Toward A Neuro-inspired Creative Decoder

Payel Das* , **Brian Quanz*** , **Pin-Yu Chen** , **Jae-wook Ahn** and **Dhruv Shah**
IBM Research, Yorktown Heights, NY, USA
{daspa,blquanz}@us.ibm.com, pin-yu.chen@ibm.com, jaewook.ahn@us.ibm.com, dhruv.shah@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, in this study 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 can be used off-the-shelf. Several novelty metrics and human evaluation were used to evaluate the creative capacity of the deep decoder. Our experiments on different image datasets (MNIST, FMNIST, MNIST+FMNIST, WikiArt and CelebA) reveal that atypical co-activation of highly activated 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 and Banzhaf, 1995] or evolutionary algorithms with hand-crafted fitness functions [DiPaola and 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 to employ deep learning-based methods in computational creativity research. Only recently, novelty generation using powerful deep generative models, such as Variational Autoencoders (VAEs)[Kingma and Welling, 2013; Rezende and Mohamed, 2015] and Generative Adversarial Networks (GANs) [Goodfellow *et al.*, 2014], have been attempted. These are designed to model the distribution of and generate known objects (*i.e.*, images). However, such models discourage out-of-distribution generation to avoid instability and minimize spurious sample generation, limiting their potential in creativity research. In fact, [Kégl *et al.*, 2018] shows that getting rid of "spurious" samples completely can limit the generative modeling capacity. 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. A number of recent neuroimaging studies [Beaty *et al.*, 2018; Shi *et al.*, 2018; Gao *et al.*, 2017] indicate stronger coupling of 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 considered as the large-scale communities of interacting brain regions, revealed by the resting-state functional correlation pattern; these networks seem to correspond to distinct functional systems of the brain. Default (task-negative) mode network is associated with spontaneous and self-generated thought, and, therefore implicated in idea generation. Control (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.

Research Question. 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.

It is widely accepted that creativity is a combination of nov-

*Payel Das and Brian Quanz contributed equally to this work and are contact authors

