ON THE DIFFERENCE BETWEEN BUILDING AND EXTRACTING PATTERNS: A CAUSAL ANALYSIS OF DEEP GENERATIVE MODELS.

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are important tools to capture and investigate the properties of complex empirical data. Recent developments such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs) use two very similar, but *reverse*, deep convolutional architectures, one to generate and one to extract information from data. Does learning the parameters of both architectures obey the same rules? We exploit the causality principle of independence of mechanisms to quantify how the weights of successive layers adapt to each other. Using the recently introduced Spectral Independence Criterion, we quantify the dependencies between the kernels of successive convolutional layers and show that those are more independent for the generative process than for information extraction, in line with results from the field of causal inference. In addition, our experiments on generation of human faces suggests that enforcing more independence between successive layers of generators may lead to better performance and modularity of these architectures.

1 INTRODUCTION

Deep generative models have proven powerful in learning to design realistic images in a variety of complex domains (handwritten digits, human faces, interior scenes). In particular, two approaches have recently emerged: Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which train an image generator by having it fool a discriminator that should tell apart real from artificially generated images; and Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) which learn both a mapping from latent variables to the data (the decoder) and the converse mapping of the data to latent variables (the encoder), such that correspondences between latent variables and data features can be easily investigated. Although these architecture have been lately the subject of extensive investigations, understanding why and how they work, and how they can be improved, remains elusive.

An interesting feature of GANs and VAEs is that they both involve the learning of two deep sub-networks. These sub-networks have a "mirrored" architecture, as they both consists in a hierarchy of convolutional layers, but information flows in opposite ways: generators and decoders mapping to the data space, while discriminators and encoders extract information from the same space. Interestingly, this difference could be framed in a causal perspective, with information flowing in the causal direction from the putative causes of variations in the observed data, while extracting high level properties from observations would operate in the anti-causal direction.

Generative models in machine learning are usually not required to be causal, as modeling the data distribution is considered to be the goal to achieve. However, the idea that a generative model able to capture the causal structure of the data by disentangling the contribution of independent factors should perform better has been suggested in the literature (Bengio et al., 2013a; Mathieu et al., 2016) and evidence supports that this can help the learning procedure (Bengio et al., 2013b). Although many approaches have been implemented on specific examples, principles and automated ways of learning disentangled representation from data remains largely an open problem both for learning representations (targeting supervised learning applications) and for fitting good generative models. GANs have for example recently been a subject of intensive work in this direction, leading

to algorithms disentangling high level properties of the data such as InfoGans (Chen et al., 2016) or conditional GANs (Mirza & Osindero, 2014). However such models require supervision (e.g. feeding digit labels as additional inputs) to disentangle factors of interest.

We propose that the coupling between high dimensional parameters can be quantified and exploited in a causal framework to infer whether the model correctly disentangles independent aspects of the deep generative model. This hypothesis relies on recent work exploiting the postulate of Independence of Cause and Mechanism stating that Nature chooses independently the properties of a cause and those of the mechanism that generate effects from the cause (Janzing & Schölkopf, 2010; Lemeire & Janzing, 2012). Several methods relying on this principle have been proposed in the literature in association to different model classes (Janzing et al., 2010; Zscheischler et al., 2011; Daniusis et al., 2010; Janzing et al., 2012; Shajarisales et al., 2015; Sgouritsa et al., 2015; Schölkopf et al., 2012). Among these methods, the Spectral Independence Criterion (SIC) (Shajarisales et al., 2015) can be used in the context of linear dynamical systems, which involve a convolution mechanism.

In this paper, we show how SIC can be adapted to investigate the coupling between the parameters of successive convolutional layers.Empirical investigation shows that SIC is approximately valid between successive layers of generative models, and suggest SIC violations indicate deficiencies in the learning algorithm or the architecture of the network. Interestingly, and in line with theoretical predictions (Shajarisales et al., 2015), SIC tends to be more satisfied for generative sub-networks (mapping latent variables to data), than for the part of the system that map the anti-causal direction (data to latent variables). Overall, our study suggests that enforcing Independence of Mechanisms can help design better generative models.

