Unlocking Feature Visualization for Deeper Networks with MAgnitude Constrained Optimization

Thomas Fel[*1,2,4], **Thibaut Boissin**[*2,3], **Victor Boutin**[*1,2], **Agustin Picard**[*2,3], **Paul Novello**[*2,3]
Julien Colin[1,5], **Drew Linsley**[1], **Tom Rousseau**[4], **Rémi Cadène**[1], **Laurent Gardes**[4],
Thomas Serre[1,2]
[1]Carney Institute for Brain Science, Brown University
[2]Artificial and Natural Intelligence Toulouse Institute
[3]Institut de Recherche Technologique Saint-Exupery
[4]Innovation & Research Division, SNCF
[5]ELLIS Alicante, Spain.
{thomas_fel@brown.edu, thibaut.boissin@irt-saintexupery.com}

Abstract

With the development of increasingly large neural architectures, there is a pressing need to develop explainability methods that can scale up to the demand. Yet, standard methods for feature visualization fail entirely on neural architectures developed after 2014 and require resorting to strong prior image models to be usable – raising questions about their validity. Here, we describe a relatively simple trick to finally unlock feature visualization for large neural networks: beyond searching for maximally activating images in the Fourier domain as in standard methods [19], we find that optimizing solely an image's phase spectrum while keeping its magnitude constant yields significantly better results – both qualitative and quantitative. Indeed, in addition to producing more compelling visualizations, our method exhibits an attribution mechanism that we leverage to encode spatial importance in the explanation. To our knowledge, our study is the first to unlock feature visualizations for the largest, state-of-the-art classification networks without resorting to any parametric prior image model, effectively advancing a field that has been stagnating since 2017 [19]. In this demo, we will showcase our results in the 1000 classes of ImageNet, which are also publicly available in our website, Loupe.

1. Introduction

The field of Explainable Artificial Intelligence (XAI) has largely focused on characterizing the intricacies of computer vision models through the use of attribution methods [6, 18, 22–24]. These methods aim to explain the decision strategy of a network by assigning an importance score each input pixel (or group of input pixels), according to their contribution to the overall decision. Such approaches only offer a partial understanding of the learned decision process as they aim to identify the location of the most discriminative features in an image, the "where", leaving open the "what" question, *i.e.* the semantic meaning of those features. Recent work [3] has highlighted the intrinsic limitations of attribution methods, calling for the development of methods that provide a complementary explanation regarding the "what".

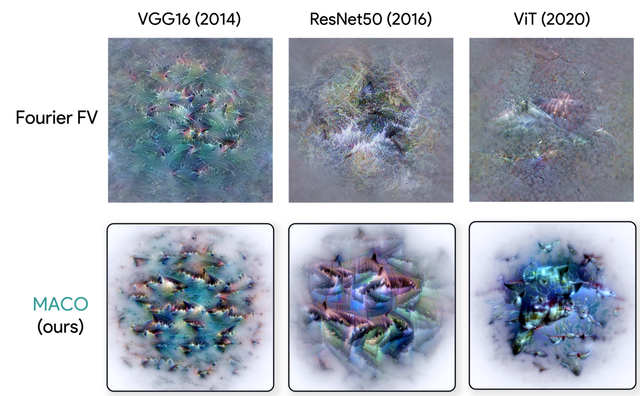

Figure 1. **Comparison between feature visualization Methods for "White Shark" Classification. (Top)** Standard Fourier preconditioning-based method for feature visualization [19]. **(Bottom)** Proposed approach, **MACO**, which incorporates a Fourier spectrum magnitude constraint.

Feature visualizations provide a bridge to fill this gap via the generation of images that elicit a strong response from a specifically targeted neuron (or a group of neurons). One of the simplest approaches uses gradient ascent to search for such an image. In the absence of regularization, this optimization is known to yield highly noisy images – sometimes considered adversarial [25]. Hence, regularization methods are essential to produce more acceptable candidate images. Such regularizations can consist of penalizing high frequencies in the Fourier domain [1, 12, 16, 19, 27], regularizing the optimization process with data augmentation [5,13,19,21,26] or restricting the search space to a subspace parameterized by a generative model [14, 15, 17, 28]. The first two approaches provide faithful visualizations, as