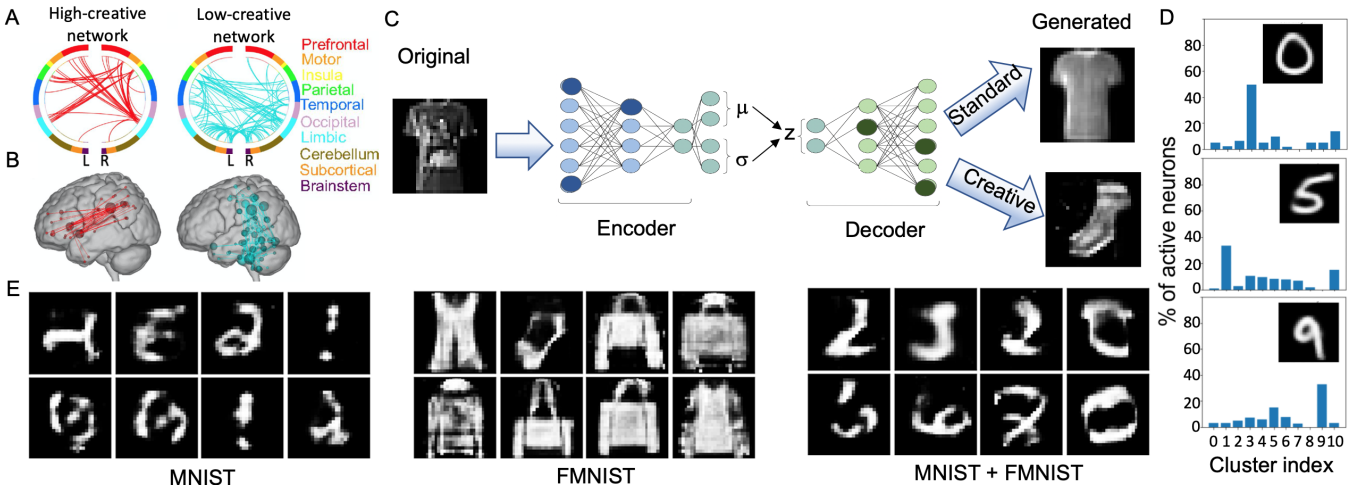

Figure 1: **A-B.** Depictions (A: circle plots, B: glass brains) of high- and low-creative networks in human brains with their highest degree nodes. Circle plot colors correspond to brain lobes: L, left hemisphere; R, right hemisphere. Adapted from [Beaty *et al.*, 2018]. **C.** Depiction of a VAE model with our neuro-inspired *creative* decoder. Normally, a small fraction of neurons in each hidden layer are low-active (dark color). Inspired by neural basis of creativity, we activate those "low-active" (task-negative) neurons to induce coupling between task-positive and task-negative neurons during "creative" decoding. **D.** Class activation of neuronal clusters in an MNIST VAE, used by the cluster-level activation version of our method that models the neuronal networks with activation clusters and works by co-activating "low-active" neuronal activation cluster(s) (representing the task-negative network), suggesting cluster activation patterns are associated with high-level concepts (e.g., digit classes / components). **E.** Samples generated by the proposed *creative* decoder that were human-annotated as creative with high confidence.

elty and value. However, value determination is non-trivial and design of clear-cut creativity evaluation schemes is believed to be as important as developing *creative* generative methods [Cherti *et al.*,]. As the goal is to generate meaningful novelty, not trivial noise, traditional metrics based on likelihood/distance are of limited use. Nevertheless, novelty alone has been found to be a better predictor of creativity than value [Diedrich *et al.*, 2015], and consistently, stringent evaluation is found to hinder creativity in the brain [Beaty *et al.*, 2017].

In the present study, we employ human annotation for creativity evaluation in addition to using a number of surrogate metrics (supervised and unsupervised) for novelty estimation of the samples generated by the *creative* decoder. A VAE model was used as the base generative framework (see Fig 1C). 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.

Our main contributions are:

• Inspired by neuroimaging findings, we propose a VAE model with a *creative* decoder, which generates novel and meaningful samples from the learned representation space by employing an atypical activation pattern during decoding. This modified generation scheme *does not require any data labels* and can be adapted to any decoder/generator.

• Different schemes (*correlation-based*, *cluster-based*, and *low-active*) of stochastic co-activation of "on" and "off" neurons during decoding are introduced and tested in terms of creative capacity and compared against a number of unsupervised baseline decoding methods. Results show that enhanced creativity can result from the neuro-inspired atypical activation as opposed to a simple random or structured noise effect.

• The creativity of generated samples was evaluated using human annotation as well as several surrogate metrics for novelty detection (*e.g.*, in-domain classifier score and reconstruction distance). Our analyses suggest that a combination of existing surrogate metrics largely captures the notion of creativity.

• A key advantage of our method is its immediate applicability to any off-the-shelf decoder/generator model, as it requires no retraining, access to abnormal samples, or additional categorical information for creative generation. To our knowledge, this is the first work that aims to enhance creative capacity of a deep generative model in a neuro-inspired manner.

2 Related Work

Deep models for novelty generation: [Nguyen *et al.*, 2015] proposed the deep neural network based *innovation engine* algorithm, demonstrating that evolving toward many objectives simultaneously approximates divergent search and encourages novel generation. It is composed of: (1) a diversity-promoting evolutionary algorithm (EA) that generates novel instances based on pre-set features, and (2) a deep classifier to determine if generated instances are interesting and should be retained. The main differences between *innovation engine* and *creative* decoder are use of a data-driven generative model instead of EA and inducetion of novelty without a set of pre-set features.

Another line of work is few/one-shot learning and generation [Lake *et al.*, 2015; Rezende *et al.*, 2016; Clouâtre and Demers, 2019], a step toward out-of-class generation. [Lake *et al.*, 2015] trains a model with Bayesian Program Learning, which represents concepts (pen strokes) as simple probabilistic programs and hierarchically combines them to generate images (letters). They showed that the model can be trained on a single image of a previously unseen class and generate novel samples of it. [Rezende *et al.*, 2016] developed a class

of sequential generative models to achieve one-shot generation. They showed that a combination of sequential generation and inference in a VAE model and attention [Bahdanau *et al.*, 2014] during inference enables generation of compelling alternative variations of images after seeing them once. Recently, [Clouâtre and Demers, 2019] combined GANs with meta-learning [Nichol *et al.*, 2018] to find GAN parameters that can generate a random instance with little data and training. The difference with this line of work is that ours does not need any novel exemplar during test time for novel generation.