2 BACKGROUND

2.1 WHY A CAUSAL PERSPECTIVE?

An insightful description of causal model is based on Structural Equations (SEs) of the form

$$Y \coloneqq f(X_1, X_2, \cdots, X_N, \epsilon)$$

The right hand side variable in this equation may or may not be random, and the additional independent noise term ϵ representing exogenous effects (originating from outside the system under consideration) may be absent. The $\coloneqq$ sign indicates the asymmetry of this expression and signifies that the left-hand-side variable is computed from the right-hand-side expression but not the opposite. This expression stays valid if something selectively changes on the right hand side variables, for example if the value of X_1 is externally forced to stay constant (hard intervention) or if the shape of f or of an input distribution changes (soft intervention). These properties account for the robustness or invariance that is expected from causal models with respect to purely probalistic ones. Assume now X_1 itself is determined by other variables according to

$$X_1 \coloneqq g(U_1, U_2)$$

then the resulting structural causal model summarized by this system of equations also implies a modularity assumption: one can intervene on the second equation while the first one stays valid, baring the changes in the distribution of X_1 entailed by the intervention.

This structural equation framework describes well what would be expected from a robust generative model. Assume the model generates faces, one would like to be able to intervene on e.g. the pose without changing the rest of the top-level parameters (hair color,...), and still be able to observe a realistic output. One could also expect intervening on specific operations performed at intermediate levels by keeping the output "qualitatively" of the same nature. For example, we can imagine that by slightly modifying the mapping that positions the eyes with respect to the nose on a face, one generates different heads that do not match exactly human standards, but are human-like. What we would not want is instead to see artifacts emerging all over the generated image, or edges being blurred.

2.2 THE PRINCIPLE OF INDEPENDENCE OF MECHANISMS

Assume we have two variables X and Y, possibly multidimensional and neither necessarily from a vector space, nor necessary random. Assume the data generating mechanism obeys the following

structural equation:

$$Y := m(X)$$

with m the mechanism, X the cause and Y the effect. We rely on the postulate that properties of X and m are "independent" in a sense that can be formalized in various ways (Peters et al., 2017). The present analysis will be based on a formalization explained below. Classical Independence of Cause and Mechanism(ICM)-based methods address the problem of cause-effect pair inference, where both directions of causation are plausible. They then take advantage of specific settings for which it is possible to prove that if the assumption is valid for the true generative model, the converse independence assumption is very likely violated (with high probability) for the anticausal or backward model

$$X := m^{-1}(Y)$$

(independence is then computed between m^{-1} and Y). As a consequence, the true causal direction can be identified by evaluating the independence assumptions of both models and pick the direction for which independence is the most likely.

2.3 Spectral Independence

We introduce a specific formalization of independence of mechanisms that is well suited to study convolutional layers of neural network. This relies on representing signals or images in the Fourier domain.

2.3.1 Background on discrete signals and images

The Discrete-time Fourier Transform (DTFT) of a sequence $a = \{a[k], k \in \mathbb{Z}\}$ is defined as

$$\widehat{\mathrm{a}}(\nu) = \textstyle\sum_{k\in\mathbb{Z}} a[k] \exp(-\mathbf{i}2\pi\nu k),\ \nu \in \mathbb{R}$$

Note that the DTFT of such sequence is a continuous 1-periodic function of the normalized frequency ν. By Parseval's theorem, the energy can be expressed in the Fourier domain by $\|\mathbf{a}\|_2^2 = \int_{-1/2}^{1/2} |\widehat{a}(\nu)|^2 d\nu$. To simplify notations, we will denote by $\langle . \rangle$ the integral (or average) of a function over the unit interval $\mathcal{I}$, that is, $\|a\|_2^2 = \langle |\widehat{\mathrm{a}}|^2 \rangle$.

The Fourier transform can be easily generalized to 2D signals, leading to a 2D function 1-periodic with respect to both arguments

$$\widehat{\mathrm{b}}(u,v) = \textstyle\sum_{k\in\mathbb{Z}, l\in\mathbb{Z}} a[k,l] \exp(-\mathbf{i}2\pi(uk+vl),\ (u,v) \in \mathbb{R}^2\,.$$

2.3.2 SIC postulate

Assume now that our cause-effect pairs X and Y are weakly stationary time series. This means that the power of these signals can be decomposed in the frequency domain using their Power Spectral Densities (PSD) $S_x(\nu)$ and $S_y(\nu)$. We assume that Y is the output of a Linear Time (translation) Invariant Filter with convolution kernel h receiving input X such that

$$Y = \{\textstyle\sum_{\tau\in\mathbb{Z}} h_\tau X_{t-\tau}\} = h * X, \tag{1}$$

