2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

An Analysis of Multi-Modal Deep Learning for Art Price Appraisal

Bin Li* and Teng Liu†
*Department of Computer Science
Columbia University, New York, New York 10027
Email: bl2899@columbia.edu
†160 Morgan Street, Jersey City, New Jersey 07302
Email: tengliu2@gmail.com

***Abstract*—This study evaluates the predictability of art prices using deep neural network models of different modalities. We experiment with Bidirectional Encoder Representations from Transformers (BERT) model for textual analysis and Convolutional Neural Network (CNN) with and without Siamese network structure for image analysis. We find that the societal mechanism of fine-art evaluation is better captured in the textual dimension than visual representation, as the former provides a richer account of cultural-historical information. Specifically, BERT, after fine-tuning, performs significantly better than CNNs on predicting the price range of artwork. The CNN with Siamese architecture does not perform better than the one without. Nonetheless, the post hoc regression analysis shows that the features extracted from the Siamese CNN complement BERT's features to provide the best model fit, while those from the regular CNN do not have the same positive contribution to the model fitness. The results provide novel insights into the cross-modal mechanism underlying the process of aesthetic evaluation in today's art world.**

1. Introduction

Art, especially the appreciation of art, has become an increasingly cryptic subject in the modern world. An art piece whose immediate aesthetic value is obscure to laymen can be sold for a jaw-dropping price, while many paintings that nonprofessionals find beautiful are consistently undervalued by fine art collectors. This topic has been the center of an endless debate on the essence of art [1], [2], [3], [4].

It seems natural to presume that the image of artwork should contain more information than the text regarding the pricing and appraisal of the art. Undoubtedly, a good number of people and art critics believe that art is related to the judgment of beauty and beauty is in appearance, which contributes to the form and the formalism of the art [5]. However, in today's art world, what is deemed beautiful is not necessarily the most sought-after. Oftentimes the social context behind an artwork plays a significant role in deciding its value [3], and this kind of information is arguably better encapsulated in text, a natural vehicle for carrying cultural-historical context.

The aesthetic demarcation in the art world offers a unique perspective to examine the complex social factors that underlie modern culture. This analysis aims to understand the dynamic between the two competing sources of aesthetic evaluation: the historic-textual and visual information.

For this study, we collected art auction data from the website of a widely known auction house. The dataset contains information about the artworks sold through the agency in recent years, including the names of the artworks, makers' names, descriptions, and the images of the artworks.

A two-pronged analysis is conducted to approach this problem.

Firstly, we are interested in the high-dimensional representations of artwork information in different modalities. To that end, we build two groups of deep neural network models that take in as input the textual content (artwork description) and the images of the artwork, respectively. For textual input, we use a novel neural network model—Bidirectional Encoder Representations from Transformers, also known as BERT [6]. For image input, we use two Convolutional Neural Networks (CNN) models, one with Siamese network structure [7] and one without. The models are trained to classify an artwork into a particular price category, and the classification accuracy is measured and compared.

Secondly, we are also interested in comparing the learned features of different modalities and evaluating their contribution to the overall prediction. To achieve this, we extract the transformed features from the trained models, reduce the feature dimensionality into a more computationally manageable range, and use them to fit several regression models. The models' goodness-of-fit are then measured and compared.

With this study, we contribute to the existing discussion on the social mechanism of art price appraisal by

- reaffirming the importance of socio-historical context in the form of textual information for appraising art price,
- revealing the multi-modal nature of the mechanism that underlies art price appraisal,

978-0-7381-2646-3/21/$31.00 ©2021 IEEE

DOI 10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00203

- providing some novel uses of computational methods for address similar social science issues.

The rest of the paper is organized as follows. In Section 2, we discuss the limitations of previous work and the novelty of our research. In Section 3, we discuss the architectures and training setup of the neural network models in use, followed by a specification of the regression analysis used for validating the results. In Section 4, we present the results of our experiment and our interpretations of the results. We conclude our study in Section 5 with a discussion on the importance of our findings.

2. Related Work

The argument over the source of aesthetic value has a long-standing history in philosophy, but there have not been many studies that address this issue using computational methods.

