Artificial Imagination of Architecture with Deep Convolutional Neural Network

Joaquim Silvestre, Yasushi Ikeda, François Guéna

▶ **To cite this version:**

Joaquim Silvestre, Yasushi Ikeda, François Guéna. Artificial Imagination of Architecture with Deep Convolutional Neural Network. CAADRIA 2016: Living Systems and Micro-Utopias - Towards Continuous Designing, Mar 2016, Melbourne, Australia. pp.881-890, 10.52842/conf.caadria.2016.881 . hal-03690510

HAL Id: hal-03690510
https://hal.science/hal-03690510

Submitted on 8 Jun 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

S. Chien, S. Choo, M. A. Schnabel, W. Nakapan, M. J. Kim, S. Roudavski (eds.), *Living Systems and Micro-Utopias: Towards Continuous Designing, Proceedings of the 21st International Conference of the Association for Computer-Aided Architectural Design Research in Asia CAADRIA 2016,* 881–890. © 2016, The Association for Computer-Aided Architectural Design Research in Asia (CAADRIA), Hong Kong.

ARTIFICIAL IMAGINATION OF ARCHITECTURE WITH DEEP CONVOLUTIONAL NEURAL NETWORK

"Laissez-faire": loss of control in the esquisse phase.

JOAQUIM SILVESTRE, YASUSHI IKEDA and FRANÇOIS GUÉNA
Keio University, Fujisawa, Japan
j.silvestre82@gmail.com, yasushi@sfc.keio.ac.jp, francois.guena@maac.archi.fr

Abstract. This paper attempts to determine if an Artificial Intelligence system using deep convolutional neural network (ConvNet) will be able to "imagine" architecture. Imagining architecture by means of algorithms can be affiliated to the research field of generative architecture. ConvNet makes it possible to avoid that difficulty by automatically extracting and classifying these rules as features from large example data. Moreover, image-base rendering algorithms can manipulate those abstract rules encoded in the ConvNet. From these rules and without constructing a prior 3D model, these algorithms can generate perspective of an architectural image. To conclude, establishing shape grammar with this automated system opens prospects for generative architecture with image-base rendering algorithms.

Keywords. Machine learning; convolutional neural network; generative design; image-based rendering.

1. Introduction

What does it take to make an architectural representation believable? By just putting a 1/50 figurine on a model or a mere shape, we envision it on a human scale and start to imagine it as life-size architecture. Hollow, residual volumes are imagined as spaces where we might go once a project has been built. By doing so, our imagination fills in the gaps and ignores the lack of details that are usually present in realistic representations.

This way we use our imagination is based on our experiences. We've seen thousands of buildings of various shapes and we can envision that what

we have in front of us, a mere shape is more than what we see. All the types of windows, doors, roofs we have seen are registered in our memory. They are elements that we articulate with non-formalised grammar while we perceive and create. Therefore, chief characteristic of an artificial imagination system that mirrors human ability is a shape grammar (Stiny, 1980), which formalise various types of architectural styles, materials and structural systems. Developing this system presents worthy outcomes for the design process studies and it is obvious how appealing it would be to automatise the formalisation of such system of rules.

Recently, progress in ConvNet shows that it is possible to use existing architecture building pictures to extract rules and then recombine these rules into a new architecture building image. Indeed, rules of classification are implicitly embedded in the images' compositions, and a system such as ConvNet extracts them. From a series of training pictures, the "back propagation" algorithm progressively sets weights and biases on branches and nodes of the network in a bottom-up way (Nielsen, 2015). The strength of this process is autodetermination, letting the system figure out by itself how to select criteria in order to perform the correct classification. This contrasts with the top-down way experts use to formalise such grammar.

Similar to our imagination process, this system needs something to trigger and orient its "imagination". As front-end interfaces, the user can orient and influence the generated image through pictures that serve as hints about what he expects from the generative system. What raises interest for design research is that non-human logic offers designers a new point of view on their object. By envisioning the Algorithmic architecture definition of Kostas Terzidis (2006), this ability could improve the control loop feedback by providing a new way to interact with computers. The paper will focus on the way we could build this system and more specifically about the interface to control it while keeping its ability to surprise us.