Next, the creative adversarial network (CAN), a GAN variant, attempts to generate novel artistic images by minimizing deviation from an art distribution while maximizing style ambiguity [Elgammal *et al.*, 2017]. CAN is based on a psychology concept that exploits style ambiguity as a means to increase arousal potential. Unlike our proposed model, CAN needs data-label pairs as input to generate novel samples.

[Cherti and Kégl, 2016] generated new images that are not digit-like by performing operations like crossover and mutation, in the latent space of MNIST digits learned using an autoencoder, or by iteratively refining a random input such that it can be easily reconstructed by the trained autoencoder. Their method needs human feedback to make the network creative-explorative, which is different from our method, as *creative* decoder exploits a reported neural basis of creativity.

3 Evaluation of Creativity

Human Evaluation of Creativity. The ultimate test of creativity is through inspection by humans. Additionally, human labeling has been used to evaluate deep generative models [Dosovitskiy *et al.*, 2016; Lopez and Tucker, 2018] or as a part of the generative pipeline [Lake *et al.*, 2015; Salimans *et al.*,]. Although human judgment of creativity suffers from several drawbacks (it is costly, cumbersome, and is subjective), it is still crucial to check how humans perceive and judge generated images. We used an in-house annotation tool to evaluate creativity of the generated samples. Given an image, the tool had four options to choose from - 'not novel or creative (similar to training data)', 'novel but not creative (different from training data but does not seem meaningful or useful)', 'creative (different from training data and is meaningful or useful)', 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, and the annotation category specifics.

Surrogate Metrics of Novelty. Novelty detection techniques can be broadly categorized in five categories: (i) probabilistic, (ii) distance-based, (iii) reconstruction-based, (iv) domain-based, and (v) information-theoretic techniques. Below, we outline the novelty metric families used for evaluation.

Reconstruction distance. Reconstruction distance based on encoder-decoder architectures has been leveraged for novelty detection [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.

k-Novelty score. We compute the kth-nearest-neighbor distance between a generated sample and the training dataset in the latent z space and in the input space.The kNN novelty score (kNS) [Ding *et al.*, 2014] of a given data point is the normalized distance to its kth nearest neighbor (denoted as $kNN(\cdot)$), which is defined as following: $kNS = \frac{d(x,kNN(x))}{d(kNN(x),kNN(kNN(x)))}$. For $k = 1$, d is an l_2 distance. For $k > 1$, d accounts for the expected distance to the kth nearest neighbor. We compute kNS for k=1 and 5.

In-domain Classifier Entropy (ICE) or "objectness". Entropy of the probabilities p returned by a trained, multi-class, in-domain classifier has been used to estimate novelty [Hendrycks and Gimpel, 2016; Kliger and Fleishman, 2018; Cherti *et al.*,], which can be defined as $ICE = -\sum_i p_i \log_2 p_i$. A higher value of ICE implies higher novelty. This metric is similar to what [Salimans *et al.*,] has used to stabilize GAN training and [Nguyen *et al.*, 2015] has employed to encourage novelty search.

In-domain Score (IS) from a one-class classifier. One-class classifiers have also been used for novelty detection, as novel classes are often absent during training, poorly sampled or not well defined. In this study, for each dataset we train a one-class support vector machine (SVM) classifier on the latent dimensions obtained from trained VAE model. Using this one-class classifier, We classify each generated image to get an in-domain score (IS) - the signed difference from the normal-classifying hyperplane. The lower this value is, the stronger the outlierness of the image.

4 Methodologies and Preliminaries

Variational Autoencoder (VAE). We use a VAE [Kingma and Welling, 2013] as the base generative model, which trains a model $p(x, z)$ to maximize the marginal likelihood $\log p(x)$ on dataset samples x. As the marginal likelihood requires computing an intractable integral over the unobserved latent variable z, VAEs introduce an encoder network $q_\theta(z|x)$ and optimize a tractable lower bound (the ELBO): $\log p(x) \geq E[\log p(x|z)] - D_{KL}[q(z)||p(z)]$. The first term accounts for the reconstruction loss, the second measures KL loss between the encoder's latent code distribution, $p_\theta(z)$, and the prior distribution, typically a diagonal-covariance Gaussian.