In that case input and output PSDs are related by the formula $S_y(\nu) = |\widehat{\mathrm{h}}(\nu)|^2 S_x(\nu)$ for all frequencies ν. The Spectral Independence Postulate consists then in assuming that the power amplification of the filter at each frequency does not adapt to the input power spectrum, i.e. the filter will not tend to selectively amplify or attenuate the frequencies with particularly large or low power. This can be translated into the fact that output power factorizes in the product of input power times the energy of the filter, leads to the postulate (Shajarisales et al., 2015):

Postulate 1 (Spectral Independence Criterion (SIC))**.** *Let S_x be the Power Spectral Density (PSD) of a cause X and h the system impulse response of the causal system of (1), then*

$$\textstyle\int_{-1/2}^{1/2} S_x(\nu)|\widehat{\mathrm{h}}(\nu)|^2 d\nu = \int_{-1/2}^{1/2} S_x(\nu) d\nu \cdot \int_{-1/2}^{1/2} |\widehat{\mathrm{h}}(\nu)|^2 d\nu\,, \tag{2}$$

holds approximately.

It can be shown that (2) is violated in the backward direction under mild assumptions when the filter is invertible. We can define a scale invariant quantity ρ_{SIC} measuring the departure from the SIC assumption, i.e. the dependence between input power spectrum and frequency response of the filter: the Spectral Dependency Ratio (SDR) from X to Y is defined as

$$\rho_{SIC} := \frac{\langle S_x \cdot |\widehat{\mathrm{h}}|^2 \rangle}{\langle S_x \rangle \langle |\widehat{\mathrm{h}}|^2 \rangle} \,, \tag{3}$$

where $\langle . \rangle$ denotes the integral over the unit frequency interval. The values of this ratio can be interpreted as follows: $\rho_{SIC} \approx 1$ reflects spectral independence, while $\rho_{SIC} > 1$ reflects a *correlation* between the input power and frequency response of the filter (the filter selectively amplifies input frequency peaks of large power, leading to anomalously large output power), conversely $\rho_{SIC} < 1$ reflect *anticorrelation* between these quantities. We will use these terms to analyze our experimental results. In addition (Shajarisales et al., 2015) also derived theoretical results suggesting that if $\rho_{SIC} \approx 1$ for an invertible causal system, then $\rho_{SIC} < 1$ in the anticausal direction. These interpretations of SDR values are summarized in Fig. 3b, which can be use to understand experimental results.

3 Independence of mechanisms in deep networks

We now introduce our causal reasoning in the context of deep convolutional networks, where the output of successive layers are often interpreted as different levels of representations of an image, from detailed low level features to abstract concepts. We thus investigate whether a form of modularity between successive layer can be identified using the above framework.

3.1 Strided convolution units and independence between scales

DCGANs have successfully exploited the idea of convolutional networks to generate realistic images (Radford et al., 2015). While strided convolutional units are used as pattern detectors in deep convolutional classifiers, they obviously play a different role in generative models: they are pattern producers. We provide in Fig. 1 (left) a toy example to explain their potential ability to generate independent features at multiple scales. On the picture the activation of one channel at the top layer (top left) may encode the (coarse) position of the pair of eyes in the image. After convolution by a first kernel, activations in a second layer indicate the location of each eye. Finally, a kernel encoding the shape of the eye convolves this input to generate an image that combines the three types of information, distributed over three different spatial scales.

What we mean by assuming independence of the features encoded at different scales can be phrased as follows: there should be typically no strong relationship between the shape of patterns encoding a given object at successive scales. Although counter examples may be found, we postulate that a form of independence may hold approximately, for naturalistic images that possess a large number of different multiscale features. A case where this assumption may be violated for a deep network is given in Fig. 1, where a long edge in the image cannot be captured by a single convolution kernel (because of the kernel size limitation to 3 by 3 in this case). Hence, an identical kernel needs to be learned at an upper layer in order to control the precise alignment of activations in the bottom layer. Any misalignment between the orientation of the kernels would lead to a very different pattern, and witness that information in the two successive layers is entangled.

