# TOWARD A NEURO-INSPIRED CREATIVE DECODER

**Payel Das**[*], **Brian Quanz**[*], **Pin-Yu Chen**, **Jae-wook Ahn**, **Dhruv Shah**
IBM Research - {daspa,blquanz}@us.ibm.com

## ABSTRACT

Creativity, a process that generates novel and meaningful ideas, involves increased association between task-positive (control) and task-negative (default) networks in the human brain. Inspired by this seminal finding, we propose a creative decoder within a deep generative framework, which involves direct modulation of the neuronal activation pattern after sampling from the learned latent space. The proposed approach is fully unsupervised and usable off-the-shelf. Novelty metrics and human evaluation were used to evaluate the creative capacity. Our experiments on different image datasets (MNIST, FMNIST, MNIST+FMNIST, WikiArt and CelebA) reveal that atypical co-activation of highly and weakly activated neurons in a deep decoder promotes generation of novel and meaningful artifacts.

## 1 INTRODUCTION

Creativity is defined as a process that produces novel and valuable (*aka* meaningful) ideas Boden (2004). In early days of computational creativity research, expert feedback Graf & Banzhaf (1995) or evolutionary algorithms with hand-crafted fitness functions DiPaola & Gabora (2009) were used to guide a model's search process to make it creative. However, those methods reportedly lack exploration capability.

Data-driven approaches like deep learning open a new direction – enabling the study of creativity from a knowledge acquisition perspective. Deep Dream Szegedy et al. (2015) and Deep Style Transfer Gatys et al. (2015) have aroused substantial interest in computational creativity research. Only recently, novelty generation using powerful deep generative models, such as Variational Autoencoders (VAEs)Kingma & Welling (2013); Rezende & Mohamed (2015) and Generative Adversarial Networks (GANs) Goodfellow et al. (2014), have been attempted. These are designed to build a model from and generate known objects (*i.e.*, images), and discourage out-of-distribution generation to avoid instability and minimize spurious sample generation Kégl et al. (2018), limiting their potential in creativity research. Therefore, new approaches to enhance the creative capacity of generative models are needed.

One path is to get inspiration from the cognitive processes associated with human creativity. How does the human brain produce creative ideas? It is an interesting question and a central topic in cognitive neuroscience research. Recent neuroimaging studies Beaty et al. (2018); Shi et al. (2018); Gao et al. (2017) suggest stronger coupling of the default mode network and executive control network in creative brains across a range of creative tasks and domains, from divergent thinking to poetry composition to musical improvisation (see Fig. 1A-B). Brain networks are the large-scale communities of interacting brain regions, as revealed by the resting-state functional correlation pattern; these networks correspond to distinct functional systems of the brain. The default mode or task-negative network is associated with spontaneous and self-generated thought, and, therefore implicated in idea generation. The control or task-positive network, in contrast, is associated with cognitive processes requiring externally directed attention.

Default and control networks often exhibit an antagonistic relation during rest and many cognitive tasks, including working memory Anticevic et al. (2012). This antagonistic relation likely reflects suppression of task-unrelated thoughts during cognitive control. Dynamic coupling of default and control networks has been previously reported during goal-directed, self-generated thought processes Spreng et al. (2015). Recently Beaty et al. (2016) proposed that stronger coordination between default and control networks contributes to creative idea generation.

[*] Payel Das and Brian Quanz contributed equally to this work and are corresponding authors

