Content-Based Image Classification for Sheet Music Books Recognition

Diego Jesús Lozano-Mejía
Universidad Peruanas de Ciencias Aplicadas (UPC)
Lima, Peru
u201413062@upc.edu.pe

Enrique Paul Vega-Uribe
Universidad Peruanas de Ciencias Aplicadas (UPC)
Lima, Peru
u201411809@upc.edu.pe

Willy Ugarte
Universidad Peruanas de Ciencias Aplicadas (UPC)
Lima, Peru
willy.ugarte@upc.pe

***Abstract*—Modern digital music libraries have grown to contain a very large number of musical representation and retrieving images from them may be difficult for people with no prior experience. This study presents a comparison of several convolutional neural networks (CNN) architectures performance on music sheet classification, which are state-of-the-art computer vision methods to perform classification tasks. The models were trained using randomly selected sheets from different sheet music books and used to classify the source book of the validation data. To evaluate the models with incomplete images, we divide each image of our dataset in nine equal parts, then test the models with them. Performance evaluation of the CNNs prove that they can be very effective in this task.**

***Keywords*—CNN, Sheet Music, Deep Learning**

I. Introduction

Through the years, digital music libraries have grown to amass a large number of musical representations on their repositories in the form of sheet music. In recent years, the increase of sheet music images on internet, due to its widespread accessibility in most devices, have allowed the birth of several public domain digital music libraries like the International Music Score Library Project (IMSLP).

Nowadays, modern digital music libraries have several methods of image retrieval. One of them are the bibliographic values, like the name of a sheet music, composer name or the year of publication. There is also the information on music form, which consists in searching using the sheet's musical concepts [1]. These text-based retrieval methods may result complex for users who lack enough knowledge to perform an optimal search; because of this, Content-Based Image Retrieval (CBIR) allows the user to use a query image as input and retrieve the category based on the image content [2].

During the last few years, content-based systems focus on lower-level features of color, texture, shape and spatial layout [3], [2], [4], and basic machine learning techniques i.e support vector machine, artificial neural networks, decision tree, Bayesian method, non-parametric and parametric approaches [1]. However, recent studies show that the use of deep learning approaches like convolutional neural networks (CNN) for features extraction and transfer learning approach achieved big success in computer vision applications [5]. Based on the above mentioned, this paper proposes the use of transfer learning using current state-of-art top-accuracy CNN architectures to classify sheet music books using sheet music. To the extent of our knowledge, there are no previous studies that use fine-tuning for sheet music books classification.

The rest of this paper is organized as follows. Section 2 describes the main concepts related to our research. Section 3 presents the main contribution of this study. Section 4 discusses the related works. Section 5 describes the experiments we performed comparing several pre-trained CNN architectures. Section 6 discusses the results and future works.

978-1-7281-8367-1/20/$31.00 ©2020 IEEE

II. Background

Now, we describe the key elements related to deep learning and other techniques used for documents classification.

Machine learning approaches can be divided in three general categories: supervised learning where the algorithms are trained using labeled data used to predict outcomes with new data, unsupervised learning where the training data is unlabeled using the systems to find hidden patterns in it and reinforcement learning where the method interacts with an environment and learns by making errors or getting rewards.

Definition 1 (Deep Learning [6])**.** *Deep learning allows computers to learn from experience and understand the world in terms of a hierarchy of concepts, with each concept defined through its relation to simpler concepts.*

This approach avoids the need for human operators to specify all the knowledge the computer needs to learn something because it is gathered from experience. Deep learning can be see as neural networks with a large number of parameters and layers in one of four fundamental network architectures.

- Unsupervised Pretrained Networks (UPNs)
- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks
- Recursive Neural Networks

Deep learning also offers the advantage of automatic feature extraction over traditional machine learning algorithms, these networks decide which characteristics of a dataset can be used as indicator to label data reliable, so we can spend less time manually creating feature sets for data classification.