3.2 SIC between successive (de)convolutional units

To connect dependency between convolutional units to SIC, that is intuitively to this context, we must take into account two things: the striding adds spacing between input pixels in order to progressively increase the dimension and resolution of the image from one layer to the next, and there is a non-linearity between successive layers. Striding can be easily modeled as it amounts to upsampling the input image before convolution (see illustration Fig. 1). We denote $.^{\uparrow s}$ the upsampling operation with integer factor[1] s that turns the 2D tensor x into

$$x^{\uparrow s}[k,l] = \begin{cases} x[k/s, l/s], & k \text{ and } l \text{ multiple of s} \\ 0 & \text{otherwise.} \end{cases}$$

[1] s is the inverse of the stride parameter, which is fractional in that case

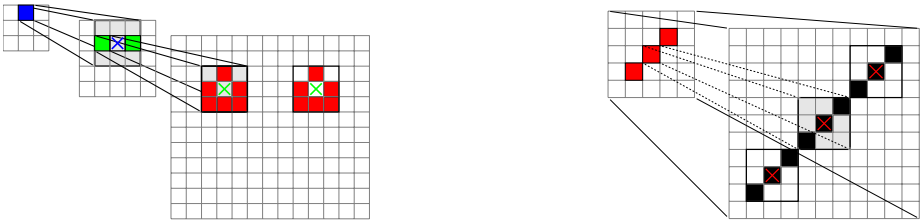


Figure 1: Left: example of combination of coarse and finer scale features using two convolution kernels in successive layers to form the eyes of a human face. Right: Example of violation of independence of mechanisms between two successive layers. Crosses indicate center of patch where each active pixel of the previous layer maps to.

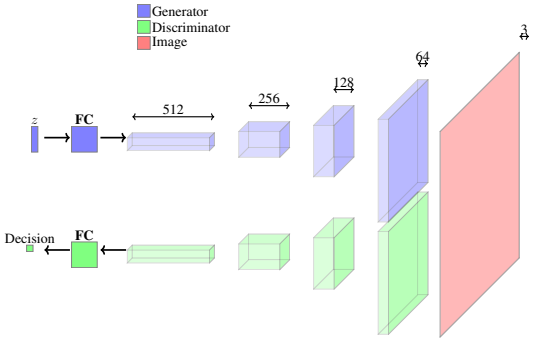


Figure 2: Architeture of the pretrained DCGAN generator used in our experiments

Interestingly, the definition implies that striding amounts to a compression of the normalized frequency axis in the Fourier domain with $\widehat{x^{\uparrow s}}(u,v) = \hat{x}(su, sv)$.

Next, the non-linear activation between successive layers is more challenging to study. We will thus make the simplifying assumption that rectified linear Units are used (ReLU), such that a pixel is either linearly dependent on small variations of its synaptic inputs, or the pixel is not active at all. We then follow the idea that ReLU activations may be used as a switch that controls the flow of relevant information in the network (Tsai et al., 2016; Choi & Kim, 2017). . Hence, for convolution kernels in successive layers that encode different aspects of the same object, we make the assumption that their corresponding outputs will be frequently coactivated, such that the linearity assumption may hold. If pairs of kernels are not frequently coactivated, this suggests that they do not encode the same object and it is thus unlikely that their weights would be dependent.

We thus write down the mathematical expression for the application of two successive deconvolutional units to a given input tensor using linearity assumptions leading to $y = g * x^{\uparrow s} = g * (h * z)^{\uparrow s}$. In the Fourier domain, this implies

$$\hat{x}(u,v) = \hat{g}(u,v)\hat{y}(su,sv) = \hat{g}(u,v)\hat{h}(su,sv)\hat{z}(s^2u, s^2.v)$$