Much of the previous research efforts have been focusing on the economic phenomena in the art auction market [8], [9], with little or no interest in the origin of the aesthetic value. While those approaches have generated some fine-grained and highly predictive models, they rarely provide any insight into the mechanism of aesthetic evaluation beyond the social-historical setting.

In recent years, there have been some studies investigating this topic using state-of-the-art machine learning methods. In [10], a multi-task model is developed to assist the categorical analysis of artworks via artist attribution, creation date estimation, type prediction, etc. Ayub et al. [11] implement an unsuccessful end-to-end CNN to predict art price based on image, while Worth [12] improves upon this approach by combining a Siamese CNN with an LSTM model to achieve a better model performance. However, these studies only focus on the visual domain of the valuation issue. Audry et al. [13] address this problem by combining artwork data from two modalities into a single neural network model to achieve state-of-the-art accuracy over simple hedonic regression model and other machine learning models. However, the study fails to tease apart the respective contributions of textual and visual information.

3. Methods

Our analysis has two parts. First, we train different neural network models on a classification task and compare the models' validation accuracy. In the classification task, the models predict which price category an artwork belongs to. Second, we compare and contrast the models' learned features by building regression models using different extracted features and comparing model performance.

3.1. Model Specifications

Models are trained on a single Nvidia GTX 1080ti graphic card with 11GB memory. The training of BERT takes roughly two hours, while the CNN and the Siamese each takes around an hour. After model convergence, we report the model with the best validation performance and use it for the later analysis.

In this study, we adopt transfer learning in training the models. Transfer learning is a machine learning technique where a model initially developed for a task is reappropriated on a second task by fine-tuning the pre-trained model on another dataset [14]. Both BERT and our CNN models capitalize on some pre-trained models that are readapted for our particular task.

For all models, we use cross-entropy loss L when training, which is defined as,

$$L = -\sum_i p_i log(q_i), i = 1, ..., n$$

where n is the total number of classes, $log(q_i)$ is the model output prediction head with log probability of belonging to class i, and $p_i \in \{1, 0\}$ is the probability of an item actually falls into class i. No softmax activation function is used in any of the three models for calculating the output prediction probability q_i; instead, the output prediction head is directly interpreted as log probability to increase numerical stability during computation.

For training optimization, Adam optimizer, short for adaptive moment estimation [15], is used for all three models.

3.1.1. BERT. BERT is an instance of a class of transformer-based neural network models that utilizes the attention mechanism to achieve contextual understanding of natural language. The model architecture description is omitted for brevity, as the details can be found in Devlin et al. [6].

In this study, we fine-tune a BERT model with 12 transformer blocks and 768 hidden states (i.e., a BERT-base model) for the price category classification task.

The pre-trained model is obtained from the *BertForSequenceClassification* class of the *Transformers* python library. The text tokenization is done in the same procedure described in [6]. A helper program is conveniently implemented by the *AutoTokenizer* class of the *Transformers* library and is used for this purpose.

The model is implemented in PyTorch 1.7. We trained the model for two epochs with a batch size of 8, a learning rate of $2e^{-5}$, and the epsilon of $1e^{-8}$.

3.1.2. CNN. The CNN model is built on top of a pre-trained backbone, the ResNet50 [16], with some modifications (see Figure 1).

The input image was resized to 128 by 128 pixels with three color channels (RGB). Then the color value is normalized to $(-1, 1)$, resulting in a tensor of shape $(128, 128, 3)$. For modification, the original prediction layer of ResNet50 was removed, and a new layer was appended after the curtailed ResNet50 via global max pooling to extract relevant features. A new prediction head with the price label is added to the end.

When we first start the training, we freeze the weights of the backbone and train the newly added layer with the

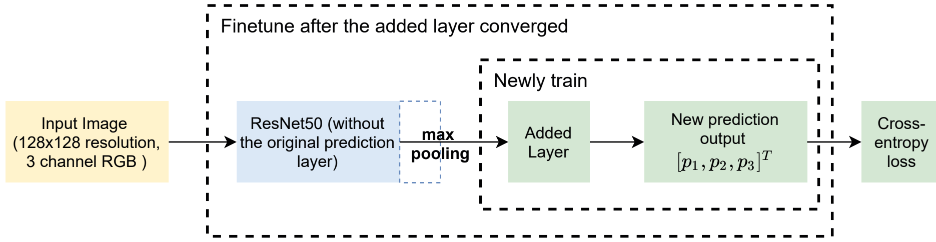

Figure 1. Visualization of ResNet50-based CNN's network structure.

learning rate of $1e^{-5}$ and a batch size of 32. After that, the whole model is fine-tuned from end to end at a lower learning rate. We use dropout regularization to prevent overfitting, with the dropout rate set to 0.2. The whole training process takes 32 epochs to converge.

