Rate Against the Machine: Image Based Etsy Price Estimation

Jeremy Warner, Zining Wang, and Jules Pommies
{jeremy.warner, wangzining, jules_pommies}@berkeley.edu
University of California, Berkeley
Computer Vision Final Project

Figure 1: *A painting from Etsy.com*

Abstract

Estimating the art prices inside the Etsy art community is challenging due to both the subjective assessment and quantitative differences in the art. Real-world art has surface textures, geometry, lighting conditions, and shapes which combine to make the problem particularly difficult. In this paper, we (1) introduce a model to process the Etsy website data, (2) combine this dataset with deep learning to build a large-scale price estimation system and (3) provide a web service to get access to it (etsyprice.com). Using many images scraped from Etsy, we train convolutional neural networks (CNNs) for two tasks: classifying images from Etsy, and beyond images, we provide a text encoder to use and refine our model with every item's description. For price estimation, we combine a CNN architecture and a Random Forest Regression achieving an 81.6% accuracy. After tuning and refining our price estimation model, we used it to develop a web-service that enables anybody to get an estimation of her/his item before publishing an offer on Etsy, or simply to compare their art with others. This service provides links back to the most similar items in the scraped Etsy database to get an idea of the validity of the prediction.

1 Introduction

This project aims to estimate the value of an artist's submission by comparing to it to relevant artwork through a neural network trained on images and background data scraped off of Etsy, a popular platform for selling art that independent artists and larger collectives both employ. It recognizes that the value of art is somewhat subjective but nonetheless seeks to profile some quantitative measures of the collective value placed on art by the artists themselves, though therein lies a flaw of the project: since we do not have actual sale data, we only know the artists' subjective value of their work. We do not have access to artwork which was sold, so the agreed-upon price for the creator and buyer is still hidden. This value would paint a truer picture of what the accepted value of a piece of art is.

2 Related Work

Convolutional Neural Network

Convolutional neural networks. While CNNs have been around for a few decades, with early successes such as LeNet, they have only recently led to state-of-the art results in object classification and detection, leading to enormous progress. Driven by the Pr. Alyosha Efros Computer Vision class, we have seen many successful CNN architectures, led by the work of Krizhevsky et al. on their SuperVision (a.k.a. AlexNet) network, with more recent architectures including GoogLeNet. In addition to image classification, CNNs are the state-of-the-art for detection and localization of objects, with recent work including R-CNNs, Overfeat, and VGG. We build on this body of work in deep learning to solve our problem of price estimation and art classification.

Random Forest Predictor

Breiman's random forest relies on Bagging and Randomization techniques, as the main idea is to train many classification decision trees with the largest extent possible without pruning. Random Forest is used in several cases but the most popular situations when it is especially attractive are the following:

(1) First, when the real world data are noisy and might contain many missing values, or when some of the features are categorical, or semi-continuous.

(2) When integrating different data sources which face the issue of weighting them is a main aspect of the project.

(3) Problems with a high number of dimensions with highly correlated features.

Based on those previous body of work we decided to combine both the RandomForest and the CNN architecture in order to get information from and process the pieces of art from `etsy.com`.

3 Learning the Price of the Art

Artworks are unlike other industrial products, which are very likely to have comparable attributes with other products and a

corresponding reasonable price. Artworks are unique and their prices are usually depending on many subjective aspects like author's reputation, appearance, and design. Images must be a significant factor in the price of artworks like paintings and accessories. The recent development of deep convolutional neural networks [He et al. 2016] and large image datasets [Russakovsky et al. 2015; Krizhevsky et al. 2012a] have pushed the capacity and flexibility of visual perception to a next stage. Object detection and classification focus on learning visual perceptions that are certain and invariant among humans, which means the dataset is less noisy and variant in terms of labeling. In this work, the network is supposed to learn artistic perception from ambiguous price labeling. The dataset is much noisier and the pattern is not even clear to people, which makes the learning difficult.

