deep learning models for the classification. More in detail, in Fig. 4 we show some correctly

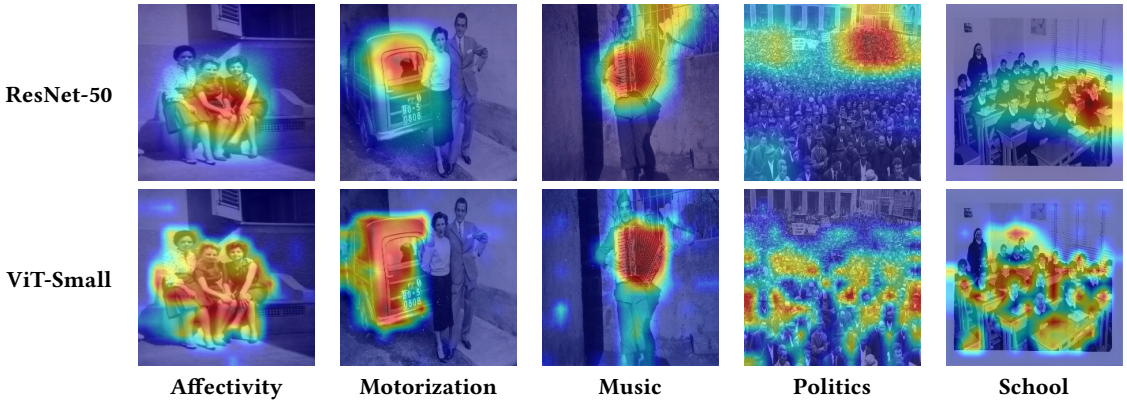
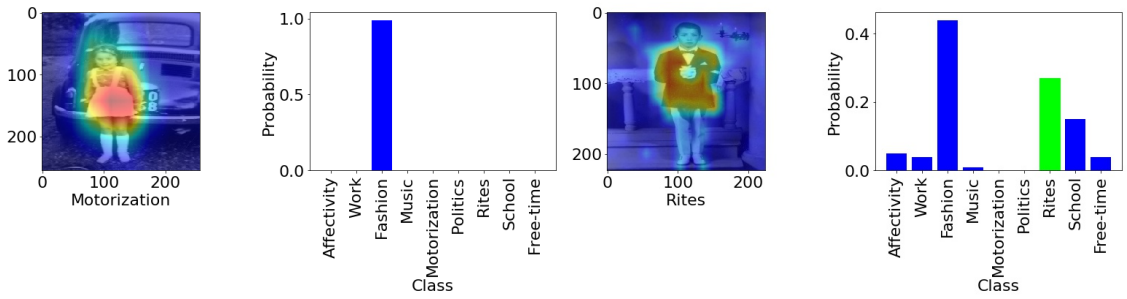


Fig. 4. Grad-Cam analysis of socio-historical contexts, of pictures within IMAGO, using ResNet-50 and ViT-Small.



(a) ResNet-50 (b) ViT-Small

Fig. 5. Grad-Cam analysis related to wrong classifications of ResNet-50 and ViT-Small models: these images exhibit the bias of the neural network for different socio-historical classes (*Motorization* and *Rites* recognized as *Fashion*).

classified IMAGO pictures processed by the Grad-Cam algorithm. Per each row, five images belonging respectively to the *Affectivity*, *Motorization*, *Music*, *Politics* and *School* classes are shown. Here, we have the same images processed by the Grad-Cam algorithm applied to ResNet-50 (first row) and ViT-Small (second row). Such images are representative of the characteristics that the classifier learned for each class. More in detail, specific poses which characterize the affection between people, as shown in the first image, are typical of the *Affectivity* class. Vehicles, or parts of them, as well as musical instruments, help recognize the pictures that belong to the *Motorization* or the *Music* class, as shown in the second and third image, respectively. The presence of crowds of people is typical of pictures in the *Politics* class (fourth image). Instead, a limited group of children, as well as children wearing school uniforms, may be exploited by the model to recognize pictures in the *School* class (last image). Examining these samples, it is not surprising that the model was able to correctly classify pictures belonging to the *Motorization* and *Music* classes, as these are clearly characterized by specific objects and, more importantly, already part of the model pre-trained on ImageNet. However, also for the majority of the other classes, the model seems to be able to learn some discriminative features.