3.1.3. Siamese CNN. The Siamese neural network, also known as the twin neural network, is an artificial neural network that uses the same weights while working in tandem on a pair of input vectors to learn the similarity/dissimilarity measure of the inputs.

A Siamese network usually receives two inputs and passes the inputs through the same network. The outputs are compared, for example, by taking the absolute difference between the tensors. The network then predicts whether or not the two inputs belong to the same class. This modification makes the network effective in learning the representation of the similarity measure between different inputs in accordance with the differences of their classification labels.

Here we take advantage of the Siamese network structure to learn the similarity between the images of artworks in relation to their difference in price (see Figure 2 for the full network structure).

Similar to the regular CNN model above, we use ResNet50 as the backbone for the Siamese network and follow the same preprocessing steps. For input, we feed two images into the same network to obtain two extracted feature vectors x_1 and x_2.

We then use three similarity/difference measures, the square of absolute difference, the difference of squared value, and the cosine similarity, to measure the distance between x_1 and x_2. The results are concatenated into a single feature vector.

The vector is then passed through an extra fully connected layer with dense representation and the Relu activation. Dropout is applied to the fully connected layer to regularize the model before generating the classification probability vector. Please note that the classification probability vector is different from the previous models. Here the output label represents the difference between inputs' price tags, indicating whether input 1 is more or less expensive than input 2.

3.2. Regression Analysis

This analysis aims to evaluate the usefulness of different network features by using them to build a regression model to predict art price as a continuous variable.

Here we compare models in a nested fashion; we fit a model using features from an individual network and fit a new one by combining features from different networks. Then we report and compare the fitted models' R^2 score, a widely used measure of the goodness-of-fit of regression models.

To extract word features from the BERT model, we combine the last four output hidden states into a single vector per suggestion from the ablation study in [6]. Due to the constraint of the computational power, we are limited to only use the first 100 words for extraction. For CNN and Siamese CNN, the features are extracted from the last layer before the prediction head.

Before using the features for regression analysis, the dimensionality of the features is reduced using Principal Component Analysis (PCA). For computational efficiency, the target size of the reduction is set to half of the size of features with the smallest dimension. We also include the length of artwork description in the regression as a nuisance regressor to disambiguate any potential effect from this kind of shallow descriptive statistics.

With the use of PCA, the regression analysis in using features from a single network is equivalent to conducting a Principal Component Regression (PCR). For the combined models, ridge regression, i.e., linear regression with L2 regularization, is used to prevent overfitting; the weight on the regularization penalty term is determined using 10-fold cross-validation.

3.3. Experiment and Dataset Descriptions

The data used in this study are collected from a website of a publicly known auction house. Each auctioned entry is numbered and categorized under the auction event where it was sold. An information retrieval crawler program is written to query the auction event and item number successively by numerical order to find any potential entry in the database. As the company does not provide any open-source API for data collection, the virtual private network and IP switching technology are used to circumvent any restrictions and guarantee effective data retrieval.

The resulting dataset contains more than 100,000 artwork entries from auction events spanning two decades. The recorded variables include auction date, price, artwork's name, description, transaction currency type, artwork's im-