The paper presents experiments with ConvNet that explore what kind of ConvNet architecture, training data sets and, rendering algorithms can generate perspective view of the architectural scene. The first part is focused on the explanation of the process that ponders the weight of the neural network. Through the use of Deep Dream (Mordvintsev et al, 2015) algorithms, the importance of the learning material will be emphasised. Then, in the second part will be pointed out the default of current pre-trained deep neural networks when they are used for architecture image generation purpose. This observation will guide us in the need for a better understanding of inner workings of ConvNet in the third part. This part will develop on how features are extracted and how they can be crafted and used to orient the generation of more coherent pictures. As conclusion, regarding what existing

image-based rendering (Fitzgibbon, 2003) algorithms can produce, guidelines to build architecture image generation system and its control interface will be proposed. The generative system presented in this paper is limited to 2D perspective representation of architecture. This is mainly due to the input data format: pictures of architecture buildings. We authors understand that a solely using one picture to represent a building or a room is a very truncated expression of architectural projects.

2. First experiment: generate architecture with deep dream and an architecture image

2.1 REVEAL THE SUBCONSCIOUS OF THE NETWORK

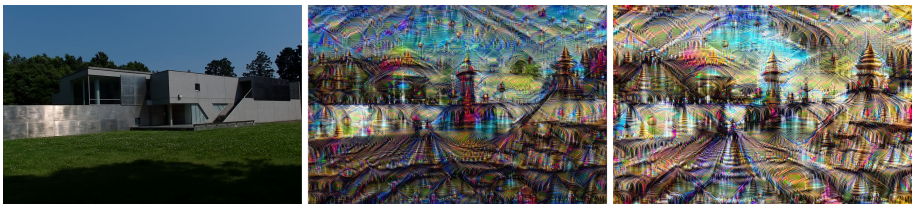

Figure 1. Details appear gradually at each step. They are based on shapes of the input image.

The central component of this system is the neural network. It concentrates and encodes all the information implicitly embedded in pictures. The network is composed of various hidden layers of computational neurons. Each performs an abstract classification task. In order to visualise what is being detected in each layer, the "Deep Dreaming" algorithm has been created by Google researchers. When Deep Dream, exploits a ConvNet it performs "gradient ascent process that tries to maximise the L2 norm"(Mordvintsev et al, 2015). In simpler terms, it modifies the input image in a way so that the detection score with the features encoded in the ConvNet is higher. Then, the output image with features exacerbated is again processed in the ConvNet with the same process. On each pass in the loop, the likeness with the ConvNet encoded features get higher. It's important to understand that features are abstract representation; they are not image components.

As we can see, it provides unclear psychedelic images that are difficult to recognize. Nonetheless, we still feel some architectural tectonic underneath. A noticeable aspect of these images is the part of the meaning is derived from what is projected by the observer. In the stance of the Rorschach test, it may reveal the viewer's mind. These dreamed images are halfway from our imagination and the "imagination" of the ConvNet. The vision of this odd dream awake us from the "world created by smart advertisers …" (Wood, 2007) since the original is close to the unimaginable.

2.2 THE DATASET

Figure 2. Left: image produced with GoogleLeNet; Right: with MIT Places (Zhou 2014).

A downside of the Deep Dream algorithm is its obsession in making us see animals in everything. By default, the program uses the GoogleLeNet trained on ImageNet Dataset. This Dataset contains a large amount of animal pictures and it explains this strange obsession. For architecture design purposes, it's more suited to use models train on MIT Places Dataset. By using it, there is a strong tendency to reveal more "architecture expressive" friendly forms.

We clearly see more biological tendencies with GoogleLeNet. This author can see a chimera of Scorpion and insects on figure 2 left. Instead of that, the Places205 ConvNet changes trees into Pagodas, and generate fractal straight roads with vanishing point wherever possible. Further experimentations of deep dream based on Places ConvNet show that that kind of dream is not so specific to the building input image used in the figure 1.