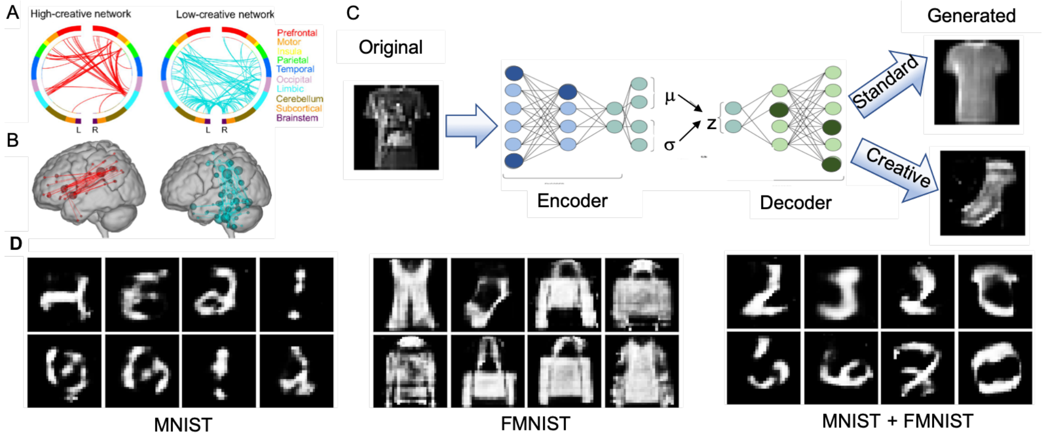

Figure 1: **A-B.** Depictions (A: circle plots, B: glass brains) of high- and low-creative networks in human brain with their highest degree nodes. Circle plot colors correspond to brain lobes in left(L) and right (R) hemispheres. Adapted from Beaty et al. (2018). **C.** Depiction of a VAE model with neuro-inspired *creative* decoder. Normally, a small % of neurons in each hidden layer is low-active (dark color). Inspired by neural basis of creativity, we activate those “low-active” neurons to induce coupling with task-positive neurons during “creative” decoding. **D.** Samples generated by *creative* decoding and human-annotated as creative with high confidence.

Motivated by neuroimaging findings suggesting stronger coupling between task-positive and task-negative neurons in creative brains, this work attempts to induce creativity in deep generative models by proposing a *creative* decoder. In a nutshell, the *creative* decoder aims to generate creative samples from the original latent (concept) space by favoring atypical co-activation of high- and low-active neurons (neuron groups derived by roughly modeling the task-negative and task-positive concepts), while the generative model training remains unchanged. To our knowledge, this is the first work that aims to enhance creative capacity of a deep generative model in a neuro-inspired manner, and further can be adapted to any decoder/generator and applied without additional training or data.

We employ human annotation, in addition to using surrogate metrics such as reconstruction distance, for creativity evaluation of the samples generated by the *creative* decoder and compare to a number of baseline approaches. A VAE model was used as the base generative framework (see Fig 1C). Results show that enhanced creativity can result from the neuro-inspired atypical activation as opposed to a simple random or structured noise effect. We show the performance of the proposed method against MNIST digits , FMINST fashion objects , and on a combined MNIST plus FMINST dataset. We also present results on the WikiArt art images and CelebA faces.

## 2 Related Work

There are several related works on novelty generation with deep generative models. One line of work is on systems designed to produce creative outcomes, but these require supervision either in the form of pre-set features to compose (Nguyen et al. (2015)), creative inputs and labels (Elgammal et al. (2017)), or human feedback to make the network creative-explorative (Cherti & Kégl (2016)), whereas our method is wholly unsupervised, exploiting a reported neural basis of creativity, and does not require alternate training or inference. Another line of related work is is few/one-shot learning and generation (Lake et al. (2015); Rezende et al. (2016); Clouâtre & Demers (2019)), a step toward out-of-class generation. However, these require novel exemplars at test time, unlike our method.

## 3 Methodology

***Creative* decoding scheme.** To capture the spirit of the atypical neuronal activation pattern observed in a creative human brain, *i.e.*, dynamic interaction between a task-positive (control) and a task-negative (default) brain network, we propose a probabilistic decoding scheme. After sampling a latent representation $z$ (e.g., in a GAN or VAE), the proposed method selects decoder neurons based on their activation correlation pattern, and then activates a few “off" neurons along with the “on” neurons.

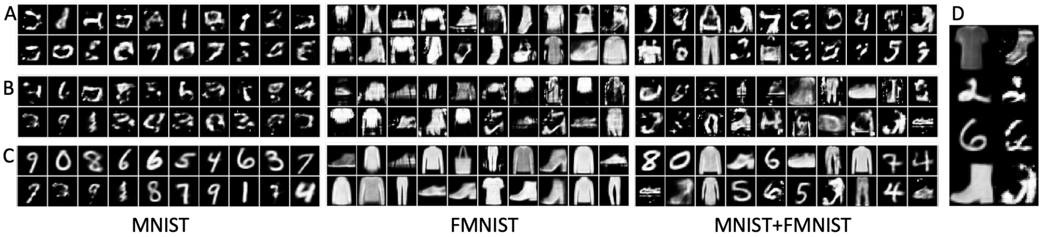

Figure 2: Samples human-annotated as (A) creative, (B) novel but not-creative, and (C) not novel or creative. D: How “creative” decoding modifies generation (combined dataset): **left:** regular; **right:** low-active decoding.