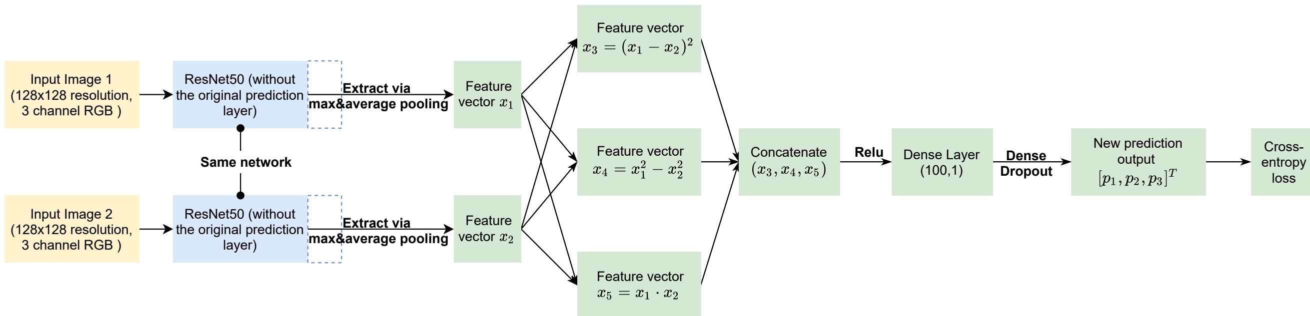

Figure 2. Visualization of Siamese CNN's network structure.

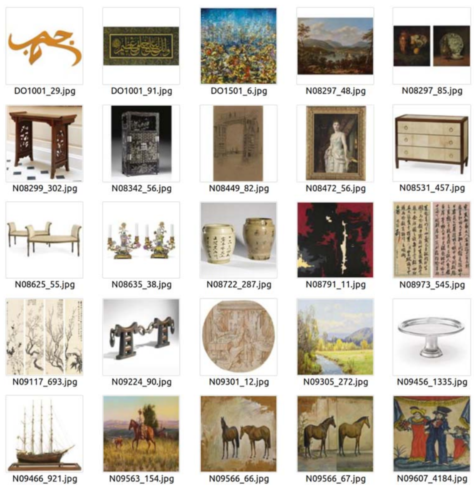

Figure 3. Some images in the dataset.

age (Figure 3), and others. The resulting dataset is further processed by performing the following steps:

- All entries whose transaction currency is not in US Dollar are removed. These include artworks that have auction price in different currency units or description in non-English languages, making them unsuitable for the analysis;
- Price unit is converted to the logarithmic unit, which yields a more normalized price distribution;
- Price is discretized into three categories: affordable (priced at less than \$8000), expensive (less than \$100,000), and luxury(\$100,000 and more). The price cut point is determined by a k-means classification algorithm with $k = 3$ using Python scikit-learn's discretization module;
- Special characters and symbols in the description text are removed. Alphabetic characters are cast to lower case;

After the above preprocessing steps, the dataset was left with more than 50,000 entries. The dataset was further split into training, validation, and test set, each containing roughly 40,000 data points, 4,000 data points, and 5,000 data points, respectively. The ratio of different price bins in each split was balanced to allow for a more straightforward comparison of model performance using flat accuracy.

TABLE 1. MODEL PERFORMANCE COMPARISON

Model	Validation loss	Validation accuracy
CNN	0.96	0.52
Siamese	1.06	0.46
BERT	**0.66**	**0.70**

4. Results

4.1. Performance Comparison

As seen in Table 1, the fine-tuned BERT performs significantly better than both CNN models on the validation set. The performance of the more complex Siamese CNN model is worse than the regular CNN.

Our analysis shows that BERT performs significantly better than CNN models on the price category prediction task. The results lend some credibility to the claim that the appearance of an artwork is not as crucial as the socio-historical narrative behind that artwork when it comes to the pricing mechanism. This finding may not be too surprising as some previous works [11], [17] have shown that art, in general, cannot be successfully appraised using convolutional neural networks on just the raw image data representing the artworks. Nonetheless, the result does not rule out all the possible roles that visual information might play in aesthetic evaluation.

It is also noteworthy that the regular CNN outperformed the more complex Siamese CNN model. This may come as a surprise as a recent finding reports favorable results from training Siamese CNN to perform a similar task [12], [18]. This is another lesson or platitude for the neural network power-seekers—a complex model is not always better than a simple one.

