PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

DEEP NEURAL NETWORK MODELS WITH EXPLAINABLE COMPONENTS FOR URBAN SPACE PERCEPTION.

ANDRÉS CÁDIZ VIDAL

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Advisor:
HANS LÖBEL

Santiago de Chile, July 2020

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE
ESCUELA DE INGENIERÍA

DEEP NEURAL NETWORK MODELS WITH EXPLAINABLE COMPONENTS FOR URBAN SPACE PERCEPTION.

ANDRÉS CÁDIZ VIDAL

Members of the Committee:
HANS LÖBEL
PATRICIO DE LA CUADRA
MEMBER B
MEMBER C

Thesis submitted to the Office of Research and Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science in Engineering

Santiago de Chile, July 2020

Gratefully to my parents and siblings

ACKNOWLEDGEMENTS

Write in a sober style your acknowledgements to those persons that contributed to the development and preparation of your thesis.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

LIST OF FIGURES vii

LIST OF TABLES viii

ABSTRACT ix

RESUMEN x

1. INTRODUCTION 1

2. RELATED WORK 4
   - 2.1. Solutions for estimating urban perception. 4
     - 2.1.1. Classic approaches. 4
     - 2.1.2. Pure machine learning approaches. 5
     - 2.1.3. Mixed approaches. 7
   - 2.2. Explainability in machine learning. 9

3. PROPOSED ARCHITECTURE 13
   - 3.1. Problem Definition 13
   - 3.2. Network architecture 13
   - 3.3. Loss function 13

4. METHODOLOGY 14
   - 4.1. Implementation 14
   - 4.2. Training 14

5. RESULTS 15

6. CONCLUSIONS 16

REFERENCES 17

APPENDIX 22

A. First Appendix 23

LIST OF FIGURES

2.1 Place pulse 2.0 survey 6

2.2 Beta Coefficients 9

2.3 Attention on VQA 12

LIST OF TABLES

ABSTRACT

The abstract must contain between 100 and 300 words. The abstract must be written in English and Spanish. In the case of doctoral theses, the layout of the abstract page is different, so please check the template provided by the OGRS.

Keywords: thesis template, document writing, **(Write here the keywords relevant and strictly related to the topic of the thesis)**.

RESUMEN

El resumen debe contener entre 100 y 300 palabras. El resumen debe ser escrito en inglés y español. En el caso de tesis de doctorado, el formato de la página del resumen es distinta, por favor verifique la plantilla entregada por la Dirección de Postrgrado.

Palabras Claves: plantilla de tesis, escritura de documentos, (**Colocar aquí las palabras claves relevantes y estríctamente relacionadas al tema de la tesis**).

1. INTRODUCTION

Urban perception is a feeling held by people about a location. These feelings can be and are often related to a particular characteristic, like happiness or beauty, or also inherently negative ones, like insecurity or fear (Ordonez & Berg, 2014). Understanding the cause of these feelings is a complex task, since unique social and psychological aspects of each individual affect how they perceive and the spaces they observe (Nasar, 1990).

Visual urban perception is responsible for a large parte of the experience that people go through while being at or using an urban space, this not only affects how much the spaces themselves are used (Khisty, 1994) but also the use of related means of transport (Antonakos, 1995). Other studies have also found correlations between urban perception, crime statistics (Ordonez & Berg, 2014) and wealth, and therefore used it as a proxy measure of inequality (Ordonez & Berg, 2014; Salesses, Schechtner, & Hidalgo, 2013; Rossetti, Lobel, Rocco, & Hurtubia, 2019).

On the other hand, being able to understand a community's need and perception of a city at scale is something of key importance on developing cities, so that the limited resources of local governments can be applied more efficiently (Santani, Ruiz-Correa, & Gatica-Perez, 2018).

Traditional methods for obtaining this type of data, consist of hand made polls about specific locations making systematic evaluation of perception an extremely costly and hard to escalate task (Nasar, 1990; Clifton & Ewing, 2008). Other approach consist of surveys based on computer generated images of simulated spaces, this is more scalable, but is limited to experimental design and it doesn't apply to a real space (Laing et al., 2009; Iglesias, Greene, & de Dios Ortúzar, 2013).