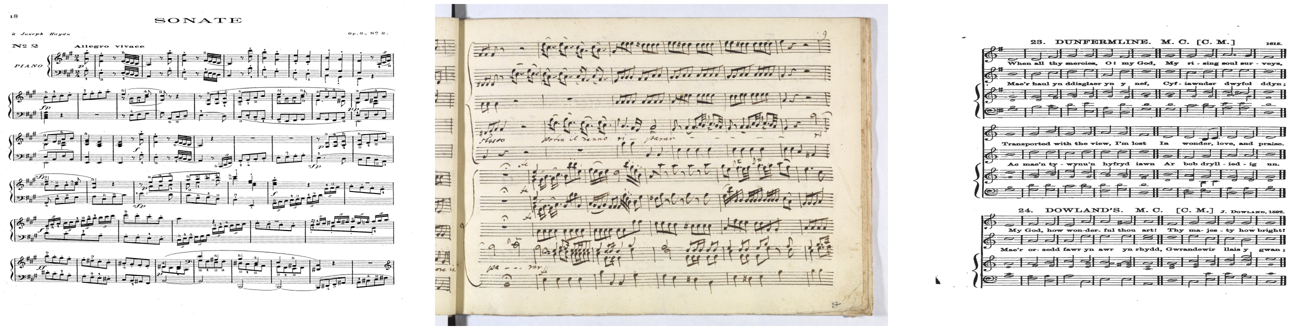


Fig. 1: Example of three sheet music books

Convolutional networks, also known as convolutional neural networks, or CNNs, are defined in [6] as a specialized kind of neural network for processing data that has a known grid-like topology like a time-series data, which can be thought of as a 1-D grid, and image data, which can be thought of as a 2-D grid of pixels. CNNs use convolutions as specialized linear operation, it can be said that they are simply neural networks that use convolutions in place of general matrix multiplication in at least one of their layers. In [6], it's explained that the CNN architectures are based on three major groups.

The input layer is configured to accept a three dimensional input in the form of the image size (width×size) and its color channels, the feature extraction layers is composed of the convolutional layer and the pooling layer which find the images features and uses them to construct higher-order features and the classification layer where one or more fully connected layers take the higher-order features.

Definition 2 (Digital Image Processing [7])**.** *An image may be defined as a two-dimensional function. $f(x, y)$, where x and y are spatial coordinates, and the amplitude of f at any pair of coordinates (x, y) is called the intensity or gray scale level of the image at that point. When x, y, and the values of f are all finite, discrete quantities, we call the image a digital image. The field of digital image processing refers to processing digital images by means of a digital computer.*

There are several techniques for digital processing images, each with different objectives: *image restoration* which attempts to reconstruct or recover an image that has been degraded; *color image processing* used in problems like coloring a monochrome image; image compression for reducing the amount of data required to represent a digital image; *morphological image processing* used as a tool for extracting image components that are useful in the representation and description of region shapes such as boundaries or skeletons.

III. Main Contribution

According to [1], content-based image retrieval (CBIR) is a framework based on the visual analysis of contents that are part of the query image. Using a query image as an input is the main requirement of CBIR, matching the visual contents of query image with other images in an archive, and visual similarity provides a base to find images with similar contents.

In CBIR, the low-level visual features of the image (like color, shape or texture) are computed from the image and matching of these features is performed to sort the output. The selection of visual features for any system depends on the requirements of the end user. To make the features more robust and unique in terms of representation, the fusion of low-level visual features is used but high computational cost is required to obtain more reliable results, also, the improper selection of features can decrease the performance of image retrieval model. These features are automatically computed to search and sort similar images from archives with minimum human interaction. Because of this, CBIR and feature extraction approaches are applied in various applications such as medical image analysis, remote sensing, crime detection, video analysis, military surveillance and textile industry [1].

Dataset: For this research, a labeled dataset with 13,763 scanned images from a total of 33 books was built. This group of images are in the public domain and were obtained from the International Music Score Library Project (IMSLP).
Getting a dataset with only sheets, we got rid of blank pages and pages composed only of text or images like covers or indexes, thus the dataset only contains images of sheets which represent books. Fig. 1 shows examples of the dataset. Furthermore, based in [8], another dataset was created with each sheet divided in nine parts, we called this dataset "Split-9". This process was done outside of the CNNs, with the objective of preserving the pre-trained CNNs architectures. The images are normalized and resized to 224×224 with Lanczos resampling filter to preserve the highest quality.