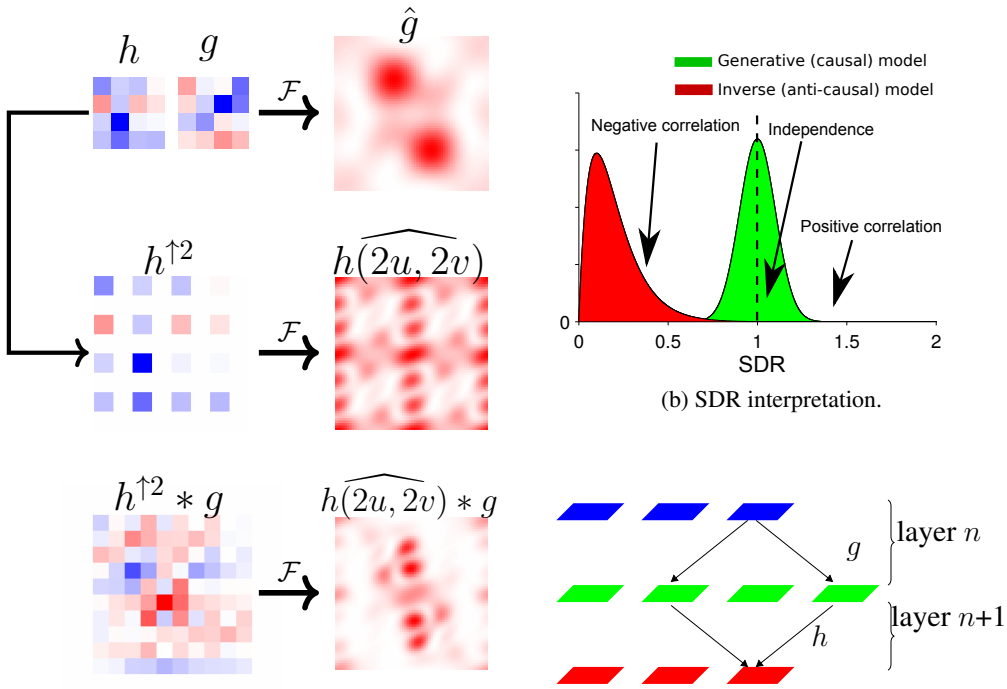
In order to have a criterion independent from the incoming data, we assume the input z has no spatial structure (e.g. z is sampled from a white noise), and thus its power is uniformly distributed across spatial frequencies (the PSD is flat). Then the spatial properties of x are entirely determined by h and the SIC criterion of equation 2 writes

$$\left\langle |\hat{\mathrm{g}}(u,v).\hat{h}(su,sv)|^2 \right\rangle = \left\langle |\widehat{\mathrm{g}}(u,v)|^2 \right\rangle \left\langle |\widehat{\mathrm{h}}(su,sv)|^2 \right\rangle , \tag{4}$$

by denoting the double integral by angular brackets. We can write down an updated version of the SDR in equation 3 corresponding to testing SIC between a cascade of two filter (each taken from one of two successive layers) as

$$\rho_{SIC} = \frac{\left\langle |\hat{\mathrm{g}}(u,v).\hat{h}(su,sv)|^2 \right\rangle}{\left\langle |\widehat{\mathrm{g}}(u,v)|^2 \right\rangle \cdot \left\langle |\widehat{\mathrm{h}}(u,v)|^2 \right\rangle} ,$$

which we will evaluate in the experiments.



(b) SDR interpretation.

(c) Convolution pathways between layers.

(a) Illustration of successive convolutions.

Figure 3: 3a Kernels and corresponding Fourier transforms (zero frequency at the center of each picture). We can see that there is a slight anticorrelation between the absolute values of Fourier transforms of both kernel, resulting in an SDR of 0.88. 3b Illustration of the meaning of SDR values. 3c Illustration of the multiple combination of convolution kernels between successive layers. The pathway depend on the input (blue), output (red) and intermediate or interlayer (green) channels.

4 EXPERIMENTS

We used a version of the DCGANs pretrained on the CelebFaces Attributes Dataset (CelebA)[2]. The structures of the generator and discriminator are summarized in Fig. 2 and contains 4 convolutional layers. We also performed experiments with a plain VAEs[3] with a similar convolutional architecture. Each layer uses 4 pixels wide convolution kernels for the GAN, and 5 pixels wide for the VAE. In all cases, layers are enumerated following the direction of information flow during test. We will talk about coarse scale for layers that a far from the image, and fine scale for the ones that are close to it in the hierarchy.