Currently, thanks to the great volumes of data generated by web platforms (Salesses et al., 2013) and to modern deep learning (DL) and computer vision techniques (LeCun, Bengio, & Hinton, 2015), new solutions for estimating urban perception have become feasible, and some previous studies have achieved significant results, either by applying traditional deep learning (Dubey, Naik, Parikh, Raskar, & Hidalgo, 2016) or by combining it with other approaches (Rossetti et al., 2019; F. Zhang et al., 2018). The solutions consist mainly of training deep convolutional neural network models (DCNN) (LeCun et al., 1989) with datasets of urban images that have some sort of label that is used as an estimator for the perception of that urban space. Most of the research is based on the place pulse dataset (Dubey et al., 2016), which consists of pairs of images along with labels that indicate which of the images is more representative of a particular attribute.

However, current deep learning methodologies, have the disadvantage of being ”black boxes”, in other words, they lack a direct or systematic way to explain or interpret the obtained results. This problem comes from the end to end nature of the neural network models and from the millions of learnable parameters they contain. Many of the problems in which these models are used would greatly benefit of more human understandable explanations of the results, making this a very important area of research for the deep learning field (Adadi & Berrada, 2018).

For the particular case of urban perception, explainability of the results is highly relevant, since the added information is valuable for the design of public policy, for example, it could be use to better discriminate which locations would be better recipients of an intervention, and which elements to modify so it convenes an effective improvement of perception.

Current research in explainability is primarily moving in two directions: one is to design novel neural network architectures and training methods so the models are more interpretable, such as the work by Dong, Su, Zhu, and Zhang (2017), the other direction is to create post-hoc algorithms (Adadi & Berrada, 2018) that analyze the results given by the neural network, these algorithms sometimes use other machine learning models, including neural networks (Ghorbani, Wexler, Zou, & Kim, 2019).

The work by Rossetti et al. (2019), presents an approach to this problem for the urban perception case by using semantic segmentation of the images (Badrinarayanan, Kendall, & Cipolla, 2015) as input for a discrete choice model that estimates the perception. The approach allows for a post-hoc aggregated analysis of the results, since the weights of the model are measure of the importance of each class of the semantic segmentation in the calculation of the perception.

The objective of this work is to design and train a model for the urban perception problem, that can give explainable insights on an instance level. For that it proposes a novel solution, consisting of a neural network architecture, that is end-to-end trainable and by using semantic segmentation (Zhao, Shi, Qi, Wang, & Jia, 2016) and self attention mechanisms (Vaswani et al., 2017) can show explainable insights for each of the input images.

ESTO HAY QUE ARREGLARLO AL FINAL
The remainder of this manuscript is organized as follows, Chapter 3 summarizes relevant previous research. In chapter 4 the problem is formally defined and the proposed model is described. Chapter 5 gives details on model implementation and training. Finally, in chapter 6 presents the research results and 7 the final conclusion.

2. RELATED WORK

This chapter consists of two sections, the first one shows an overview some of the different methods that have been previously used in the literature for understanding or estimating urban perception, these methods are separated into 3 types: the classic approaches (all the methods not relying on massive amounts of data are grouped here), approaches based on machine learning and approaches consisting of machine learning models combined with other techniques. The different methods are explained briefly and a short discussion is presented. Section two summarizes the main aspects of the research on explainability on deep learning, and describes some techniques that have been applied in urban perception or other domains that are relevant for this work.

