Explainable neural image recommendation using Network Dissection visual concepts

Antonio Ossa-Guerra[1,2] Denis Parra[1,2] Hans Löbel[1,2]
[1] Pontificia Universidad Católica de Chile, Santiago de Chile, Chile
[2] Instituto Milenio Fundamentos de los Datos, Santiago de Chile, Chile
aaossa@uc.cl, {dparra,halobel}@ing.puc.cl

Introduction

Problem. Recommendation models that work with visual data, mostly rely on latent image features, which are not understandable on their own. In order to deal with this problem, explanation mechanisms have been developed to provide a visual explanation, but it is hard to translate those explanations into human understandable terms [1, 2, 3, 4, 5].

Motivation. Explanations in a recommendation context increase the users' trust in the system, and therefore the users' satisfaction [5]. By using human understandable concepts, explanations could find new applications and reach new users.

Related work. Current research has applied attention models over images, but those models generate visual but no explicit explanations. Also, different techniques have been developed to identify human understandable concepts in models, but not in single instances [6, 7, 8, 5, 9].

Contributions. We develop a local explanator as an extension of Network Dissection [6] to identify visual concepts in images, and propose and implement a method to transform state-of-the-art visually aware recommendation systems into explainable models that reach performance levels comparable to state-of-the-art models.

Research questions

(**RQ1**) Is it possible to build an **interpretable item representation**?

(**RQ2**) Can we deliver **accurate recommendation** using concept-based representations?

(**RQ3**) Can we provide **explanations** of recommendations in terms of **visual concepts**?

Dataset

UGallery[1] is an online art gallery implemented as an e-commerce platform, where artists can showcase their and sell their art pieces to the platform users. The dataset consists of 2919 users, 13297 items, and 4897 individual purchases or transaions on different art pieces.

In its majority, artworks correspond to physical pieces, meaning that they can only be sold once by the platform. This causes the number of interactions to be significantly lower when compared to other datasets.

Figure 1. Sample UGallery image

[1]: https://www.ugallery.com

Extraction of Visual Concepts

Concept extraction To extract human understandable concepts from a model, a modified NetDissect [6] implementation is used to measure the IoU between the ground-truth segmentation and each unit activated image area. The top IoU value per category for each unit is stored to create a profile for each unit in each analyzed layer.

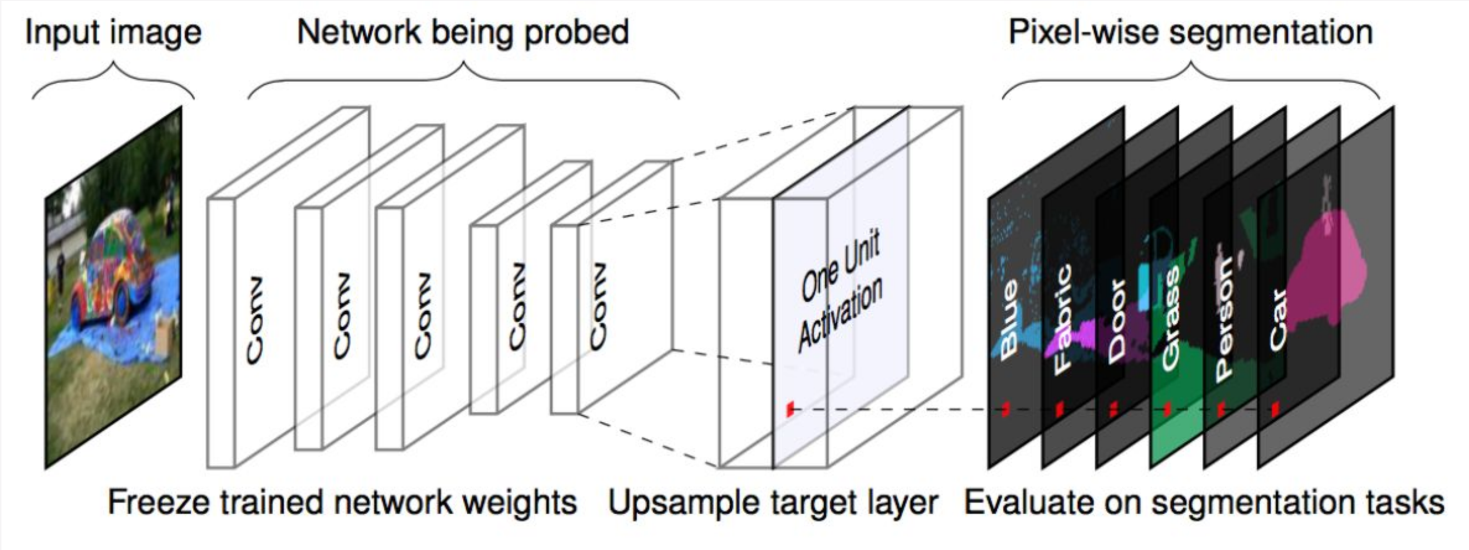

Figure 2. Forward pass of segmentation dataset [6]