4.1 SIC BETWEEN SUCCESSIVE DECONVOLUTIONAL UNITS

We first illustrate the theory by showing in Fig. 3a an example of two deconvolution kernels between successive layers (1 and 2). It shows a slight anti-correlation between the absolute values of the Fourier transforms of both kernels, resulting I a SDR of .88. One can notice the effect of kernels g (upper layer) and h (lower layer) on the resulting convolved kernel in the Fourier domain: the kernel from the upper layer tends to modulate the fast variations of $g * h$ in the Fourier domain, while h affects the 'slow' variations. This is a consequence of the design of the strided fractional convolution, which has for consequence to foster a decoupling between the kernel parameters of each layers. We use this approach to characterize the amount of dependency between successive layers by plotting the histogram of the SDR that we get for all possible combination of kernels belonging to each layer, which involve computing it across all possible pathways between input and output channels,

[2] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[3] https://github.com/yzwxx/vae-celebA

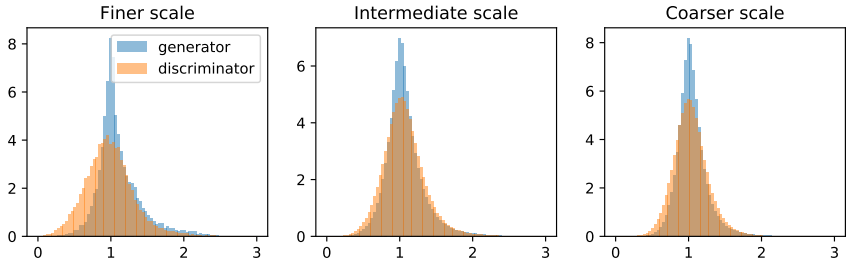


Figure 4: Superimposed histograms of SIC generic ratios for layers at the same level of resolution (from lower to higher).

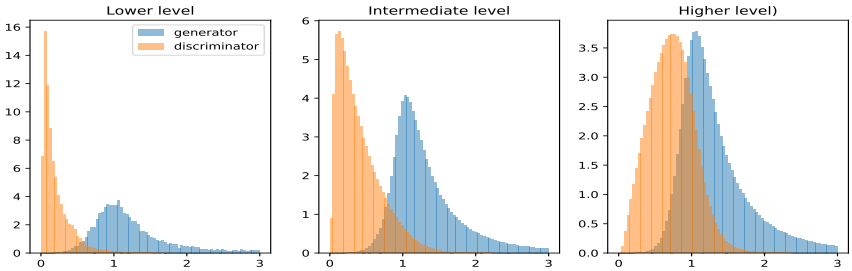


Figure 5: Comparison of SDR statistics between decoder (causal) and encoder (anticausal) of a trained VAE.

as described in Fig. 3c. The result is shown in Fig. 4, witnessing a rather good concentration of the SDR around 1, which suggest are rather good independence of the convolution kernels between successive layers. It is interesting to compare these histograms to the values computed for the discriminator of the GAN, which implements convolutional layers of the same dimensions in reverse order. The result also shown in Fig. 4 exhibits a broader distribution of the SDR, especially for layers encoding lower level image features. This is in line with the ICM principle, as the discriminator is operating in the anticausal direction. However, the difference between the generator and discriminator is not strong, suggesting that ICM might be improved for the generator, possibly leading to better quality images. In addition, the weak difference between causal and anti-causal direction may also be due to the fact that the discriminator does not implement an inversion of the putative causal model, but only extract relevant latent information for discrimination. In order to check our method on a generative model including a network mapping back the input to latent variable, we applied it to a trained VAE. The results presented in Fig. 5 show much sharper differences between generator (decoder) and encoder. The shape of the histograms are matching prediction from (Shajarisales et al., 2015) shown in Fig. 3b. The difference with GANs can be explained by the fact that VAEs are indeed performing an invertion of the generative model, leading to very small SDR values in the anticausal direction. We also note overall a much broder distribution of VAE SDRs in the causal direction (decoder). Interestingly, the training of the VAE did not lead to values as satisfactory are what we obtained with the GAN. Examples generated from the VAE are shown in appendix (Fig. 8). This suggests that dependence between cause and mechanism may reflect a suboptimal performance of the generator.