Training CNN model: The use of transfer learning is a method used frequently in CNN-based models due to their effectiveness in computer vision problems because they allow training at a faster pace, increase the model's generalization and counteract an unbalanced dataset [5]. Several CNN architectures with pre-trained weights derived from ImageNet were used in the due to their optimal training using a great number of labeled images [5]. From this, several state-of-art top-accuracy CNN architectures were chosen according to their architecture type [9] to evaluate their performance with our datasets i.e. VGG16 [10], MobileNet [11], ResNet50 [5], Inception V3 [12] and Inception-ResNet-V2 [13]. To fine-tune pre-trained CNNs according to our datasets, the original networks architectures were kept until the first fully-connected layer, which are the feature extractors, then the fully-connected layers are replaced with two other layers: a global average pooling layer and a last fully-connected layer with 33 outputs.

Every fine-tuned CNN is configured with cross-entropy loss function and Adam optimizer with a learning rate of .001. To evaluate the performance, the dataset was randomly split in: training (70%), validation (15%) and testing (15%) due to the limited quantity of images per class in our dataset. We trained the CNNs with 30 epochs using 9,619 images, for validation 2,050 images and finally we tested with 2,094 images.

IV. Related Works

To the best of our knowledge, prior works on sheet music classification are very limited, usually music is read from sheets using OMR techniques [14], which recognize music symbols and their semantics in order to capture the music digitally. Works using convolutional neural networks use full images for classification. In [15], a deep-learning algorithm based on a color-classification approach is proposed for the real-time analysis of endoscopic HSI (hyperspectral imaging) of human biopsies taken from the esophagus.

Transfer learning is a technique used in CNNs because they lead to a better generalization of data and faster training. In [9], they developed a CNN model using transfer learning for gastric lesions classifications by removing the fully-connected layers and the last classifier layer of a regular CNN, then adding two feature extraction layers for classification.

To develop a CNN model capable of recognizing sheet music books efficiently, we used several pre-trained CNN, similarly to [9], thus finding which one works better with sheet music. To evaluate the performance, we use accuracy for training [5] and for classification we use precision, recall and F1 which are the most common validation metrics [9].

V. Experiments

In this section we perform test to evaluate the performance of transfer-learning in a sheet music dataset. In the first one, the fine-tuned CNNs are evaluated with the original dataset. In the second one, each image was divided in 9 equally sized parts in our dataset and then it is used to train the fine-tuned CNNs. Finally, we compare the most efficient models.

Experimental Protocol: For the experiments performance evaluation, Accuracy is used for the training and validation sets and Accuracy, Precision, Recall and F1-score for the test dataset. The experiments were done in Google Colab Pro with Intel Xeon Quad-Core Processor and a GPU Tesla P100-PCIE-16GB using Python 3.6 with Keras and TensorFlow backend.

Experiment 1: In this experiment, the fine-tuned CNNs (VGG16, MobileNet, ResNet50, InceptionResNetV2 and InceptionV3) performance was evaluated using the original dataset. The classification accuracy and cross-entropy loss values for each model are presented in Table Ia. According to this table and Fig. 2, MobileNet training is the most efficient architecture with a CA and CEL of 99.96% and .0142 respectively. Next are InceptionV3, InceptionResnetV2, VGG16 and ResNet50 respectively. In the validation phase, MobileNet stays as the most efficient architecture with 97.56% CA and .0906 CEL. However, now VGG16 is in second place followed by InceptionResnetV2, InceptionV3 and ResNet5.