On the other hand, learning to regress artwork with only images can be misleading. Pictures that are drawn by hand and taken by cameras clearly have different values in most cases, but it is even hard for a people to tell a painting from a photo if it is painted realistically. Moreover, it is almost impossible to differ a printed painting from hand-made painting just from images, not to say similar artworks from authors with different reputations usually have different prices. More information is needed to assess the value of an artwork.

In this work, both image and text descriptions are extracted from Etsy. Two datasets are built for comparison, paintings, and jewelry respectively. The deep convolutional neural network is implemented to process the image and other classic learning structures are utilized to process the text description. In order to obtain the effect of these two features, image and text, to the price regression, predictions are made with each feature separately as baselines. The results are compared with the learning method that merges both features. The influence of images and text are demonstrated through the comparison.

3.1 Regression with Text Descriptions

Etsy provides an `Overview` section for the sellers to describe their items. There are common categories like reviews, favorites, and materials for each artwork. As mentioned before, the reputation of the author is an important factor and thus is taken as a categorical feature. Reviews and favorites are one measure of the popularity of the product and should also be related to the price. Materials contain the composites of the item as well as how the artwork is produced, such as painted, hand-made, printed and manufactured. Sellers can also write paragraphs to describe their items. Here only 'reviews', 'author' and 'materials' are taken into account as text features. Natural language processing can be applied to the long paragraph for more informative details but is not implemented due to time constraint.

3.1.1 Encoding Text Descriptions

The text features have different properties and are pre-processed as follows:

- 'reviews' is treated as a continuous feature here. This feature only takes one dimension and is normalized to the origin and scaled to unit variance for learning.
- 'author' is a categorical discrete feature. An artwork can only have one author, so it is transformed to a binary vector by one-hot encoding.
- 'materials' is also a categorical feature. It is different from 'author' in that an item can have multiple materials. So the 'materials' is transformed to a binary vector where each bit represents a material. If the item has one material, the corresponding bit is set as 1 otherwise 0.

There are numerous kinds of materials and authors on Etsy, approximately of the same scale as the size of the dataset. It turns out that most art we scraped are owned by a small set of authors and only these popular authors have significant influence to their products. Similar things happen for the materials used in the art. As a result, we only choose those dominant authors and materials as unique categories. Others are labeled as 'unknown' and treated as one category in the feature.

The first 1000 authors and 600 materials are chosen and each has an additional 'unknown' class. The total vector size is 1603 for the text feature.

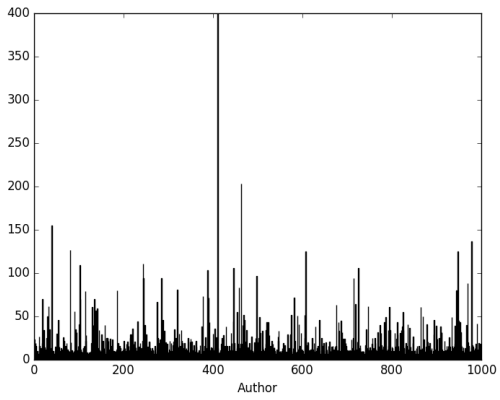

Figure 2: *The distribution of authors.*

3.1.2 Random Forest Regressor for Text

A random forest regressor is built to predict price from text. The prices are logged and the regressor minimizes the mean-square loss.

3.2 Regression with Image using Convolutional Neural Network

Each item on Etsy is associated with one image. The images are first padded to square with zeros and then resized bilinearly to a fixed size. No stretching for shrinking is performed so as to keep the original appearance of the original painting/jewelry. Since color, lightness, and integrity are important for an artwork, random data augmentation are not applied to images.