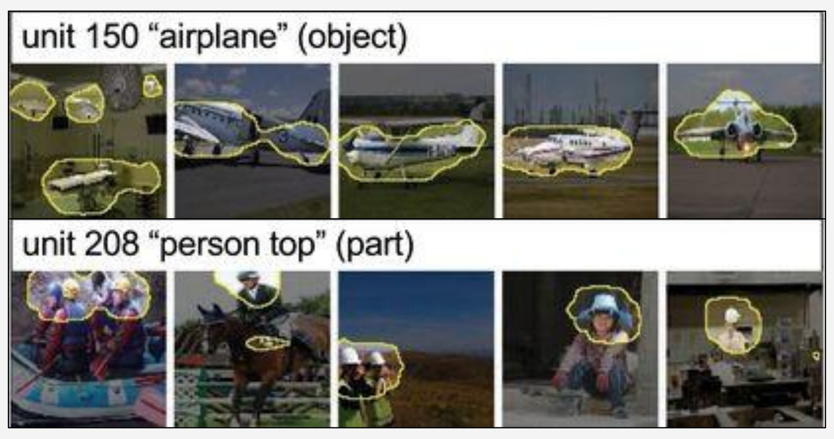

Figure 3. Sample unique detectors and their activations [6]

Model Analysis (Preparation) Before using a model on the selected dataset, the proposed method requires two pieces of information from the NetDissection analysis: (1) the threshold of activation for each convolutional unit, and (2) the units considered as unique detectors (Figure 3).

Details of Proposed Solution

A Representation of Visual Concepts To aggregate the information obtained, using Network Dissection, from a pretrained model from unit-level to model-level, we define 4 criteria: unit score computation (3 options), consider unique detectors only (2), layer weight computation (5), and aggregation of same-concept units (2).

Algorithm 1: Build representation $V(y)$ of visual concepts for image y

Input : An image y, and a set L of convolutional layers in a DNN
Output: A vectorial representation $V(y)$ of visual concepts in image y

```
Initialize V as an empty array;
foreach visual concept c do
    Initialize V_c as an empty array;
    foreach convolutional layer l do
        w_l ← Criteria3(l);
        foreach unit u_k ∈ Criteria2(l) do
            s_{k,c} ← Criteria1(x,k,c);
            V_c[k] ← s_{k,c} · w_l;
    V_c ← Criteria4(V_c);
    V[c] ← V_c;
return V;
```

Algorithmic Definition To build a representation *V(y)* of an image *y* from visual concepts, each criteria can be considered as a function encapsulating a decision. In this way, each combination of criteria results in a slightly different algorithm to transform both the activations of image *y* and the global information provided by NetDissect into a vector *V(y)* of visual concepts.

Embedding Construction We construct embeddings by stacking the representation of every item but using our method as a concept extractor. The proposed embedding can be treated similarly to the obtained using a DNN as a feature extractor, which allow us to train a visually aware recommendation system without changing its architecture. Because of our guided contruction, our data in inherently interpretable.

Due to the number of possible configurations of our method, the baseline model was be used to select the best representations.

Models Trained We use some of the state-of-the-art image recommendation models in our recommendation tasks: VBPR [10] and CuratorNet [11], and VisRank [10, 11] as baseline model. After the training phase, we perform an offline evaluation to compare the performance of the proposed method against its non-explainable counterpart.

Model	Configuration	AUC	MRR	R@20	P@20	N@20	R@100	P@100	N@100
VBPR	Traditional	*.71603*	*.05241*	*.12211*	*.00728*	*.06874*	*.17988*	*.00214*	*.07924*
VBPR	Proposed	**.71964**	**.06505**	**.13600**	**.00817**	**.08180**	**.19297**	**.00232**	**.09320**
CuratorNet	Traditional	*.72226*	*.03681*	*.09499*	*.00563*	*.04998*	*.16004*	*.00192*	*.06325*
CuratorNet	Proposed	.71138	**.03881**	**.10212**	**.00604**	**.05228**	**.17421**	**.00210**	**.06632**
Random	Random	.49868	.00066	.00137	.00007	.00032	.00904	.00011	.00200

Table 1. AUC, Mean Reciprocal Rank (MRR), Recall (R), Precision (P), and nDCG (N) at different recommendation list length (20, 100).