The usual explanation for this behaviour is the different types of pictures in the dataset are unbalanced and the ground truth categories in which they are classified are oriented for a specific detection task. If large parts of the pictures in the dataset contain eyes and if this visual detail is an asset to recognise in which category the image should be classified, this leads to create a tendency of popping eyes when Deep Dream is used with this ConvNet. If we look closely the M.I.T. Places dataset, we can find lots of images with humans on it. These images are poor in architectural or environmental detail but they still inform us on the plausible place.

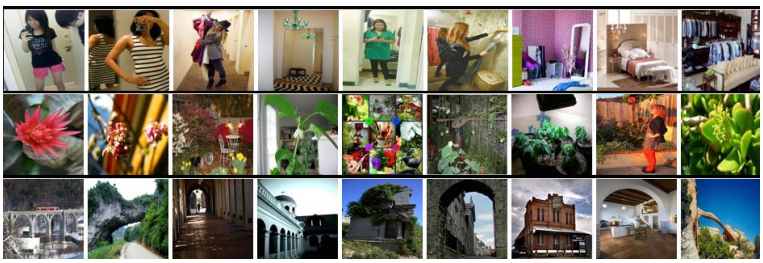

Figure 3. Samples of the Places Data set: labels for each row starting to the top, fitting room, floral shop indoor, arch. The last row could be used in a more architecture oriented Dataset.

Since places are partially defined by human activity, the data set contains lots of pictures with humans inside. But a humanoid shape doesn't pop when Deep Dream because to classify images in the demanded category, it seems that the network discovers more valuable abstract features than the one present in eyes. In short, what links these images for us can be the human behaviour factor but, for the trained ConvNet, it's something else, such as maybe a few values of colour pixels in a specific pattern.

2.3 CRITICS OF DREAMING

As said before, Deep Dream has been created to envision what is encoded inside the ConvNet. The purest expression of what is encoded is done by using an RGB noise as the input image (see Figure 4). It reveals the tendency of the ConvNet. Whatever we use Place ConvNet or GoogleLeNet ConvNet, the tendencies are different but we can notice familial likeness in their dream: Images tend to have a "rainbow" texture that prevents generated architectural pictures the ability to express materiality. Secondly, elements are unrelated and are floating next to each other without structure.

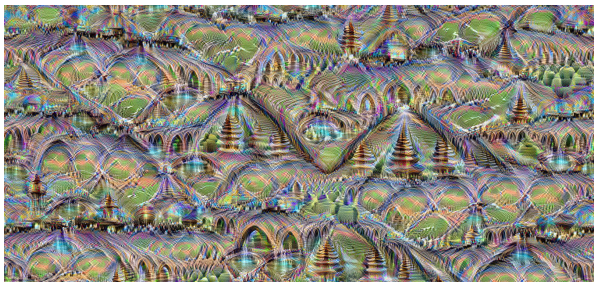

Figure 4. "Natural expression" of the ConvNet trained with places. From an RGB noise image the algorithm is more prone to generate this kind of shape.

Often, pictures of architectures present a background that fills the pictures' space more than an iconic element that stands in the middle. From our experimentation, actual Deep Dream is more suited to help us invent a new chimera than an unseen architecture environment. Last, the elements don't fit into any layout context: the upper part is filled with elements similar to the one on the lower part: Context of the canvas is not taken in consideration.

Unlike with this noise input image, the generation can be structured in a particular context as shown in Figure 1. It doesn't state clearly a design intention, but that's the designer's responsibility to transform it in a reachable "micro-utopia" (Wood, 2007). What does it take to produce more realistic images of an architectural environment? Firstly, there is a lack of scale hier-

archy between elements. In the element itself, relations of proportion are respected, but through the space of the picture, a pagoda can be the size of a mountain. This can be due to a lack of relation to the human scale and rules of perspective that aren't encoded in some ways in the network. This will be the theme of further research.

Moreover, pictures of the data set are framed to show something specific: It doesn't try to detect something in the background. This technical limitation creates an implicit connection between pictures: mainly centred on the one central element it should learn to detect, and then discarding the background. This is not a problem for their initial goal, but to hijack the model to generate architecture image, the lack of features encoding the context is a pitfall.