Fig. 5, instead, reports two misclassified examples. Both of them were classified as belonging to the *Fashion* class, but the leftmost picture belongs to *Motorization* while the rightmost to *Rites*. The classes provided by the model, in this case,

are off-target since the leftmost picture depicts a child standing in front of a car, while the rightmost reports a child with an elegant jacket and bow tie. These errors may be due to different reasons. The model that analyzed the first image focused on secondary details of the pictures (i.e., fashion aspects) rather than the main ones (i.e., car). For what concerns the second image, instead, the misclassification may be due to the very specific point of view of its owner. This amounts to a situation where it may be hard for the deep learning models to predict the correct class. In such a case, only the owner of the photo knows whether the child there portrayed was elegantly dressed for a rite or another special occasion. The classifier is driven by those clothes to choose the *Fashion* class. Nevertheless, the owner's point of view amounts to the ground truth, according to the method adopted in this work. This proves the intrinsic challenge that the socio-historical context classification poses, since any classifier, including a socio-historian scholar, may be subject to such kinds of errors. For this reason, we further investigated such phenomenon in Section 5, analyzing the differences between the results obtained with the deep learning model (quantitative) and the analysis made by a socio-historian (qualitative).

5 A COMPARISON BETWEEN QUANTITATIVE AND QUALITATIVE METHODS

To assess how deep learning models performed in comparison to a scholar, we designed a specific experiment where a socio-historian was asked to categorize all the 3,327 images belonging to the IMAGO test set, providing for each picture its socio-historical context. It is worth noticing that the socio-historian employed a qualitative approach to classify each image, exploiting his/her sociological and historical background, and so not only based on the contents of the picture.

The deep learning model can provide a Top-k accuracy while, for this experiment, the socio-historian could freely select multiple categories per photo. Although unrestricted to use as many labels as desired, it is interesting to note that no more than three have been considered at once. Then, to make a fair comparison, we considered the k most probable classes chosen by the models in comparison with the k selected by the socio-historian, where k ranged from 1 to 3, depending on the choice made by the socio-historian.

Considering the test set, an accuracy of **72.24**% was obtained by ResNet-50, compared to **68.98**% obtained by ViT-Small and **66.93**% of the socio-historian. It is possible to observe that the deep learning models obtained a higher performance than the socio-historical scholar. The socio-historian could fail at recognizing particular details that only the photo owner could have known, while the deep learning model focuses on specific visual characteristics that, in some cases, led to correct classification. The results of this experiment highlight how quantitative methods could support qualitative analyses, at least in the context of socio-historical studies.

6 CONCLUSION

With this work, we delved into the debate concerning how the integration of quantitative and qualitative methods should occur, experimenting on how quantitative methods may support qualitative ones. To this aim, we analyzed the IMAGO dataset, composed of family album photographs, built adopting a well-founded socio-historical labeling protocol based on the definitions provided by socio-historical scholars [7].

In particular, we introduced a quantitative method based on deep learning to predict the socio-historical context classification of family album images. We trained and tested different deep learning models based on CNNs and Transformers. We then proceeded to compare the performance of the trained models with the one of a socio-historical scholar. The results of such assessment proved that quantitative methods could not only speed up the cataloging processes but also support socio-historians in carrying out qualitative analyses of complex or large catalogs of visual information. Clearly, this is only one step in the direction of exploiting quantitative models to support qualitative

analysis, which may take into account all the processes involved in the complex socio-historical domain. Finally, this line of work may benefit from the adoption of a multi-modal approach, which considers also other sources usually analyzed by means of qualitative methods (e.g., historical texts).

ACKNOWLEDGMENTS

This work was supported by the University of Bologna with the Alma Attrezzature 2017 grant and by AEFFE S.p.a. and the Golinelli Foundation with the funding of two Ph.D. scholarships.