***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 z, the proposed method selects decoder neurons based on their activation correlation pattern, and then activates a few "off" neurons along with the "on" neurons. Three different co-activation schemes were tested, each selecting a set of "off" neurons to turn "on" using different grouping criteria to represent "task-negative": (1) correlation-based – random selection from the pool of "off" neurons that are most anti-correlated with the firing neurons across the training data; (2) cluster-based (Fig. 1D) – "off" neurons were chosen at random from the neuronal clusters (obtained by clustering the neuronal activation map) with low percent of cluster active during decoding; and (3)

"low-active" based – random selection from "off" neurons that were weakly-activated in the trained decoder across most of the training data. Another version of the "low-active" method only selected low-active neurons without any significant class specificity (non-specific low-active method). Note: the "off" neurons are fundamentally different from task-negative brain ones due to implication in self-generated thoughts.

In the following, due to the space constraint, we only provide a complete description and evaluation of the low-active method; other variants (correlation, cluster, and non-specific low-active method) were found to perform in a similar fashion.

Preliminaries. Let $d_j^k(z)$ represent the output of the jth neuron of the kth layer, given input $z \in \mathcal{Z}$ to the decoder p_θ. We further use the short-hand, $d_{ji}^k = d_j^k(z_i)$, where z_i is the encoding of the ith training data point. Then the *percentage activation* of neuron j in layer k is $a_j^k = (1/n)\sum_{i=1}^n \mathbb{1}_{d_{ji}^k > \tau}$, for a given activation threshold τ. For RELU activation functions we set τ to $1\mathrm{e}{-7}$. Neurons with $a_j^k \approx 0$ are classified as *dead* neurons and excluded from neuronal manipulation. Additionally, for the ith input we call neuron j of layer k "active" or "on" if $d_{ji}^k > \tau$, and "inactive" or "off" otherwise. Let $\vec{d_j^k}$ be the vector with ith entry $= d_{ji}^k$. Given neuron j and neuron h in layer k, let $C^{kjh} = \mathrm{Cov}[\vec{d_j^k}, \vec{d_h^k}]$ (the covariance matrix). Then we define their correlation $R_{jh}^k = C_{01}^{kjh}/\sqrt{C_{00}^{kjh} C_{11}^{kjh}}$ - which are the entries of the layer correlation matrix R^k.

Neuron flipping. We define flipping a neuron "off" by setting it to a minimum activation value, *i.e.*, 0 for RELU activation. We define the "on" value of a neuron j in layer k as $o_j^k = \lambda \cdot s(\{d_{ji}^k | i = 1 \ldots n\})$, *i.e.*, λ denotes a scaling factor for the statistic of training activation values, *e.g.*, s equal to mean, max, *etc*. Since, task-negative neurons are activated at higher levels than usual during creative processes in the human brain, we set $s = \max$ and $\lambda = 2$ in experiments.

We start with a sample $z \sim p(z)$ and obtain the neuronal activations for a selected layer k of the decoder, for which we now use shorthand $d_{jz}^k = d_j^k(z)$. Additionally let $\mathcal{A}$ be the corresponding set of active or "on" neurons and $\mathcal{D}$ the set of inactive or "off" (non-dead) neurons. During creative decoding, we flip some number or percentage, ρ, of a group of "on" and/or "off" neurons in a layer k, either randomly or selectively. Each method modifies $d_j^k(z)$, and this modified layer output is then passed through the remainder of the decoder, to obtain the final generated values.

Correlation method The correlation method (1) randomly selects an "off" neuron from the group in $\mathcal{D}$ least-correlated with the most-activated group of neurons; (2) selects additional neurons correlated with the selected deactivated neuron; and (3) turns them on. The idea is that these deactivated neurons are not correlated with the most active ones, so can be viewed as instance-specific task-negative neurons. Using correlation, concepts encoded in multiple neurons might be better captured than pure random selection.