2.1. Solutions for estimating urban perception.

2.1.1. Classic approaches.

Methods for measuring perception of urban spaces have appeared in the literature of several disciplines for many years, with some of the most influential studies dating back to 1960 (Lynch, 1960). Due to technological limits the literature consisted mainly of several types of qualitative surveys for a long time. This surveys consisted in having subjects, complete different tasks such as drawing maps of a certain place (Lynch, 1960), evaluating fundamental aspects of a neighborhood (Nasar, 1990), or in more recent approaches evaluating the impact of transformations generated with edited images (Jiang, Mak, Larsen, & Zhong, 2017). Most of these surveys were conducted in person or by phone, and then the results were analyzed manually, making it very difficult and costly to scale to multiple locations, or larger amounts of samples. The main benefit of this approach, is that it permits a very refined control of the observation process since both the subjects being interviewed and the spaces in question are chosen by the researcher. Added to that, the experiments conducted in person allow for the observer to use senses different than vision to analyze the subject space, resulting in a richer appreciation.

Other methodology, more common in economics and engineering, consists of using discrete choice models and stated choice surveys to model the effect of different variables in perception or other urban related variables (Rose & Bliemer, 2009; Iglesias et al., 2013; Torres, Greene, & Ortúzar, 2013). The amount and complexity of the variables measured depends on the model design. To have and exact control of the variables that have an effect on the survey, computer generated images of urban spaces can be used (Iglesias et al., 2013; Torres et al., 2013).

The advantage of this method is that through the estimated parameters of the model, the effect of each of the studied variables on the perception estimation can be measured, allowing for quantitative results and an understanding of the impact different elements have on the perception of the urban landscape. The main disadvantage of this approach comes from the difficulty of the survey design, variables need to be chosen carefully and the process its vulnerable to biases from the model designer.

2.1.2. Pure machine learning approaches.

Thanks to the massive adoption of web and mobile technologies such as google maps, new types of data are available in considerably large volumes, and new highly scalable ways of generating data can be designed and implemented quickly. That fact allows for some very data dependent machine learning algorithms to be applied to new problems, including urban perception estimation. Several different datasets have been proposed for this problem, most of them based on surveys over large amounts of urban images (Salesses et al., 2013; Dubey et al., 2016; Quercia, O'Hare, & Cramer, 2014; Liu, Silva, Wu, & Wang, 2017; Santani et al., 2018). The most important of them, all consisting of pairwise comparisons of street view images, are *Place pulse 1.0* (PP 1) (Salesses et al., 2013) with measures of safety, class and uniqueness over images of 4 cities, *Urban Gems* with measures of beauty, quietness and happiness over images of London and *Place pulse 2.0* (PP 2) (Dubey et al., 2016), the largest dataset available, with measures of six different attributes over images of 56 different cities, the models proposed on this work are trained on this dataset. All of these were collected through public online surveys of large scale, where the users are asked to choose the image most representative of an attribute of a pair, see figure 2.1 for an example.

Figure 2.1. Snapshot of the place pulse 2.0 survey. Extracted from Dubey et al. (2016)

Earlier attempts at using this data for training models tried to turn the problem into a classification problem by ranking the images from the votes with manually engineered methods such as the one suggested on the place pulse 1.0 paper (Salesses et al., 2013) and use the rank to split the data in two halves with a different label, Ordonez and Berg (2014) use this approach to train SVM models on PP 1 using different types of visual features as input, including a deep neural network. On the PP 2 paper, the authors present the first end to end deep learning model for urban perception regression, which uses a typical transfer learning technique (Pan & Yang, 2010), a Imagenet (Deng et al., 2009) pretrained network for the base of the model, which is used as input for by two parallel modules, one for classification and one for regression. They train the architecture separately on the 6 different attributes of the dataset, the models learn to emulate human voting and to output a urban perception score (through the regression module) on the image for the correspondent attribute. Other works (Porzi, Rota Bulò, Lepri, & Ricci, 2015; Santani et al., 2018) take similar approaches but pretrain models or use features based on the places dataset (Zhou, Lapedriza, Xiao, Torralba, & Oliva, 2014), which provides better performance according to their results.