	Training		Validation	
Model	Accuracy	Loss	Accuracy	Loss
VGG16	.9631	.2214	.9682	.2036
ResNet50	.8558	.7209	.8394	.7424
MobileNet	**.9996**	**.0142**	**.9756**	**.0906**
InceptionResnetV2	.9775	.0986	.9473	.1991
InceptionV3	.9935	.0468	.9058	.2966

(a) Original dataset

	Training		Validation	
Model	Accuracy	Loss	Accuracy	Loss
VGG16	.9511	.1816	.9449	.1984
ResNet50	.7913	.7421	.7748	.7798
MobileNet	**.9826**	**.0509**	**.9513**	**.1839**
InceptionResnetV2	.9637	.1087	.9432	.1868
InceptionV3	.9436	.1634	.8959	.3681

(b) Split-9 dataset

TABLE I: Fine-tuned CNNs performance comparison using the a) original dataset and b) split-9 dataset

Model	Accuracy	Precision	Recall	F1
VGG16	.9608	.9692	.9566	.9608
ResNet50	.8428	.8481	.8255	.8259
MobileNet	**.9713**	**.9742**	**.9689**	**.9710**
InceptionResnetV2	.9379	.9487	.9315	.9366
InceptionV3	.9026	.9140	.9026	.9037

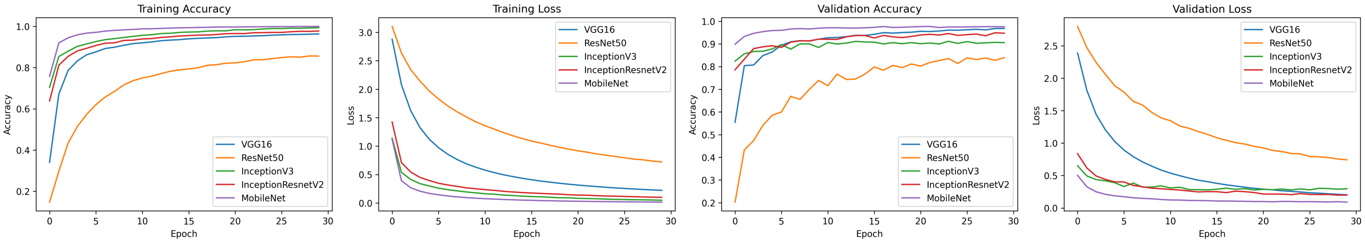
(a) Original dataset

Model	Accuracy	Precision	Recall	F1
VGG16	.9459	.9457	.9420	.9432
ResNet50	.7647	.7719	.7642	.7598
MobileNet	**.9460**	**.9527**	**.9379**	**.9435**
InceptionResnetV2	.9407	.9387	.9397	.9378
InceptionV3	.8920	.8843	.8959	.8884

(b) Split-9 dataset

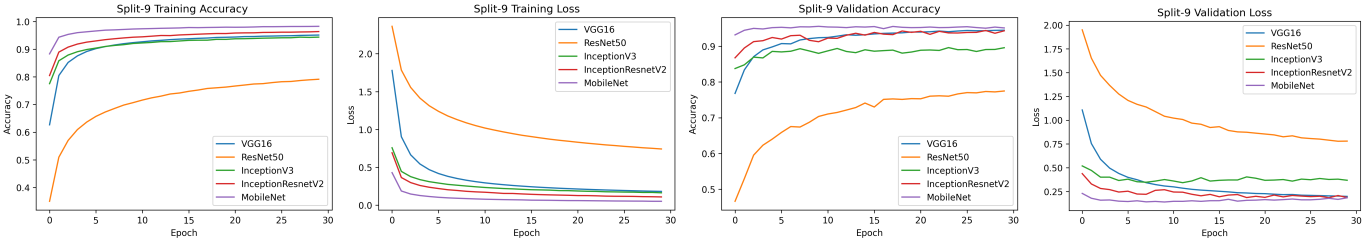
TABLE II: Fine-tuned CNNs metrics comparison using the a) original dataset and b) split-9 dataset

Experiment 2: For this experiment, we use the Split-9 dataset. The objective of this experiment is to test the fine-tuned CNNs performance with divided images. For this evaluation, we used the same five fine-tuned CNNs used in Experiment 1 (VGG16, MobileNet, ResNet50, Inception-ResNet-V2 and InceptionV3) with the same configuration. With this dataset, according to Table Ib and Fig. 3, in training, MobileNet provided the best efficiency with a CA and CEL values of 98.26% and .0509 respectively, followed by InceptionResnetV2, VGG16,



(a) CA on training data (b) CEL of training data (c) CA on validation data (d) CEL of validation data

Fig. 2: Results on training and validation data for the original dataset.