4.2 Removing correlated units

We saw in the above histograms that while many successive convolutional unit had a SDR close to one, there are tails composed of units exhibiting either a ratio inferior to one (reflecting negative correlation between the Fourier transform of the corresponding kernels) or a ratio larger than one (reflecting positive correlation). Interestingly, if we superimpose the histograms (see Fig. 4) of the lower level layer of the generator and discriminator networks, we see that these tails are quite different between networks, showing more negative correlations for the discriminative networks, while the positive correlation tail of the generative networks remains rather strong. This suggests

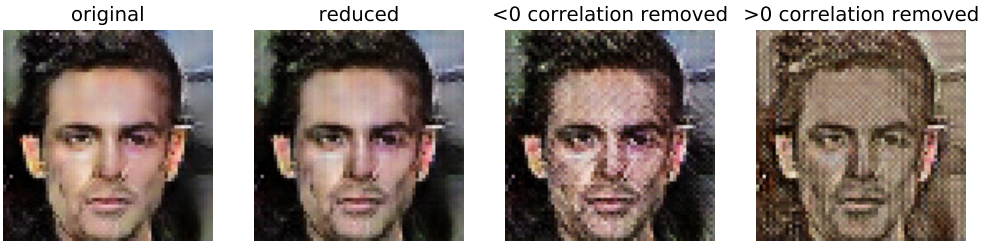


Figure 6: Example generated figures using a pretrained DGAN (left column). Second column: the output of the same network when removing low energy filters from the third layer (reduction of the number of filters from 8,192 to 4,260). Third column: the output of the same network when removing the filters leading to the lowest average generic ratio ($\rho < .9$, leading to 3,684 filters). Fourth column: same when removing the filters leading to the largest average generic ratio ($\rho > 1.45$, leading to 3,660 filters). More examples are shown in appendix Fig. 7.

that negative and positive correlation are qualitatively different phenomena, in line with our analysis of the matrix factorization algorithms. In order to investigate the nature of filters exhibiting different signs of correlation, we selectively removed filters of the third layer of the generator (the last but one), based on the magnitude (above or below one) of the average SDR that each filter achieved when combined with any of the filters in the last layer. In order to check that our results were not influenced by filters with very small weights (that still can exhibit correlations), we zeroed the filters having the kernels with the smallest energy, while maintaining an acceptable quality of the output of the network (see Fig. 6 second column). This removed around half of the filters of the third layer. Then we remove additional filters exhibiting either large or small (anti-correlated) generic ratio, such that the same proportion of filters is removed (see Fig. 6 third and fourth column). It appears clearly from this result that filters exhibiting large positive or negative correlation do not play the same role in the network. From the quality of the generated images, filters from the third layer negatively correlated to filters from the fourth seem allow correction of checkerboard artifacts, potentially generated as side effect of the fractional strided convolution mechanisms[4]. Despite a decrease in texture quality, the removal of such weight does not distinctively affect an informative aspect of the generated images. Conversely, removing positively correlated filters lead to a disappearance of the color information in a majority of the regions of the image. This suggests that such filters do encode the color structure of images, and the introduced positive correlation between the filters from the third and fourth layer may result from the fact that uniform color patches corresponding to specific parts of the image (hair, skin) need a tight coordination of the filters at several scales of the image in order to produce large uniform areas with sharp border.

5 Discussion

In this work, we derive a measure of independence between the weights learned by convolutional layers of deep networks. The results suggests, as predicted in many cases in the field of causal inference, that generative models that map latent variables to data tend to have more independence between successive layers than discriminative or encoding networks. This suggests that enforcing independence during training may help build better generative models. Moreover, it also indicates on which layers should be modified to improve the performance. Finally, we speculate that independence between successive layers, by favoring modularity of the network, may help build architecture that can be easily adapted to new purposed. In particular, separation of spatial scales in such model may help build architectures where on can intervene on one scale without affecting others, with applications such as style transfer Gatys et al. (2015). One specific feature of our approach is that this quantitative measure of the networks in not statistical and as such does not require test samples to be computed. Only the parameters of the model are used, which makes the approach easy to apply to any neural network equipped with convolutional layers.

[4] https://distill.pub/2016/deconv-checkerboard/

REFERENCES

Y Bengio, A Courville, and P Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013a.

Y Bengio, G Mesnil, Y Dauphin, and S Rifai. Better mixing via deep representations. In *ICML 2013*, 2013b.

X Chen, Y Duan, R Houthooft, J Schulman, I Sutskever, and P Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. jun 2016.

Jae-Seok Choi and Munchurl Kim. A deep convolutional neural network with selection units for super-resolution. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pp. 1150–1156. IEEE, 2017.

P. Daniusis, D. Janzing, K. Mooij, J. Zscheischler, B. Steudel, K. Zhang, and B. Schölkopf. Inferring deterministic causal relations. In *UAI 2010*, 2010.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

D. Janzing and B. Schölkopf. Causal inference using the algorithmic Markov condition. *Information Theory, IEEE Transactions on*, 56(10):5168–5194, 2010.