F. Zhang et al. (2018), train models on PP 2 by combining a DCNN features and a SVM classifier, they use this model to obtain perception indicators of Beijing, they also use a semantic segmentation model (Cordts et al., 2016) on the images and used the results as input to a linear regression, interpreting the regression weights as an indication of importance of the different segmentation classes on perception. On a following work (F. Zhang et al., 2020) they train one deep network to predict all 6 attributes of PP 2 in one forward pass, they do this using an end-to-end architecture similar to Dubey et al. (2016) but adding one output and loss component for each attribute.

Is important to note that most of the literature so far is more focused on applying the models to new cities (F. Zhang et al., 2018; Santani et al., 2018; Costa, Soares, & Marques, 2019; Rossetti et al., 2019) or generating new datasets with new attributes (Santani et al., 2018; F. Zhang et al., 2020), than it is on improving model design and performance. This is consistent with the fact that so far no good measures of performance for this problem have been defined, due to the fact that the datasets don't provide a measure of perception per se but a proxy through the survey votes. The objective of the models in the literature is to rank the images by the estimated perception of an attribute, but they measure performance using accuracy on classification of the human votes, which doesn't necessarily correlate with the models capacity to generalize and rank well, especially in conflicted cases where even human voters would have difficulties (F. Zhang et al., 2018). Despite the fact that models in the literature don't surpass 70% classification accuracy on PP2, the actual ranking task seems to have correct results either by visual inspection, or by comparing with metrics from other domains such as crime rates or wealth indicators (Rossetti et al., 2019; F. Zhang et al., 2018; Ordonez & Berg, 2014).

2.1.3. Mixed approaches.

With the intention of generating more or different insights, usually more explainability, some work in the literature consists of combinations of computer vision or machine learning methods with other techniques. In Rossetti et al. (2019) the authors use a combination of low and high level features of the images as input for a discrete choice model that calculates perception. They extract low level features with traditional computer vision methods like edges or blobs and the high level features with a pretrained neural network for semantic segmentation. The semantic segmentation features allow for a posthoc analysis of the results, the authors reach conclusions like ”Images with more sidewalks were deemed to be safer, livelier and wealthier, but less beautiful on average” and they present a table with the significance of each of the segmentation classes in each of the six PP 2 attributes according to the discrete model parameters. On a similar line, as was mentioned earlier F. Zhang et al. (2018) in addition to their main method, use semantic segmentation features (they aggregate them by percentage of pixels on the image) as an input for multivariate linear regression allowing for similar conclusions to those of Rossetti et al. (2019) but using the beta coefficients (see figure 2.2).

On another work Seresinhe, Preis, and Moat (2017) train a DCNN to calculate the beauty of outdoor images, using transfer learning from the Places dataset, but separately they use a places trained model to obtain text tags from the scenes such as ’Mountain’ or ’Tower’, and similarly to F. Zhang et al. (2018) they use a regression model (elastic net) to make conclusions about the significance of the concepts on the perception of beauty. The disadvantage of this approaches is that they give more insights of the results only at a general level, and therefore do not allow for conclusions on a per instance level, which is what this work intends to do.

Authors of Costa et al. (2019) do an agreement analysis for this type of datasets, they built their own dataset of pairwise comparisons for safety, but used it for generating clusters of users based on the semantic segmentation of the images they voted for. They conclude that most clusters are due to lack of enough comparisons to do a good characterization and that given enough votes all users converge to one generic profile. Is important to note that authors don’t provide any social or demographic information of the 439 users