Specifically, we randomly select from “off” neurons that were weakly-activated in the trained decoder across most of the training data - referred to as the “low-active” method. “Off” low-active neurons most-correlated with the first one selected are selected until a specified number to turn on is reached. It should be noted that the “off” neurons in the decoder are actually *not task-positive*, and not “default task-negative" neurons as in the brain. *I.e.*, decoder neurons are not used for self- generated thought.

In the following, we provide results and evaluation for the low-active method. We compared those results with the samples generated from (1) the unmodified decoder as well as by (2) Noisy-decoding: during decoding a random Gaussian noise was infused in a fraction of neurons, and (3) Random flipping: activation of randomly selected “off” neurons (same as our proposed method but does not apply special criteria to select “off” neurons). Linear interpolation in the latent space between training samples that belong to different classes, followed by standard decoding did not yield creative samples (result not shown).

**Evaluation of Creativity** The ultimate test of creativity is through human inspection, and human labeling has been used to evaluate deep generative models Dosovitskiy et al. (2016); Lopez & Tucker (2018) or as a part of the generative pipeline Lake et al. (2015); Salimans et al. (2016). We used an in-house annotation tool to evaluate creativity of the generated samples, with four options to choose from per image - ‘not novel or creative’, ‘novel but not creative’, ‘creative’, and ‘inconclusive’. Annotators did not have access to the knowledge of the decoding scheme at the time of annotation, but were primed on the training dataset.

We also used a number of surrogate novelty metrics. One key one we focus on for this paper is reconstruction distance based on encoder-decoder architectures Wang et al. (2018). For image $x$ and corresponding latent (encoded) vector $z$, novelty can be estimated from the distance between $x$ and the closest sample the VAE can produce from $z$. Therefore, $D_r = \min_z ||x - E[_\theta(x|z)]||_2$. Since a trained VAE has a narrow bottleneck, the reconstruction distance of any novel image will be large.

## 4 Experiment Details and Results

We used a VAE as the generative model. For F/MNIST datasets the encoder network consisted of 3 fully-connected layers (1000, 500, 250) before the $z$ output (50 for F/MNIST and 100 for the combination), with the decoder architecture mirroring it. RELU activations and training dropout of $0.10$ were used. We performed modifications at the decoder’s 3rd hidden layer, however, modifying lower layers also produced a variety of creative results (not shown). Results were obtained by perturbing five neurons during decoding. For the low-active method, we used neurons whose activations were within the 1st and 15th percentiles of the neuron percent activations for the layer.

Reconstruction distance is averaged over 10K generated samples. For human evaluation of creativity, 9 annotators annotated a pool of $\approx 500$ samples per dataset (we used agreement amongst 3 or more annotators as consensus), generated from either using neuro-inspired creative decoding (low-active), baseline decoding (noisy decoding and random activation) or regular decoding.

**Human annotation results.** Figures 1D and 2A-C present the analysis of human annotations of the generated samples. Visually, creative samples generated from the latent space of MNIST digits do not appear digit-like, rather look more like symbols. In contrast, those generated from FMNIST still resemble fashion objects; however, novel objects, *e.g.*, a shirt with one sleeve, sock, bag with asymmetric handle, or front-slit kurta were found. Comparison with “novel but not creative” images