REFERENCES

[1] Amazon. 2021. Amazon SageMaker Ground Truth. https://aws.amazon.com/it/sagemaker/groundtruth/.
[2] F Basura, M Damien, R Khan, and T Tuytelaars. 2014. Color features for dating historical color images. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2589–2593.
[3] Eric PS Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K Gay. 2017. Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? *Journal of the Association for Information Science and Technology* 68, 6 (2017), 1397–1410.
[4] Nick Black. 1994. Why we need qualitative research. *Journal of epidemiology and community health* 48, 5 (1994), 425.
[5] D Calanca. 2004. Percorsi di storia della famiglia. *Rivista di storia e storiografia* 5, 5 (Nov. 2004), 203–210.
[6] D Calanca. 2005. Album di famiglia. Autorappresentazioni tra pubblico e privato (1870-1950). *Storia e Futuro - N° 8-9* (2005).
[7] D Calanca. 2011. Italians posing between public and private. Theories and practices of Social Heritage. *Almatourism-Journal of Tourism, Culture and Territorial Development* 2, 3 (2011), 1–9.
[8] Looi Theam Choy. 2014. The strengths and weaknesses of research methodology: Comparison and complimentary between qualitative and quantitative approaches. *IOSR Journal of Humanities and Social Science* 19, 4 (2014), 99–104.
[9] J Deng, W Dong, R Socher, L J Li, K Li, and L Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848
[10] Andrew Dewdney. 1991. More than black and white: The extended and shared family album. *Family Snaps: The Meaning of Domestic Photograph, London: Virago* (1991).
[11] A Dosovitskiy, L Beyer, A Kolesnikov, D Weissenborn, X Zhai, T Unterthiner, M Dehghani, M Minderer, G Heigold, S Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[12] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.
[13] Google. 2021. AI Platform Data Labeling Service. https://cloud.google.com/ai-platform/data-labeling/docs.
[14] K Han, Y Wang, H Chen, X Chen, J Guo, Z Liu, Y Tang, A Xiao, C Xu, Y Xu, et al. 2020. A survey on visual transformer. *arXiv preprint arXiv:2012.12556* (2020).
[15] G Huang, Z Liu, L Van Der Maaten, and K Q Weinberger. 2018. Densely Connected Convolutional Networks. arXiv:1608.06993 [cs.CV]
[16] Lei Jiang, Panote Siriaraya, Dongeun Choi, and Noriaki Kuwahara. 2021. A Library of Old Photos Supporting Conversation of Two Generations Serving Reminiscence Therapy. *Frontiers in Psychology* (2021), 3633.
[17] H Kaiming, Z Xiangyu, R Shaoqing, and S Jian. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV]
[18] Joel Michell. 1999. *Measurement in psychology: A critical history of a methodological concept*. Vol. 53. Cambridge University Press.
[19] G Mitman and K Wilder. 2016. *Documenting the world: film, photography, and the scientific record*. Univ. of Chicago Press.
[20] MoMA. 2020. Vernacular photography. https://www.moma.org/collection/terms/vernacular-photography.
[21] Robin Notshulwana and Naydene de Lange. 2019. "I'm me and that is enough": Reconfiguring the family photo album to explore gender constructions with Foundation Phase preservice teachers. *Teaching and Teacher Education* 82 (2019), 106–116.
[22] F Palermo, J Hays, and A A Efros. 2012. Dating historical color images. In *European Conference on Computer Vision*. Springer, 499–512.
[23] T H Phan and K Yamamoto. 2020. Resolving Class Imbalance in Object Detection with Weighted Cross Entropy Losses. arXiv:2006.01413 [cs.CV]
[24] Jon Prosser. 1998. The status of image-based research. *Image-based research: A sourcebook for qualitative researchers* (1998), 97–112.
[25] Jason Radford and Kenneth Joseph. 2020. Theory in, theory out: the uses of social theory in machine learning for social science. *Frontiers in big Data* 3 (2020), 18.
[26] M Raghu, T Unterthiner, S Kornblith, C Zhang, and A Dosovitskiy. 2021. Do Vision Transformers See Like Convolutional Neural Networks?. In *Thirty-Fifth Conference on Neural Information Processing Systems*.
[27] Johnny Saldaña. 2021. *The coding manual for qualitative researchers*. sage.
[28] M Sandbye. 2014. Looking at the family photo album: a resumed theoretical discussion of why and how. *Journal of Aesthetics & Culture* 6, 1 (2014), 25419.

[29] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[30] R R Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, and D Batra. 2019. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision* 128, 2 (Oct 2019), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[31] P Sorcinelli. 2004. Imago. Laboratorio di ricerca storica e di documentazione iconografica sulla condizione giovanile nel XX secolo. *Rivista di storia e storiografia* 5, 5 (Nov. 2004), 200–202.

[32] Lorenzo Stacchio, Alessia Angeli, Giuseppe Lisanti, Daniela Calanca, and Gustavo Marfia. 2022. Towards a holistic approach to the socio-historical analysis of vernacular photos. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

[33] C Szegedy, V Vanhoucke, S Ioffe, J Shlens, and Z Wojna. 2015. Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs.CV]

[34] Abbas Tashakkori and John W Creswell. 2007. The new era of mixed methods. , 3–7 pages.

[35] H Touvron, M Cord, M Douze, F Massa, A Sablayrolles, and H Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*. PMLR, 10347–10357.

[36] A Vaswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A N Gomez, and I Polosukhin Ł Kaiser. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.