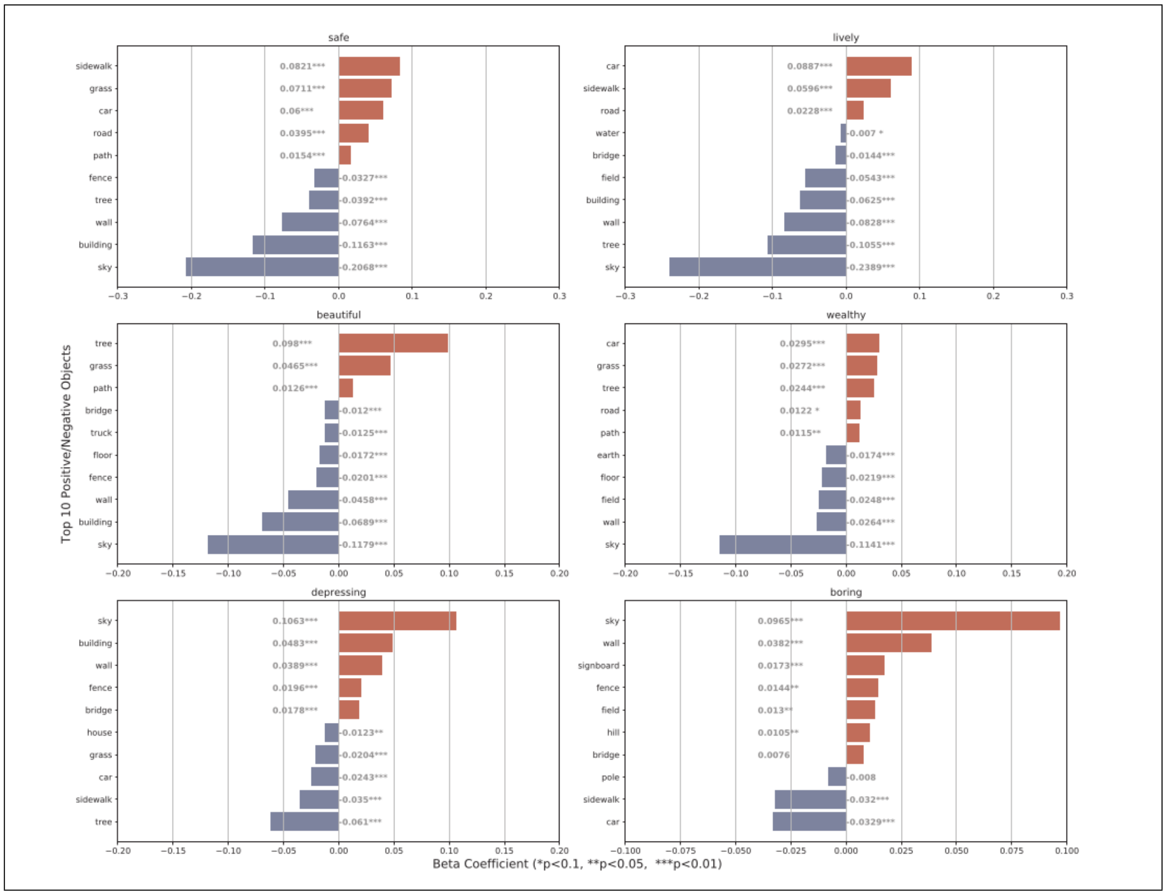

Figure 2.2. Linear regression beta coefficients for most significant objects. Extracted from F. Zhang et al. (2018)

that participated in the survey, and no other similar studies have been done so far so their conclusion hasn't been replicated.

2.2. Explainability in machine learning.

As was mentioned before, explainability has become a very active area of research in machine learning, this is due to the large increase in the usage of ML models for different day to day applications that affect the life's of thousands of people (Ras, van Gerven, & Haselager, 2018). For example, in cases where model outputs are used for analytics or decision making, explainability can make the model both more trustworthy and informative.

Adadi and Berrada (2018) summarize the reasons for enhancing explainability in four points:

(i) Explain to justify: To fullfil the need for reasons of a particular ML generated outcome.
(ii) Explain to control: To allow a better handling of model behavior.
(iii) Explain to improve: The additional understanding of model outputs is useful to design improvements on the systems.
(iv) Explain to discover: As a model overcomes human performance in a task, if its doing so in an explainable manner, then new knowledge for humans may be obtainable.

Is also important to note that laws and regulations related to this topic may become norm in the future such as with the *European Union General Data Protection Regulation (GDPR)* (2016). According to it's articles 13, 14 and 15, when personal data is collected for automated decision-making, the subject has the right to access, and the data controller is obliged to provide, "meaningful information about the logic involved as well as the significance and the envisaged consequences of such processing for the data subject", which will be very difficult to comply with, when working with something like a black box neural network.