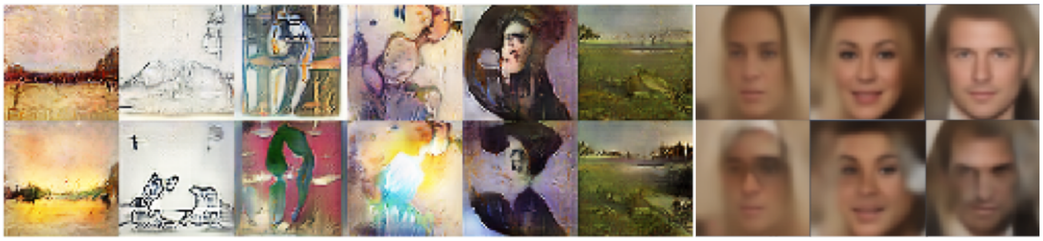

Figure 3: Results on WikiArt (with ArtGAN) and CelebA (using VAE) (top: regular, bottom: modified using low-active decoding)

confirms the known association of both novelty and value with human perception of creativity. Interestingly, VAE trained on the combined dataset outputs creative images distinct from the MNIST- and FMNIST-only cases. As the perception of “creativity” is mostly aesthetical, results on more broad and complex “artsy" datasets is shown in Figure 3; however, we refrain from evaluating them.

Table 1: Human annotation results: L1 (Creative), L2 (Novel but not creative), L3 (Not novel or creative), L4 (Inconclusive). Values are normalized fraction of annotated instances (with a consensus of 3 or more users) within each decoding scheme. We also report average reconstruction distance, $D_r$, of all generated samples by each method. Highest L1 (creative) fraction and $Dr$ are marked in bold.

| | MNIST | | | | | FMNIST | | | | | MNIST+FMNIST | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L1 | L2 | L3 | L4 | ($D_r$) | L1 | L2 | L3 | L4 | ($D_r$) | L1 | L2 | L3 | L4 | ($D_r$) |
| Low-active | **0.39** | 0.50 | 0.11 | 0.00 | **4.87** | **0.27** | 0.52 | 0.14 | 0.07 | **5.49** | **0.26** | 0.54 | 0.17 | 0.04 | **4.91** |
| Noisy | 0.17 | 0.07 | 0.76 | 0.00 | 1.78 | 0.19 | 0.09 | 0.71 | 0.01 | 1.91 | 0.25 | 0.22 | 0.51 | 0.01 | 1.97 |
| Random | 0.24 | 0.46 | 0.30 | 0.00 | 4.21 | 0.17 | 0.62 | 0.16 | 0.06 | 4.93 | 0.19 | 0.61 | 0.15 | 0.05 | 4.68 |
| Regular | 0.10 | 0.03 | 0.85 | 0.01 | 1.45 | 0.13 | 0.08 | 0.76 | 0.03 | 1.22 | 0.23 | 0.17 | 0.55 | 0.05 | 1.59 |

**Comparison with baseline methods.** Normalized fraction of creative samples with low subject variability (Table 1, L1) suggests that the low-active method constantly outperforms baseline methods (random flipping of “off” neurons, noisy decoding) and regular decoding in single dataset scenarios; the relative gain clearly depends on the dataset. Noisy decoding performs similar to regular decoding (*i.e.*, generates samples similar to training data), while random activation of “off” neurons has a higher tendency to produce novel (but not creative) samples. Therefore, the special effect of “creative” decoding cannot be replicated by simply flipping random “off” neurons (Table 1) - the neuro-inspired selection and flipping of low-active neurons is what promotes creativity in generations. Training on combined dataset enables creative generation by using baseline and regular methods as well, likely due to the extended capacity of the VAE itself (interpolating between unrelated object types). An interesting observation emerges from the average reconstruction distance ($D_r$): the low-active method consistently yields samples with higher $D_r$ on average for all datasets, demonstrating it’s enhanced ability of generating out-of-distribution samples.

## 5 Discussion and Future work

Prior work has shown that high decoder capacity enables easier posterior inference; at the same time the model becomes prone to over-fitting Kingma et al. (2016). The presence of inactive latent units in a trained VAE decoder originates from regularization. Those low-active neurons are sparsely activated, not important for reconstruction / classification, and often encode unique sample-specific features - so are likely not part of the winning ticket Frankle & Carbin (2018); effects of pruning them will be investigated in future. While our intent is not to downplay the complexity of the human brain and the creative cognition process, the fact that exploiting the extra unused capacity of a trained decoder in a brain-inspired manner provides access to novel and creative images is interesting. Future work will include testing creativity of trainable decoders with purposely added extra capacity and exposing the model to a more complicated multi-task setting. Additionally, memory retrieval in a deep neural net by disentangling “low-active” neurons and then activating them at test time will be investigated. Surrogate metric design for better capturing human perception of creativity will also be investigated in the future toward empowering machine learning models with “creative autonomy” Jennings (2010).