The AlexNet [Krizhevsky et al. 2012b] is utilized to process the images. However, AlexNet is originally designated for classification of objects. In order to regress the price, the inspiration from Faster-RCNN [Ren et al. 2015] on bounding box regression is taken. The prices are partitioned into several coarse intervals logarithmically between \$0.1 to \$100,000. The network first produces probabilities of the price intervals, then another fully connected layer is used to predict the shift from the center of the price interval.

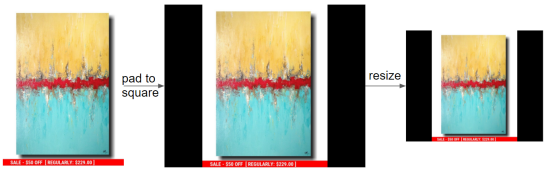

Figure 3: *Image padding and resizing*

3.3 Merging the Text and Image Features

The value of artworks cannot be determined as accurately by just images alone. As mentioned above, images need text for more accurate information. Learning by merging image and text features is performed in order to achieve best performance and further analyze the contribution of these features. Two approaches are designed to merge the image and text information after the image is encoded into a fixed length vector.

3.3.1 Neural Network Approach

This approach uses a end-to-end learning structure modified from the AlexNet's regression. The image is first processed by the convolution layers and transformed to a vector after the first fully connected layer. Then, the raw text feature is fed into the network and concatenated with the processed image feature. Then more fully connected layers are used to process the merged data and generate the predicted price in the same way as described in the image only regression. The weights of the convolution layers are pre-trained in

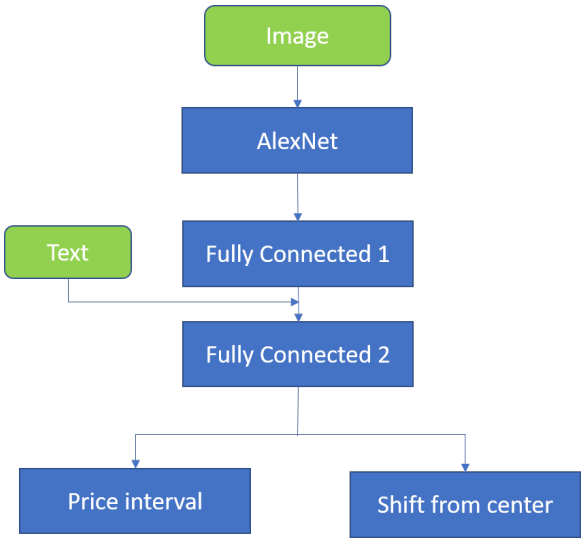

Figure 4: *Network structure for merging*

the image only regressor. The fully connected layers are initialized randomly.

3.3.2 Random Forest Approach

The image only network serves as a good encoder for the image. It is observed that the random forest is better at utilizing encoded text features compared to fully connected layers in practice. Hence another approach for merging both features is to augment the text features with the output from some fully connected layers in the image only network. The augmented input is then feed to the random forest for regression. The shortcoming of this approach is that the

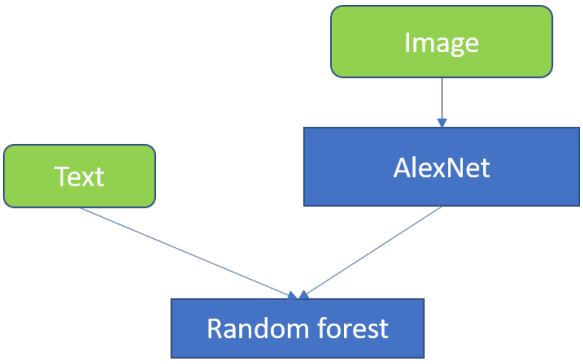

Figure 5: *Augmented text feature and random forest*

CNN is not back-propagated. Only the random forest is trained.

4 Implementation and Evaluation

A total of 94,000 items are extracted from Etsy with a scraper, including paintings and jewelries. After pre-processing and discard those items whose author and materials are both 'unknown'. The paintings is split into 35280 training and 11910 test samples. The jewelry is split into 29005 training and 9836 test samples.