One of the two most common approaches to explainability in the literature are Post-hoc methods (Adadi & Berrada, 2018) which try to obtain insights about how the models work, after the process of inference over all the dataset is completed. The methods mentioned in section 2.1.3 are examples of this approach. Other more complex methods found in the recent literature are based on analyzing model sensitivity to semantically meaningful concepts on input (Kim et al., 2017; Shi, Zhang, Wang, & Reddy, 2020), concepts that may be automatically mined from the data as in the approach proposed by Ghorbani et al. (2019). This techniques, although very promising, are still too recent and are not extensible to many domains.

The other approach, which is the one followed by this work, consists of taking advantage of the model design to improve interpretability. This can be done by either using existent features of the model or by introducing architectural changes that make them more explainable. A traditional example of this approach are rule based models like decision trees (Breiman, Friedman, Stone, & Olshen, 1984). Due to the black box nature of deep neural networks this becomes a much more complex task for deep learning, and its an important area of research.

Earlier solutions found in the literature consist of augmenting model input with semantic information, such as text, object bounding boxes or even knowledge bases (Dong et al., 2017; Zhuo, Cheng, Zhang, Wong, & Kankanhalli, 2019; G. Li, Wang, & Zhu, 2019). This methods usually require additional supervision which restricts them to densely annotated datasets such as Visual Genome (Krishna et al., 2016), and the way the neural networks actually use the additional information is not always clear.

Other very common technique in the literature is the use of attention based models (Bahdanau, Cho, & Bengio, 2014), which have layers that consist on using a part of the input (usually called query) to compute a set of weights for the rest of the input (usually called value). The attention weights are usually computed through a linear transformation and a softmax operation on the query, giving them the property of being a probability distribution over the value vector (Cordonnier, Loukas, & Jaggi, 2019), which is used to increase or decrease, parts of the value vector and therefore the layer output. A particular case of attention is self-attention, which means that the same vector is used as both query and input. A common attention architecture in the recent literature is the transformer (Vaswani et al., 2017), which has been widely adopted in both language and vision tasks (Devlin, Chang, Lee, & Toutanova, 2018; Radford et al., 2019; Bello, Zoph, Vaswani, Shlens, & Le, 2019; L. H. Li, Yatskar, Yin, Hsieh, & Chang, 2019; Carion et al., 2020).

Attention models have the additional value that the weights can be visualized and used to interpret what the network is doing, providing explainable information about the model's decision process for each data instance. Clark, Khandelwal, Levy, and Manning

(2019) analyse the attention outputs of the NLP transformer model BERT and show how they correspond well to linguistic notions of syntax, Y. Zhang, Niebles, and Soto (2019) create an explainable VQA model by adding supervised self-attention layers and visualizing their output as heatmaps over the input images (see figure 2.3). Cordonnier et al. (2019) present a theoretical relationship between between self attention and convolutional layers, and as part of their work they provide an interactive visualization of the attention weights [1].

Figure 2.3. Visualization of attention weights for the VQA problem. Extracted from Y. Zhang et al. (2019)

[1]Available at `epfml.github.io/attention-cnn/`

3. PROPOSED ARCHITECTURE

3.1. Problem Definition

3.2. Network architecture

3.3. Loss function

4. METHODOLOGY

4.1. Implementation

4.2. Training

5. RESULTS

6. CONCLUSIONS

Nothing to say. Be happy.

REFERENCES

Adadi, A., & Berrada, M. (2018, 09). Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2018.2870052

Antonakos, C. L. (1995). Environmental and travel preferences of cyclists. *Transportation Research Part A: Policy and Practice*, *29*(1), 85.

Badrinarayanan, V., Kendall, A., & Cipolla, R. (2015). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *CoRR*, *abs/1511.00561*. Retrieved from `http://arxiv.org/abs/1511.00561`

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., & Le, Q. V. (2019). Attention augmented convolutional networks. In *Proceedings of the ieee international conference on computer vision* (pp. 3286–3295).