Clustering method The clustering method is like the correlation method, except instead of considering individual neurons, clusters of neurons are considered, which represent instance-specific task-negative sub-networks. Spectral clustering is applied to the layer output correlation matrices from the training data (R^k) to get cluster memberships. For a given instance and decoder layer, one or more clusters with lowest percent activation are randomly selected, where percent activation for cluster $\mathcal{C}^k$ is $(1/|\mathcal{C}^k|)\sum_{j \in \mathcal{C}^k} \mathbb{1}_{d_{jz}^k > \tau}$. The percent activations of selected clusters are then increased by turning on more neurons in those clusters until the specified number to turn on is reached.

Algorithm 1 Low-active method

Input: Layer output d_{jz}^k; percent activation percentile κ; fraction of neurons to flip on ρ

$t \leftarrow \kappa$-percentile($\{a_j^k\}$)
$\mathcal{S} \leftarrow \{j | j \in \mathcal{D} \wedge a_j^k \leq t\}$
$s \leftarrow$ random select from $\mathcal{S}$
$\mathcal{S}_s \leftarrow \mathrm{argmax}_{\mathcal{S}' \subset \mathcal{S}, |\mathcal{S}'| = \lfloor \rho |\mathcal{S}| \rfloor} \sum_{h \in \mathcal{S}'} R_{sh}^k$
$d_{jz}^k \leftarrow o_j^k, \forall j \in \mathcal{S}_s$

"Low-active" method. The "low-active" method (Algorithm 1) imitates the task-negative concept by identifying neurons that typically have low activation across all the training data and turning some number of them on at decode time. Specifically, a neuron is selected from the pool of "off" neurons that have the lowest percent activations, a_j^k (defined using a threshold cutoff). Next, "low-active" neurons that are most correlated with the selected neuron are also turned on (randomly selecting the remainder from the low-active group provides similar results).

The **non-specific "low-active" method** variant uses two criteria for selecting neurons: (1) the max percent activation across all training data in any given class is below a threshold and (2) the entropy of percent activation across classes is above a threshold - *i.e.*, non-specific, as this scheme mimics the concept of the task negative default network in the brain. We use training data class labels for this purpose - in general if unavailable, surrogate class labels can be derived, e.g., using clustering with cluster membership as classes.

Baseline decoding schemes. The samples generated by the *creative* decoding schemes were compared with the original training distributions as well as with samples generated by (1) Linear Interpolation in the latent space between training samples that belong to different classes, followed by standard decoding (result not shown), (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 methods but does not apply special criteria to select "off" neurons).

5 Experimental Details

VAE architecture. For F/MNIST 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 the reverse of the encoder. RELU activations were used; dropout equal to 0.10 for fully-connected layers was used during training only.

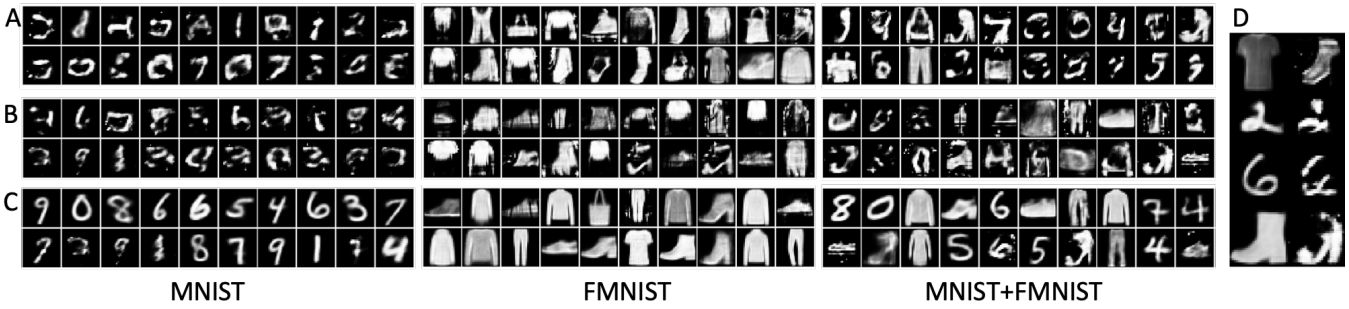

Figure 2: Random VAE-generated samples human-annotated as (A) creative, (B) novel but not-creative, and (C) not novel or creative. D: How "creative" decoding modifies generation (combined dataset): **left:** regular; **right:** low-active decoding.