The network is trained with TensorFlow and TITAN X gpu. The performance is evaluated by the prediction accuracy - error threshold curve. The error is defined as

$$\text{error} = \frac{\text{Predicted Price - True Price}}{\text{True Price}}$$

which means the percentage deviation from actual price. Accuracy is calculated given different error threshold and the curve is plotted accordingly.

5 Results and Discussion

The price of the arts sold on Etsy, though subjective, has a strong correlation with some objective features presented in the dataset. Since the price is proposed by sellers instead of being the actual market price, it is related to the author who sells the art. However, review numbers, as another feature in the text description, reflects the coherence between the seller's will and the actual demand of the market.

5.1 Single Feature Performance

The regression results with only the text or image feature is compared in 5.1. The accuracy-error curve is shown in 6 and 7. It is noticeable that the jewelry has lower performance using only the image feature. This is because of jewelries usually only occupies a small part of the image and there are a lot of distractions like noses, ears and boxes that affects the regression. The text seems to be able provide more reliable information on the price.

accuracy at error $< 25\%$		Jewelry	painting
text only	train	58.2%	56.1%
	test	47.9%	46.4%
image only	train	40.5%	58.1%
	test	27.2%	32.2%

One of the explanations that can justify the poor performances of the "Image only' over the "Text only" model is the importance of

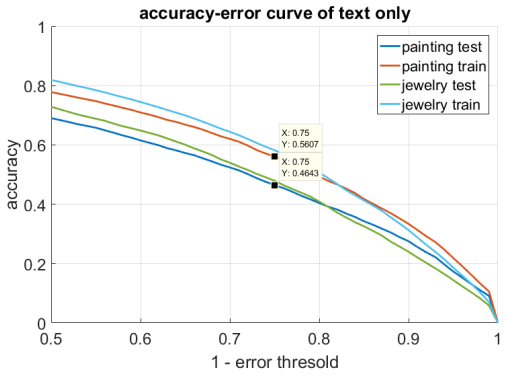

Figure 6: *Single feature performance of text*

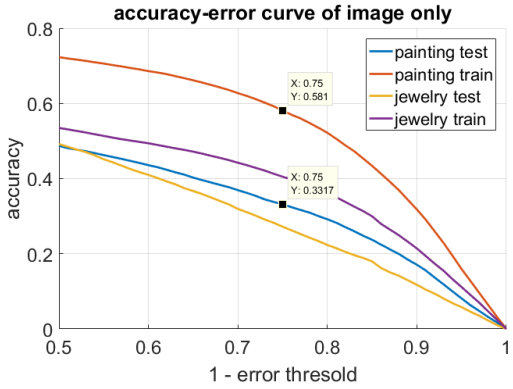

Figure 7: *Single feature performance of images*

the author feature. Indeed, it seems that every author is trying to sell all his/her work around the same price 8.

5.2 The Effect of Merging

Due to the limit of time, the merging methods are only applied to the paintings dataset. 5.2 shows the comparison of accuracy when the deviation from actual price is required to be smaller than 25%. The accuracy-error curve is shown in 9. It can be seen from the table that there is some improvement after merging the text to the CNN network. However, it is observed that even if the text input is replaced all by 'unknown', the accuracy of the merged AlexNet will only decrease by 1-2%. The network only utilizes little of the text features. That is why the second merged random forest structure is proposed. It can be seen that given the same image feature, the random forest takes more text feature into account and thus has further improvement in accuracy.

accuracy at error $< 25\%$		painting	
text only	train/test	56.1%	46.4%
image only	train/test	58.1%	33.2%
Merged AlexNet	train/test	71.1%	38.1%
Merged random forest	train/test	81.6%	40.8%

The regressions with image feature involved on the painting dataset has a significant sign amount of over-fitting. The effect is caused by multiple reasons. The main reason might be the design of the regression structure for the CNN. The loss and accuracy curve during the training is shown in 10. It can be seen that the accuracy of the classification part does not have a large gap between the training set and validation set. So the shifting part is not learned very well so