However, we should also note that, although ResNet50 is not a simple model per se, BERT is much bigger than the two CNNs in terms of size and number of parameters. The image data we collected also represents a conceptually wider array of variety. The dataset contains painting, sculpture,

TABLE 2. REGRESSION ANALYSIS

Source of Features	Method	Hyperparameter	R^2
CNN	PCR	$PC = 1024$	$< -3e^9$
Siamese	PCR	$PC = 1024$	0.512
BERT	PCR	$PC = 1024$	**0.676**
CNN + Siamese	Ridge	$\lambda = 1e^5$	0.497
CNN + BERT	Ridge	$\lambda = 1e^5$	0.682
Siamese + BERT	Ridge	$\lambda = 6e^4$	**0.705**
CNN + Siamese + BERT	Ridge	$\lambda = 1e^5$	0.676

ready-made, furniture, cultural relics, and artwork of other forms. The image collection in our dataset may be too divergent and the training size too small for the CNN to demonstrate its power of object detection. On the flip side, the diversity of visual representation also reveals a major limitation of the image-based solution. Because art comes in different forms, it can be challenging to train an image-based machine learning model to skillfully discern all the different forms in relation to their prices.

4.2. Regression Results

The results of the regression analysis in Table 2 provides a closer look into the model performance. Although using BERT features still yields the best single-network R^2 score, combining BERT features with those from the Siamese network results in the best model overall. Combining with features from the CNN contributes little to or sometimes even hurts the regression prediction.

From the post hoc regression analysis, we find that the features extracted from the Siamese network are much more useful than those from the regular CNN, despite the latter achieving better accuracy. Features from the Siamese CNN also complement BERT Features better than the regular one.

5. Conclusion

Our analysis finds that BERT outperforms CNNs on the price category prediction task, and the more complex CNN does not perform any better than the simple one on this task. These results suggest that an image-based solution might not be as effective as a text-based one when it comes to automatic art price prediction, as the latter can capitalize the rich socio-historical context of the artwork, such as artist background and artwork provenance, which comes in the form of text.

The results also challenge the common conception that art is mainly concern with the beauty of appearance. It seems that when it comes to appraising art for its monetary price, socio-historical background plays a more prominent role than pure looks.

Our findings do not entirely rule out the role that visual information might play in aesthetic evaluation. On the contrary, the regression analysis reveals that images can carry modality-exclusively information useful for art price prediction. An image-based model like CNN can capture this kind of information if the model is carefully attuned to focus on the relevant information. With its ability to generalize the similarity measure of input space, the Siamese CNN can be an ideal candidate to be incorporated into a cross-modal model for automatic art price prediction.

References

[1] A. C. Danto, "V. the end of art," in *The philosophical disenfranchisement of art*. Columbia University Press, 2005, pp. 81–116.

[2] ——, "Art, essence, history, and beauty: A reply to carrier, a response to higgins," *The Journal of Aesthetics and Art Criticism*, vol. 54, no. 3, pp. 284–287, 1996.

[3] P. Bourdieu, *Distinction a social critique of the judgement of taste*. Routledge, 2018.

[4] R. Lind, "The aesthetic essence of art," *The Journal of Aesthetics and Art Criticism*, vol. 50, no. 2, pp. 117–129, 1992.

[5] C. Greenberg, "Necessity of "formalism"," *New Literary History*, vol. 3, no. 1, pp. 171–175, 1971.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[7] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *Advances in neural information processing systems*, vol. 6, pp. 737–744, 1993.

[8] O. Ashenfelter and K. Graddy, "Auctions and the price of art," *Journal of Economic Literature*, vol. 41, no. 3, pp. 763–787, 2003.

[9] J. W. Galbraith and D. J. Hodgson, "Econometric fine art valuation by combining hedonic and repeat-sales information," *Econometrics*, vol. 6, no. 3, 2018.

[10] G. Strezoski and M. Worring, "Omniart: Multi-task deep learning for artistic data analysis," *CoRR*, vol. abs/1708.00684, 2017.

[11] R. Ayub, C. Orban, and V. Mukund, "Art appraisal using convolutional neural networks," 2017.

[12] T. Worth, "Painting2auction: Art price prediction with a siamese cnn and lstm," 2020.

[13] M. Aubry, R. Kraeussl, G. Manso, C. Spaenjers *et al.*, "Machines and masterpieces: Predicting prices in the art auction market," 2019.

[14] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[17] J. Bailey, "Can machine learning predict the price of art at auction?" *Harvard Data Science Review*, 2020.

[18] R. Bai, J. C. Lam, and V. O. Li, "Siamese-like convolutional neural network for fine-grained income estimation of developed economies," *IEEE Access*, vol. 8, pp. 162 533–162 547, 2020.