Besides changing the dataset, improving such system can be done by changing the architecture of the ConvNet. For instance, if it includes SpatialPyramid module like in the work of Clement Farabert (2013), we can expect that dreamed images will loose the tendency to imagine people in the sky or other incongruous elements. Indeed, the Spatial Pyramid module duplicates the trained image in various resolutions and process them in parallel and share the weight of neurone layer at different scales. By doing so, low-resolution images processed will detect more general and composition related features while higher resolution ones will detects components itself.

3. Second experiment: generate architecture image with 2 inputs images and neural style algorithm

Figure 5. The algorithm is far more than a Photoshop filter; it identifies how things are represented in the artist's style.

The design cognitive process could be a model to change the network architecture. If the algorithm generates architecture in a similar fashion as architects design, generated architecture could be more realistic. Moreover, it could test the cognitive process used as a model for the algorithm. Neuroscientists (Gatys et al, 2015) create a neural network to show the process involved in the creation and perception of artistic imagery. The algorithm they created is a model of artistic process that supposes a dichotomy between style and content. Basically, the algorithm takes a content input photo and

one or more style input images like a painting and applies the style on the content − i.e.: the content is drawn with the style provided.

This theory of cognitive processes can be applied to architecture, too. Same as in the painting field, there are historic architectural style and personal architectural style. But the transposition can't be so naively done: architects don't only represent reality with their style; they create it. The sole representation time is during the project rendering phase and the sketching phase. Even if they use an architectural sketch as style, this notion of content can't fit in the architecture cognitive design process theory: architects don't apply their style on content. The content can't be another architect's building since this one already has a style embedded into all the details. The closest notion that can replace content could be the program. The architect applies his or her style on the way to execute the program.

Figure 6. If the pictures are wisely chosen, it's possible to use this algorithm with architect style on content.

Like in figure 6, if we try to apply the neural style algorithm on architecture to generate an architectural hybridisation, it doesn't apply a style. Instead, patterns and materials are applied. To illustrate this, we mixed naively, the style of the architect Kengo Kuma with the "content" of Louis Khan library. Nevertheless, understand that this algorithm could be useful to improve generative architectural images. As we see, output images are cleaner and more realistic even if their novelty factor is less wild.

After experimenting with the code on some content and art pairings, it reveals that the inner mechanics can be popularised as: an understanding of the style as a summary of the textures used by the artist. Hitherto, it's appropriate for impressionist art but less efficient on textureless art. It matches content with ways to represent it in the original art.

Again, for this to work, it is based on a trained neural network. This one provides abstract categories in which style and content features match. Changing the ConvNet will influence the process by creating different features matching categories.

As has been explained before, upon building a ConvNet specifically for architecture, it will be interesting to see if the matching is more precise. Another way to use Neural Style for generative purpose is to hack the inner log-

ic of the algorithm in order to apply it to architecture. The example in figure 7 shows how the style factor can be replaced by a material set. The algorithm will apply them geometrically in almost a rational fashion. Right now, we are here in-between fulfilled expectation and surprise.

Figure 7. The Bunker shape as content, an abstract style image composed with desired material give a wooden bunker.

Neural style was intended to support the theory of "understanding of style" through the algorithmic process. In the case of architectural design, the alteration of the image in figure 7 doesn't show an understanding of the construction materials used in design. To explore that point and more about the cognitive process of architect, a specific algorithm to model the architect creation and perception of pictures of architectures should be established. Supposing that architectural design is the interaction of structural solutions, materiality and tectonics. The model that is able to test this theoretical division and how they can interact in the creative phase should stimulate the extraction of the relevant information for recomposition from input images.

Mere generation for serendipitous purpose is a less ambitious goal. If we exclude that the algorithm should fit the model of a plausible cognitive process. We could focus on merely being more genuine in following the rules of perception, rather than architectural design ones, in order to get more visually satisfying images. For further research, dropping the cognitive process likeness seems more realistic.

To conclude this part, the extracting and matching ability of this algorithm open the perspective of using multiple image input of various types in order to drive the image generation. We can imagine image equations that can be processed through an algorithm similar to Neural Style.