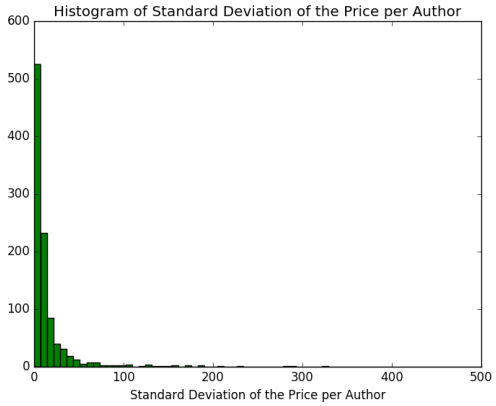

Figure 8: *Histogram of Standard deviation in price per author for paintings*

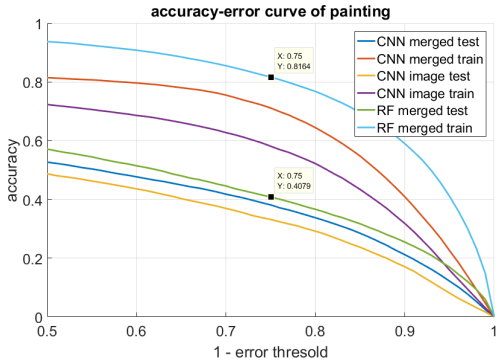

Figure 9: *Performance of the merge structure on paintings*

as to cause a big gap in the price prediction.

The over-fitting of the image feature also affects the merged random forest. Although text feature along can produce higher test accuracy, the over-fitting on the image causes the image features dominating the prediction of the prediction. The prediction accuracy is very high on the training set after merging text but it does not outperform the text only prediction on the test set. Image features with different feature lengths, namely 4096 and 256 are tested in the merged random forest structure and similar results are seen, which

5.3 Website for Prediction (etsyprice.com)

Based on the previous results, we decided to create a website, where new or old members of the Etsy community can upload a picture on their art and get an estimation of its price, as seen in Figure 12. This service also provides several pictures and links to the closest pieces of art in our scraped Etsy database to help the user appreciate the accuracy or not of the prediction. One example of this website in action can be found at this link, comparing sailboats: http://etsyprice.com/val/114196264842115.

One feature we could add to this website would be for artists to further classify their artwork to improve the accuracy of the prediction. This could be. done by having them signing into Etsy and importing data about the past items they have sold and their ratings. Another improvement would be to allow for the specification of the art's the material, size, or format.

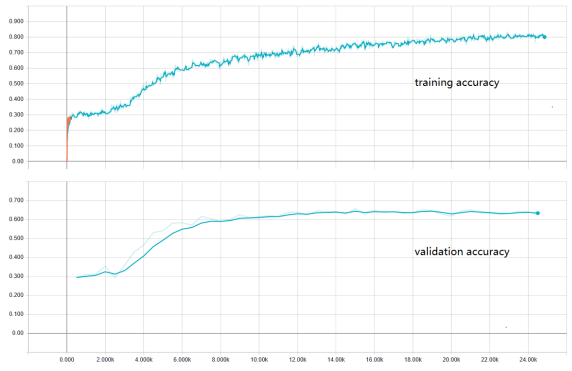

Figure 10: *Accuracy of the classification part during training of the merged AlexNet*