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A. M., & Zagoruyko, S. (2020). End-to-end object detection with transformers. *ArXiv*, *abs/2005.12872*.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? an analysis of bert's attention. *CoRR*, *abs/1906.04341*. Retrieved from `http://arxiv.org/abs/1906.04341`

Clifton, K., & Ewing, R. (2008, 03). Quantitative analysis of urban form: A multidisciplinary review. *Journal of Urbanism: International Research on Placemaking and Urban Sustainability*, *1*, 17-45. doi: 10.1080/17549170801903496

Cordonnier, J.-B., Loukas, A., & Jaggi, M. (2019). On the relationship between self-attention and convolutional layers. *arXiv preprint arXiv:1911.03584*.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., … Schiele,

B. (2016). The cityscapes dataset for semantic urban scene understanding. *CoRR*, *abs/1604.01685*. Retrieved from http://arxiv.org/abs/1604.01685

Costa, G., Soares, C., & Marques, M. (2019). Finding common image semantics for urban perceived safety based on pairwise comparisons. In *2019 27th european signal processing conference (eusipco)* (pp. 1–5).

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, & Li Fei-Fei. (2009). Imagenet: A large-scale hierarchical image database. , 248-255.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, *abs/1810.04805*. Retrieved from http://arxiv.org/abs/1810.04805

Dong, Y., Su, H., Zhu, J., & Zhang, B. (2017). Improving interpretability of deep neural networks with semantic information. *CoRR*, *abs/1703.04096*. Retrieved from http://arxiv.org/abs/1703.04096

Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). Deep learning the city: Quantifying urban perception at a global scale. *ArXiv*, *abs/1608.01769*.

European union general data protection regulation (gdpr). (2016). Retrieved from https://gdpr-info.eu/

Ghorbani, A., Wexler, J., Zou, J. Y., & Kim, B. (2019). Towards automatic concept-based explanations. In *Advances in neural information processing systems* (pp. 9273–9282).

Iglesias, P., Greene, M., & de Dios Ortúzar, J. (2013). On the perception of safety in low income neighbourhoods: using digital images in a stated choice experiment.

Jiang, B., Mak, C. N. S., Larsen, L., & Zhong, H. (2017). Minimizing the gender difference in perceived safety: Comparing the effects of urban back alley interventions. *Journal of Environmental Psychology*, *51*, 117–131.

Khisty, C. J. (1994). Evaluation of pedestrian facilities: beyond the level-of-service concept. *Transportation Research Record*, *1438*, 45-50.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., & Sayres, R. (2017). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). *arXiv preprint arXiv:1711.11279*.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... Fei-Fei, L. (2016). Visual genome: Connecting language and vision using crowdsourced dense image annotations.. Retrieved from `https://arxiv.org/abs/1602.07332`

Laing, R., Davies, A.-M., Miller, D., Conniff, A., Scott, S., & Morrice, J. (2009). The application of visual environmental economics in the study of public preference and urban greenspace. *Environment and Planning B: Planning and Design*, *36*(2), 355-375. doi: 10.1068/b33140

LeCun, Y., Bengio, Y., & Hinton, G. (2015, 05). Deep learning. *Nature*, *521*, 436-44. doi: 10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Li, G., Wang, X., & Zhu, W. (2019). Perceptual visual reasoning with knowledge propagation. In *Proceedings of the 27th acm international conference on multimedia* (p. 530–538). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3343031.3350922` doi: 10.1145/3343031.3350922

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). Visualbert: A simple and performant baseline for vision and language. *ArXiv*, *abs/1908.03557*.

Liu, L., Silva, E. A., Wu, C., & Wang, H. (2017). A machine learning-based method for the large-scale evaluation of the qualities of the urban environment. *Computers, Environment and Urban Systems*, *65*, 113–125.

Lynch, K. (1960). *The image of the city (vol. 11).* MIT press Cambridge, MA, USA.