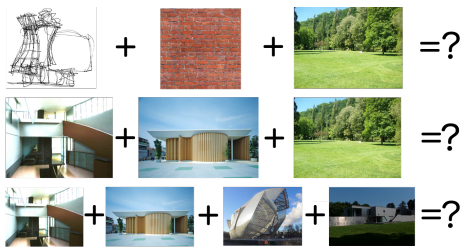

Figure 8. Various types of input equation that can be used to drive the image generation

We could imagine that the first equation is in a similar fashion to Neural Style algorithm, matching material and sketch and inlay the image in the site. The second equation will extract relations between site and architecture, elements position (like windows) and tectonics and then generate pictures of architecture within a site picture. Or, just hope that the network find a "weird logic" of abstract features associations and succeed in designing a building and the site background by taking inspiration of the 4 inputs presented in the third equation.

4. Conclusion

Figure 9. In this dataset sample we can notice the iconic and stand alone aspect of the chosen architecture building

As we see in this paper, deep neural network succeeded in grasping a few notions of architecture and managed to manipulate them in unexpected ways. Despite the fact that these notions are encoded within the black box of the ConvNet in a format we can not operate with, we still can control the building input material and parameter of the ConvNet. In future works, we aim at generating a more coherent architectural space and it seems that more notions need to be taught to the network. In order to achieve that, we will keep using the open-sourced code cited here while refining the ConvNet. In this case, ConvNet can be tuned through two parameters: curation of pictures of the training dataset and the design of ConvNet architecture.

Currently, ConvNet architectures have been designed for a detection and identification purpose. In future works, we would like to set up a network architecture for the sole purpose of generating. Then, instead of feeding it with the usual dataset focused on daily life objects or domestic animals, we would like to create an architecture specific dataset.

Since architecture has a wide range of expressions and points of view, we will segment the task into various datasets. Indeed, inside space cannot be fully grasped through one picture. It seems wiser to start simple and make the first test dataset with the front view of various individual buildings standing in a open surrounding. With this test dataset, we aim to produce images of iconic buildings in an understated context, like in the Figure 9. If we want

to work with more architecturally inner space, pictures such as the ones in Figure 10 would be used.

Figure 10. Unlike the previous figure, the background is more significant than the foreground

Beside the appropriate data set, other obstacles need to be overcome: conceptual notions of a pictures' construction need to be embedded in the network. We can imagine that notions of hierarchy of scale and perspectives need to be encoded in someway into the system. One solution could be to provide RGB+D pictures of architecture space or a set of two close views of each scene. Once such a base is set, we can start to experiment on a combination of ConvNet architecture, type of image feeding datasets and image-based rendering methods. When the combination of these three elements are settled, it will be very intuitive for designers to express an orientation of design through operations between only a few reference images.

References

Nielsen, N. M.: 2015, *Neural Networks and Deep Learning*, Detremination Press.

Mordvintsev, A., Olah, C., and Tyka, M.: 2015, "Inceptionism". Available from: <http://googleresearch.blogspot.com/2015/06/inceptionism-going-deeper-into-neural.html>.

Gatys, A. L., Ecker, S. A. and Bethge, M.: 2015, A Neural Algorithm of Artistic Style, *CoRR*.

Farabet,C., Courpie, C., Najman, L. & LeCun, Y.: 2013, Learning Hierarchical Features for Scene Labeling, *IEEE Trans*.

Fitzgibbon, A., Wexler, Y. and Zisserman, A.: 2003, Image-based rendering using image-based priors, *ICCV*.

Fuller, M. and Sónia, M.: 2011, Feral Computing: From Ubiquitous Calculation to Wild Interactions, *Fiberculture Journal*, **135**, 144–163.

Stiny, G.: 1980, Introduction to shape and shape grammars. *Environment and Planning B,* **7**(3), 343–351

Terzidis, K.: 2006, *Algorithmic Architecture*, Architectural Press, Boston.

Wood, J.: 2007, *Design for Micro-Utopias: Making the Unthinkable Possible*, Gower.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A. and Oliva, A.: 2014, Learning Deep Features for Scene Recognition using Places Database, *Advances in Neural Information Processing Systems,* 27 (NIPS).