## References

Alan Anticevic, Michael W Cole, John D Murray, Philip R Corlett, Xiao-Jing Wang, and John H Krystal. The role of default network deactivation in cognition and disease. *Trends in cognitive sciences*, 16(12):584–592, 2012.

Roger E Beaty, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative cognition and brain network dynamics. *Trends in cognitive sciences*, 20(2):87–95, 2016.

Roger E Beaty, Yoed N Kenett, Alexander P Christensen, Monica D Rosenberg, Mathias Benedek, Qunlin Chen, Andreas Fink, Jiang Qiu, Thomas R Kwapil, Michael J Kane, et al. Robust prediction of individual creative ability from brain functional connectivity. *Proceedings of the National Academy of Sciences*, pp. 1087–1092, 2018.

Margaret A Boden. *The creative mind: Myths and mechanisms*. Routledge, 2004.

Akın Kazakçıand Mehdi Cherti and Balázs Kégl. Digits that are not: Generating new types through deep neural nets. *arXiv preprint arXiv:1606.04345*, 2016.

Louis Clouâtre and Marc Demers. Figr: Few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019.

Steve DiPaola and Liane Gabora. Incorporating characteristics of human creativity into an evolutionary art algorithm. *Genetic Programming and Evolvable Machines*, 10(2):97–110, 2009.

Alexey Dosovitskiy, Jost Tobias Springenberg, Maxim Tatarchenko, and Thomas Brox. Learning to generate chairs, tables and cars with convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):692–705, 2016.

Ahmed Elgammal, Bingchen Liu, Mohamed Elhoseiny, and Marian Mazzone. Can: Creative adversarial networks, generating" art" by learning about styles and deviating from style norms. *arXiv preprint arXiv:1706.07068*, 2017.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

Zhenni Gao, Delong Zhang, Aiying Liang, Bishan Liang, Zengjian Wang, Yuxuan Cai, Junchao Li, Mengxia Gao, Xiaojin Liu, Song Chang, et al. Exploring the associations between intrinsic brain connectivity and creative ability using functional connectivity strength and connectome analysis. *Brain connectivity*, 7(9):590–601, 2017.

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

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Jeanine Graf and Wolfgang Banzhaf. Interactive evolution of images. In *Evolutionary Programming*, pp. 53–65, 1995.

Kyle E Jennings. Developing creativity: Artificial barriers in artificial intelligence. *Minds and Machines*, 20(4):489–501, 2010.

Balázs Kégl, Mehdi Cherti, and Akın Kazakçı. Spurious samples in deep generative models: bug or feature? *arXiv preprint arXiv:1810.01876*, 2018.

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

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in neural information processing systems*, pp. 4743–4751, 2016.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

CS Lopez and CE Tucker. Human validation of computer vs human generated design sketches. *ASME Paper No. DETC2018-85698*, 2018.

Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Innovation engines: Automated creativity and improved stochastic optimization via deep learning. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, pp. 959–966. ACM, 2015.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. *arXiv preprint arXiv:1603.05106*, 2016.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pp. 2234–2242, 2016.

Liang Shi, Jiangzhou Sun, Yunman Xia, Zhiting Ren, Qunlin Chen, Dongtao Wei, Wenjing Yang, and Jiang Qiu. Large-scale brain network connectivity underlying creativity in resting-state and task fmri: cooperation between default network and frontal-parietal network. *Biological psychology*, 135:102–111, 2018.

R Nathan Spreng, Kathy D Gerlach, Gary R Turner, and Daniel L Schacter. Autobiographical planning and the brain: activation and its modulation by qualitative features. *Journal of cognitive neuroscience*, 27(11):2147–2157, 2015.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Huan-gang Wang, Xin Li, and Tao Zhang. Generative adversarial network based novelty detection using minimized reconstruction error. *Frontiers of Information Technology & Electronic Engineering*, 19(1):116–125, 2018.