Performance on DL recommendation models Table 1 shows the results of both VBPR and CuratorNet models using the traditional approach (feature extraction using a pretrained ResNet50) and the proposed method (concept embedding using the best overall configuration). In VBPR, the proposed method outperformed its latent counterpart in all metrics, except Recall and Precision at 200. In CuratorNet, the traditional approach only outperformed our method in AUC.

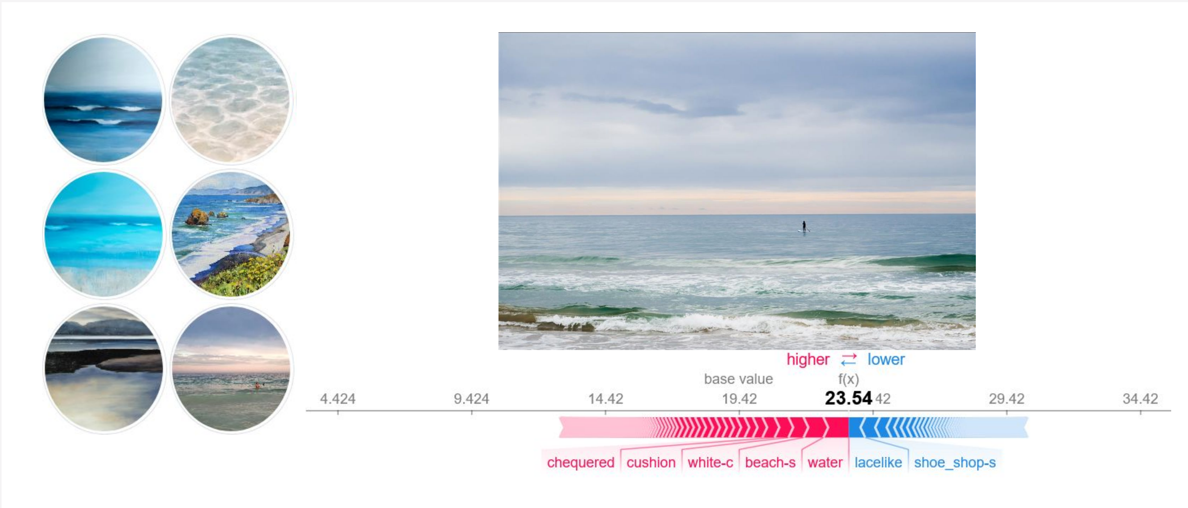

Figure 4. SHAP [12] explanation for a sample user on a "beach" scene

Explanations We apply SHAP [12] to explain how changes in the input (item) modify said score. The SHAP plot that explains the recommendation, attributes each input concept an importance value by analyzing the model internals and modeling how the presence (or absence) of a feature changes the output of the model. This explanation is personalizad, because it considers how the item interacts with the consumed items.

Conclusions

(**C1**) An interpretable item representation based on visual concepts is achievable by our extension of NetDissect.

(**C2**) Our proposed method shows competitive results in state-of-the-art models using an interpretable item representation instead of a latent representation (traditional approach).

(**C3**) Our method can deliver explanations through known feature attribution methods (such as SHAP).

References

1. Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. 2019. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In Proceedings of the 24th International Conference on Intelligent User Interfaces. ACM, 408-416.
2. Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
3. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 93.
4. Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interatction* 27, 3-5 (2017), 393-444.
5. Nava Tintarev and Jusith Masthoff. 2015. Explaining recommendations: Design and evaluation. In Recommender systems handbook. Springer, 353-382.
6. David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6541-6549.
7. Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279* (2017).
8. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 1135-1144.
9. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 20148-2017.
10. He, R., & McAuley, J. (2016, February). VBPR: visual bayesian personalized ranking from implicit feedback. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).
11. Messina, P., Cartagena, M., Cerda-Mardini, P., del Rio, F., & Parra, D. (2020). CuratorNet: Visually-aware Recommendation of Art Images. arXiv preprint arXiv:2009.04426.
12. Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In Proceedings of the 31st international conference on neural information processing systems (pp. 4768-4777).