D. Janzing, P.O. Hoyer, and B. Schölkopf. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.

D. Janzing, J. Mooij, K. Zhang, J. Lemeire, J. Zscheischler, P. Daniušis, B. Steudel, and B. Schölkopf. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182–183:1–31, 2012.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

J. Lemeire and D. Janzing. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, pp. 1–23, 7 2012. doi: 10.1007/s11023-012-9283-1.

M F Mathieu, J J Zhao, A Ramesh, P Sprechmann, and Y LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5041–5049, 2016.

M Mirza and S Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, 2017.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *ICML 2012*, 2012.

E. Sgouritsa, D. Janzing, P. Hennig, and B. Schölkopf. Inference of cause and effect with unsupervised inverse regression. In *ICML 2015*, 2015.

N. Shajarisales, D. Janzing, B. Schölkopf, and M. Besserve. Telling cause from effect in deterministic linear dynamical systems. In *ICML 2015*, 2015.

Chuan-Yung Tsai, Andrew M Saxe, and David Cox. Tensor switching networks. In *Advances in Neural Information Processing Systems*, pp. 2038–2046, 2016.

J. Zscheischler, D. Janzing, and K. Zhang. Testing whether linear equations are causal: A free probability theory approach. In *UAI 2011*, 2011.

APPENDIX

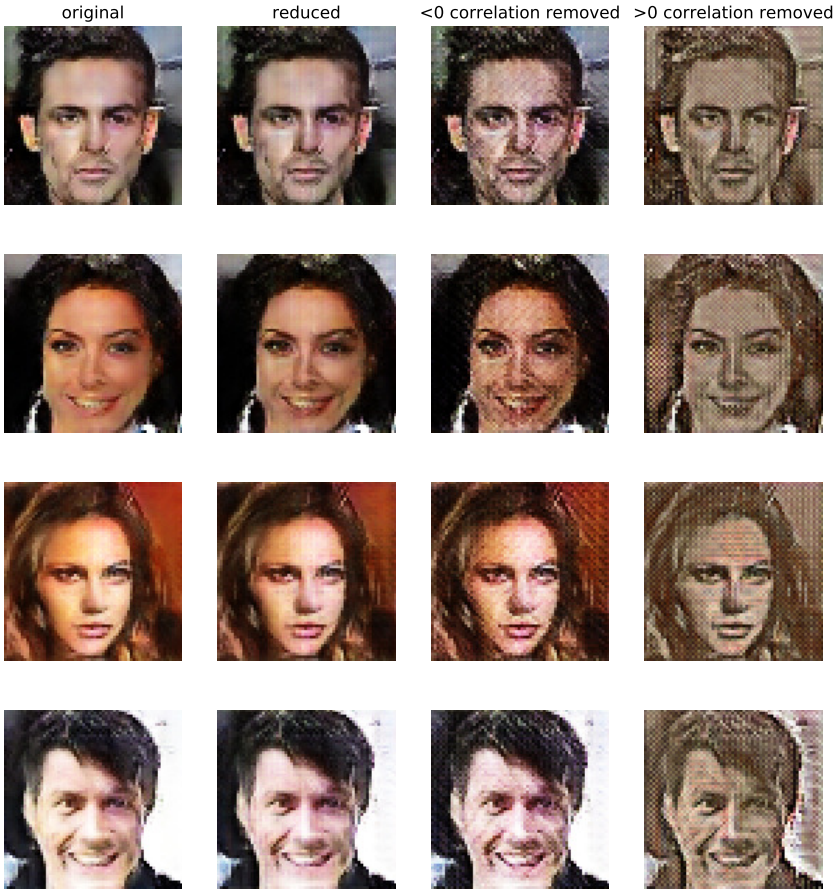


Figure 7: Example generated figures using a pretrained DGAN (left column). Second column: the output of the same network when removing low energy filters from the third layer (reduction of the number of filters from 8,192 to 4,260). Third column: the output of the same network when removing the filters leading to the lowest average generic ratio ($\rho < .9$, leading to 3,684 filters). Fourth column: same when removing the filters leading to the largest average generic ratio ($\rho > 1.45$, leading to 3,660 filters).

Figure 8: Random examples generated by the trained VAE.