* Equal contribution

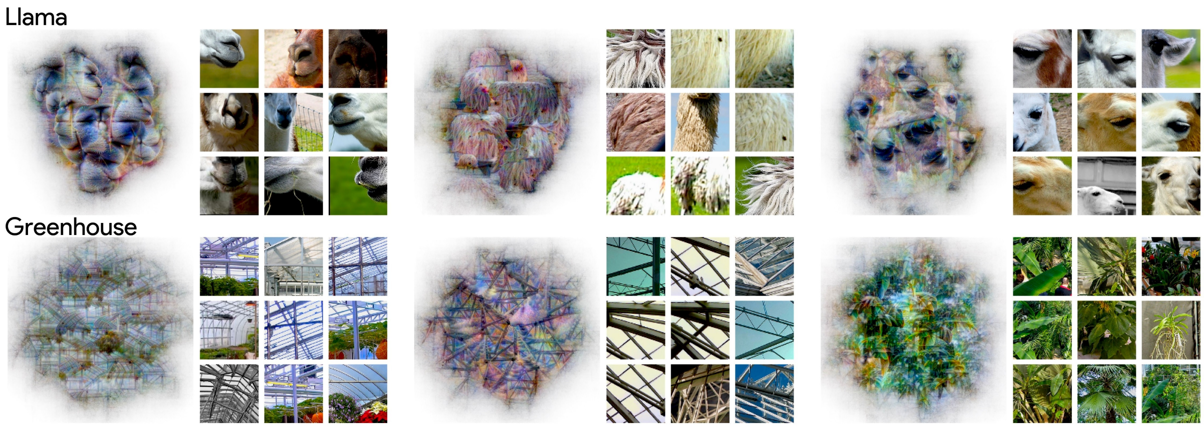

Figure 2. **Combining MACO with CRAFT [7] to visualize concepts.** We leverage CRAFT to identify significant concepts within a class – like a llama's mouth, fur, and eyes – and propose to combine these concepts with their corresponding feature visualizations that maximize the concept vector. The concepts are extracted from a ResNet50 trained on ImageNet. All visualizations for all ImageNet classes will be available in a public website.

they only depend on the model under study; unfortunately, in practice, they still fail for modern classification models (*e.g.,* ResNet50V2 [10] and ViT [4], see Figure 1). The third approach yields interpretable feature visualization even for these models but at the cost of major biases: in that case, it is impossible to untangle the true contributions of the model under study from those of the generative prior model. Herein, we present a new feature visualization method that is applicable to the largest, state-of-the-art networks without relying on any parametric prior image model.

Our proposed approach, called MAgnitude Constrained Optimization (MACO), builds up on the seminal work by Olah *et al*. who described the first method to optimize for maximally activating images in the Fourier space in order to penalize high-frequency content [19]. MACO also uses this phase/magnitude decomposition of the Fourier spectrum, but it solely optimizes the image phase while keeping its magnitude constant. Such a constraint is motivated by psychophysics experiments showing that humans are more sensitive to differences in phase than in magnitude [2, 8, 9, 11, 20].

2. Magnitude-Constrained Feature Visualization

The primary goal of a feature visualization method is to produce an image $\boldsymbol{x}^\star$ that maximizes a given criterion over some activations $\mathcal{A}$ that we denote $\mathcal{L}_{\mathcal{A}}(\boldsymbol{x}) \in \mathbb{R}$; usually some value aggregated over a structure in a neural network $\boldsymbol{f}$ (*e.g.*, neurons, channels, logits). A concrete example consists in finding a natural "prototypical" image $\boldsymbol{x}^\star$ of a class $k \in [\![1, K]\!]$ without using a dataset or generative models. However, optimizing in the pixel space $\mathbb{R}^{W\times H}$ is known to produce unrealistic images $\boldsymbol{x}^\star$, ridden with impulsional noise. Parameterizing the image in the Fourier space makes it possible to directly manipulate the image in the frequency domain. We propose to take a step further and decompose the Fourier spectrum $\boldsymbol{z}$ into its polar form $\boldsymbol{z} = \boldsymbol{r}e^{i\boldsymbol{\varphi}}$ instead of its cartesian form $\boldsymbol{z} = \boldsymbol{a} + i\boldsymbol{b}$, which allows us to disentangle the magnitude ($\boldsymbol{r}$) and the phase ($\boldsymbol{\varphi}$). To summarize, we formally introduce MACO:

Definition 2.1 (MACO)**.** *The feature visualization results from optimizing the parameter vector* $\boldsymbol{\varphi}$ *such that:*

$$\boldsymbol{\varphi}^\star = \underset{\boldsymbol{\varphi}\in\mathbb{R}^{W\times H}}{\arg\max}\, \mathbb{E}_{\boldsymbol{\tau}\sim\mathcal{T}}(\mathcal{L}_{\mathcal{A}}((\boldsymbol{\tau}\circ\mathcal{F}^{-1})(\boldsymbol{r}e^{i\boldsymbol{\varphi}}))$$
$$s.t.\quad \boldsymbol{r} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{D}}(|\mathcal{F}(\boldsymbol{x})|)$$