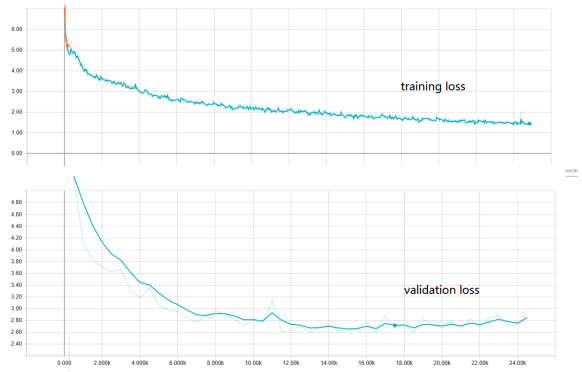

Figure 11: *Overall loss during training of the merged AlexNet*

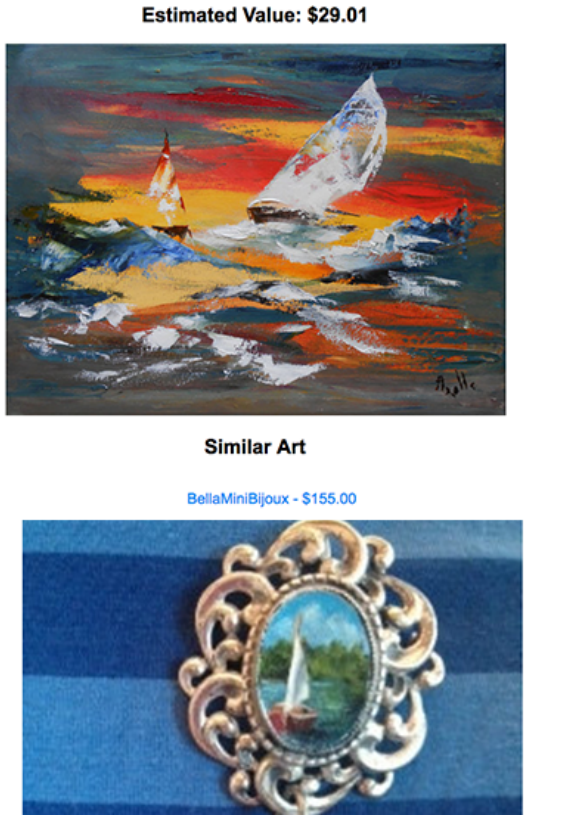

Figure 12: *Example art uploaded on etsyprice.com. The user can see both the price that our model predicts along with several of the closest pieces of art that our model predicts.*

6 Future Work and Conclusions

Future work would expand the dataset to more art types on Etsy and retrain the model. We would scrape multiple websites to gain an improved model. We also would allow users to specify text labels on their art to better estimate its value to the Etsy community. We could periodically scrape Etsy and track trends in art value over time. Finally, if there were a direct integration with Etsy that provided us with transactional data we could train on actual deals that occurred rather than just postings from artists.

In this study, for price estimation of Etsy items, transferring state-of-the-art deep Convolutional Neural Network models is explored. A generic image representation model and a text embedding model are chosen to investigate the transferability of these models and the merging possibility for a price estimation based on an image and a short description. It is observed that a text only regression is more accurate than an image only regression. Nevertheless, by combining the two different kinds of feature we successfully improved the overall accuracy of the model. Another important conclusion is that the model is extremely dependent on the kind of art we are evaluating (e.g. paintings, jewelry, clothes).

Acknowledgements

The authors are very grateful to Professors Alyosha Efros, Stella Yu and Jitendra Malik, along with Tinghui Zhou for valuable suggestions and feedback on the approach, implementation, and analysis of this project.

References

BREIMAN, L. 1996. Bagging predictors. *Machine Learning 24*, 2, 123–140.

HE, K., ZHANG, X., REN, S., AND SUN, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

KRIZHEVSKY, A., SUTSKEVER, I., AND HINTON, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 1097–1105.

REN, S., HE, K., GIRSHICK, R. B., AND SUN, J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR abs/1506.01497*.

RUSSAKOVSKY, O., DENG, J., SU, H., KRAUSE, J., SATHEESH, S., MA, S., HUANG, Z., KARPATHY, A., KHOSLA, A., BERNSTEIN, M., ET AL. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision 115*, 3, 211–252.