(a) CA on training data (b) CEL of training data (c) CA on validation data (d) CEL of validation data

Fig. 3: Results on training and validation data for the split-9 dataset.

InceptionV3 and ResNet5. For the validation phase, MobileNet is at first place again with 95.13% accuracy and .1839 loss.

Results: For the fine-tuned CNN performance evaluation with both datasets, we compared the metrics between them using test data. As observed in Table IIa MobileNet achieves the top 97.13% accuracy, 97.42% in Precision, 96.89% in Recall and 97.10% in F1-Score. And, as seen in Table IIb for the Split-9 dataset, MobileNet also obtains the highest values on Accuracy, Precision and F1-Score with 94.60%, 95.27% and 94.35% respectively; but, VGG16 has the best Recall with 94.20%. As mentioned before, in both experiments MobileNet is the most effective fine-tuned architecture for both datasets. However, when comparing the performance, it gets better result with the original dataset than the Split-9 dataset.

VI. Conclusion

Our work relies on pre-trained CNN to test their performance in sheet music classification, thus extracting features and classify them based on their content. These have proven to be very effective, MobileNet having the best results. Nevertheless, with split images, the models are slightly less effective but still with high values being MobileNet the best overall.

An interesting extension is applying K-fold Cross-Validation for fine-tuned CNNs performance evaluation. Additionally, a larger dataset with more variety of images (e.g., handwritten sheets) may lead us to further training these fine-tuned CNNs, similarly to optimization in itemset mining [16], [17], [18].

References

[1] A. Latif, A. Rasheed, U. Sajid, A. Jameel, N. Ali, N. I. Ratyal, B. Zafar, S. Dar, M. Sajid, and T. Khalil, "Content-based image retrieval and feature extraction: A comprehensive review," *Math. Prob.in Eng.*, 2019.

[2] S. Fadaei, R. Amirfattahi, and M. R. Ahmadzadeh, "Local derivative radial patterns: A new texture descriptor for content-based image retrieval," *Signal Process.*, 2017.

[3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Underst.*, 2008.

[4] S. Leutenegger, M. Chli, and R. Siegwart, "BRISK: binary robust invariant scalable keypoints," in *ICCV*, 2011.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[6] J. Patterson and A. Gibson, *Deep Learning: A Practitioner's Approach*, 1st ed. O'Reilly Media, Inc., 2017.

[7] R. C. González and R. E. Woods, *Digital image processing, 3rd Edition*. Pearson Education, 2008.

[8] T. Jin and S. Hong, "Split-cnn: Splitting window-based operations in convolutional neural networks for memory system optimization," in *ACM ASPLOS*, 2019.

[9] X. Liu, C. Wang, J. Bai, and G. Liao, "Fine-tuning pre-trained convolutional neural networks for gastric precancerous disease classification on magnification narrow-band imaging images," *Neurocomp.*, 2020.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[11] D. Sinha and M. El-Sharkawy, "Thin mobilenet: An enhanced mobilenet architecture," in *UEMCON*, 2019.

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[13] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *AAAI*, 2017.

[14] L. Mengarelli, B. Kostiuk, J. G. Vitório, M. A. Tibola, W. Wolff, and C. N. Silla, "OMR metrics and evaluation: a systematic review," *Multim. Tools Appl.*, 2020.

[15] A. Grigoroiu, J. Yoon, and S. Bohndiek, "Deep learning applied to hyperspectral endoscopy for online spectral classification," *S.R.*, 2020.

[16] W. Ugarte, P. Boizumault, S. Loudni, B. Crémilleux, and A. Lepailleur, "Soft constraints for pattern mining," *J. I. Inf. Syst.*, vol. 44, no. 2, 2015.

[17] W. Ugarte, P. Boizumault, S. Loudni, B. Crémilleux, and A. Lepailleur, *Mining (Soft-) Skypatterns Using Constraint Programming*. ADKM, 2016, vol. 5.

[18] W. Ugarte, P. Boizumault, S. Loudni, and B. Crémilleux, "Modeling and mining optimal patterns using dynamic CSP," in *ICTAI*, 2015.