The feature visualization is then obtained by applying the inverse Fourier transform to the optimal complex-valued spectrum: $\boldsymbol{x}^\star = \mathcal{F}^{-1}((\boldsymbol{r}e^{i\boldsymbol{\varphi}^\star})$

Transparency for free Visualizations often suffer from repeated patterns or unimportant elements in the generated images. This can lead to readability problems or confirmation biases. It is important to ensure that the user is looking at what is truly important in the feature visualization. We take advantage of the fact that during backpropagation and we can obtain the intermediate gradients on the input $\partial\mathcal{L}_{\mathcal{A}}(\boldsymbol{x})/\partial\boldsymbol{x}$ for free as $\frac{\partial\mathcal{L}_{\mathcal{A}}(\boldsymbol{x})}{\partial\boldsymbol{\varphi}} = \frac{\partial\mathcal{L}_{\mathcal{A}}(\boldsymbol{x})}{\partial\boldsymbol{x}}\frac{\partial\boldsymbol{x}}{\partial\boldsymbol{\varphi}}$. We store these gradients throughout the optimization process and then average them to identify the areas that have been modified/attended to by the model the most during the optimization process.

3. Results

In this demo, we will showcase our results in the 1000 classes of ImageNet, including its applications to generating class maximizing images (Fig. 1), visualizations to illustrate internal representations of state-of-the-art vision transformer models, using feature inversion to elucidate which parts of the input are lost inside the model, and illustrating concepts discovered via CRAFT [7].

These results are also be publicly available in our website (Loupe) for everyone to browse and explore at leisure.

References

[1] M. Øygard Audun. Visualizing googlenet classes. *URL: https://www.auduno.com/2015/07/29/visualizing-googlenet-classes/*, 2(3). 1

[2] Terry Caelli and Paul Bevan. Visual sensitivity to two-dimensional spatial phase. *JOSA*, 72(10):1375–1381, 1982. 2

[3] Julien Colin, Thomas Fel, Rémi Cadène, and Thomas Serre. What i cannot predict, i do not understand: A human-centered evaluation framework for explainability methods. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[4] A Dosovitskiy, L Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, M Dehghani, Matthias Minderer, G Heigold, S Gelly, Jakob Uszkoreit, and N Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2

[5] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019. 1

[6] Thomas Fel, Remi Cadene, Mathieu Chalvidal, Matthieu Cord, David Vigouroux, and Thomas Serre. Look at the variance! efficient black-box explanations with sobol-based sensitivity analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[7] Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[8] Evgeny Gladilin and Roland Eils. On the role of spatial phase and phase correlation in vision, illusion, and cognition. *Frontiers in Computational Neuroscience*, 9:45, 2015. 2

[9] Nathalie Guyader, Alan Chauvin, Carole Peyrin, Jeanny Hérault, and Christian Marendaz. Image phase or amplitude? rapid scene categorization is an amplitude-based process. *Comptes Rendus Biologies*, 327(4):313–318, 2004. 2

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[11] Olivier R Joubert, Guillaume A Rousselet, Michele Fabre-Thorpe, and Denis Fize. Rapid visual categorization of natural scene contexts with equalized amplitude spectrum and increasing phase noise. *Journal of Vision*, 9(1):2–2, 2009. 2

[12] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5188–5196, 2015. 1

[13] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015. 1

[14] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4467–4477, 2017. 1

[15] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, 29, 2016. 1

[16] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 1

[17] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. *Visualization for Deep Learning workshop, Proceedings of the International Conference on Machine Learning (ICML)*, 2016. 1

[18] Paul Novello, Thomas Fel, and David Vigouroux. Making sense of dependence: Efficient black-box explanations using dependence measure. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[19] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 1, 2

[20] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981. 2

[21] Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 1

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. In *Workshop on Visualization for Deep Learning, Proceedings of the International Conference on Machine Learning (ICML)*, 2017. 1

[25] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[26] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 1

[27] Mike Tyka. Class visualization with bilateral filters. 2016. *URL: https://mtyka. github. io/deepdream/2016/02/05/bilateral-class-vis. html*, 2(3). 1

[28] Donglai Wei, Bolei Zhou, Antonio Torrabla, and William Freeman. Understanding intra-class knowledge inside cnn. *arXiv preprint arXiv:1507.02379*, 2015. 1