Nasar, J. L. (1990). The evaluative image of the city. *Journal of the American Planning Association*, *56*(1), 41-53. doi: 10.1080/01944369008975742

Ordonez, V., & Berg, T. L. (2014). Learning high-level judgments of urban perception. In D. Fleet, T. Pajdla, B. Schiele, & T. Tuytelaars (Eds.), *Computer vision – eccv 2014* (pp. 494–510). Cham: Springer International Publishing.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345-1359.

Porzi, L., Rota Bulò, S., Lepri, B., & Ricci, E. (2015). Predicting and understanding urban perception with convolutional neural networks. In *Proceedings of the 23rd acm international conference on multimedia* (pp. 139–148).

Quercia, D., O'Hare, N. K., & Cramer, H. (2014). Aesthetic capital: what makes london look beautiful, quiet, and happy? , 945–955.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Ras, G., van Gerven, M., & Haselager, P. (2018). Explanation methods in deep learning: Users, values, concerns and challenges. In *Explainable and interpretable models in computer vision and machine learning* (pp. 19–36). Springer.

Rose, J. M., & Bliemer, M. C. J. (2009). Constructing efficient stated choice experimental designs. *Transport Reviews*, *29*(5), 587-617. Retrieved from `https://doi.org/10.1080/01441640902827623` doi: 10.1080/01441640902827623

Rossetti, T., Lobel, H., Rocco, V., & Hurtubia, R. (2019). Explaining subjective perceptions of public spaces as a function of the built environment: A massive data approach. *Landscape and Urban Planning*, *181*, 169-178.

Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013, 07). The collaborative image of the city: Mapping the inequality of urban perception. *PLOS ONE*, *8*(7), 1-12. Retrieved from `https://doi.org/10.1371/journal.pone.0068400` doi: 10.1371/journal.pone.0068400

Santani, D., Ruiz-Correa, S., & Gatica-Perez, D. (2018). Looking south: Learning urban perception in developing cities. *ACM Transactions on Social Computing*.

Seresinhe, C. I., Preis, T., & Moat, H. S. (2017). Using deep learning to quantify the beauty of outdoor places. *Royal Society open science*, *4*(7), 170170.

Shi, T., Zhang, X., Wang, P., & Reddy, C. K. (2020). A concept-based abstraction-aggregation deep neural network for interpretable document classification. *arXiv*, arXiv–2004.

Torres, I., Greene, M., & Ortúzar, J. d. D. (2013, 07). Valuation of housing and neighbourhood attributes for city centre location: A case study in santiago. *Habitat International*, *39*, 62–74. doi: 10.1016/j.habitatint.2012.10.007

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *CoRR*, *abs/1706.03762*. Retrieved from `http://arxiv.org/abs/1706.03762`

Zhang, F., Zhou, B., Liu, L., Liu, Y., Fung, H. H., Lin, H., & Ratti, C. (2018). Measuring human perceptions of a large-scale urban region using machine learning. *Landscape and Urban Planning*, *180*, 148–160.

Zhang, F., Zu, J., Hu, M., Zhu, D., Kang, Y., Gao, S., ... Huang, Z. (2020). Uncovering inconspicuous places using social media check-ins and street view images. *Computers, Environment and Urban Systems*, *81*, 101478.

Zhang, Y., Niebles, J. C., & Soto, A. (2019). Interpretable visual question answering by visual grounding from attention supervision mining. In *2019 ieee winter conference on applications of computer vision (wacv)* (pp. 349–357).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2016). Pyramid scene parsing network. *CoRR*, *abs/1612.01105*. Retrieved from `http://arxiv.org/abs/1612.01105`

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In *Advances in neural information processing systems* (pp. 487–495).

Zhuo, T., Cheng, Z., Zhang, P., Wong, Y., & Kankanhalli, M. (2019). Explainable video action reasoning via prior knowledge and state transitions. In *Proceedings of the 27th acm international conference on multimedia* (p. 521–529). New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/3343031.3351040` doi: 10.1145/3343031.3351040

APPENDIX

A. FIRST APPENDIX