We performed modifications at the decoder's 3rd hidden layer, since some transformation from the z-space should be necessary to capture underlying data invariance/structure that is then decoded to the final image. However, modifying lower layers also produced a variety of creative results - analysis of decoder layer choice impact is future work. Unless otherwise stated, results in the main paper were obtained by perturbing five neurons during decoding. For the low-active method, we used neurons whose activations (see Method Section) were within the 1st and 15th percentiles of the neuron percent activations (a_j^k) for the layer.

Sample evaluation. All novelty metrics reported are averaged over 10K generated samples per method. The trained in-domain classifiers yielded test accuracy of 99.28% for MNIST and 92.99% for FMNIST. For human evaluation, 9 evaluators annotated a pool of $\approx$ 500 samples per dataset (we used agreement amongst $>$3 annotators as consensus) [Sbai *et al.*, 2018; Elgammal *et al.*, 2017], generated from either using neuro-inspired creative decoding (low-active), baseline decoding (noisy decoding and random activation) or regular decoding.

6 Results

Human annotation. Figures 1E and 2A present the analysis of human annotations of the generated samples by 9 annotators. For comparison, samples that were annotated as "novel but not creative" and "not novel or creative" are also shown (Fig. 2B-C). Visually, creative samples generated from the latent space of MNIST digits do not appear digit-like anymore, but look more like symbols. In contrast, those generated from FMNIST still resemble fashion objects; however, novel objects such as a shirt with one sleeve, sock, bag with asymmetric handle, or front-slit kurta were found. Comparison with "novel but not creative" images confirms the known association of both novelty and value with the human perception of creativity. Interestingly, the combined dataset VAE outputs creative images that are distinct from the MNIST- and FMNIST-only cases.The present perception of "creativity" is mostly aesthetical, so experiments on more complex "artsy" datasets are needed; we show some results in Figure 3.

Comparison with baseline methods. Normalized fraction of creative samples with low subject variability (Table 1, L1 columns) suggests that the low-active method constantly outperforms the baseline methods (random flipping of "off" neurons, noisy decoding) and regular decoding in a single dataset scenario; 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 generation of creative samples 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-act method consistently yields samples with higher reconstruction distance on average for all datasets, demonstrating it's enhanced ability of generating out-of-distribution samples. However, those generated samples may not always be perceived as creative by human , *e.g.* in the combined data scenario, as the meaningfulness might disappear. Additionally, we ran categorical variable significance tests (chi-squared and G-tests) between our methods and baselines. Low-act method was found to be significantly different from all baselines (p<0.05) for all datasets and tests, except for the random baseline in the combined dataset scenario.

Relation between human judgment and novelty metrics. Next, we report the values of novelty metrics and their relation with the human judgment of creativity in Table 2. The std deviations (std) (not shown) are small for Dr and $kNSz$ (*e.g.*, std is 20% of the average Dr), while ICE and IS show slightly higher variability. Nevertheless, the overall trend of ML metrics between creative and not-creative images remains consistent across all datasets. Table 2 reveals that a combination of low in-domain score, and high D_r and kNS_z with respect to the L3 (not creative or novel) samples are key characteristics of the creative images, on average. The same holds for novel (L2) generations; however, novel samples lie somewhere in between regular (L3) and creative (L1). These results are consistent with earlier findings [Cherti *et al.*,], demonstrating that out-of-class metrics capture the creative capacity of generative models well.

We further trained a "creativity" classifier using surrogate metrics (Table 2) as features, and found a combination of metrics provides superior predictability over any single metric.

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

	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

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				
	NS_z	D_r	ICE_M	IS	NS_z	D_r	ICE_F	IS	NS_z	D_r	ICE_M	ICE_F	IS
L1	1.305	5.361	0.215	419.507	1.378	5.650	0.126	341.557	1.429	4.716	0.251	0.126	279.086
L2	1.291	5.486	0.265	434.823	1.367	5.751	0.218	488.675	1.516	5.207	0.293	0.199	323.443
L3	1.100	3.084	0.100	492.340	1.224	3.667	0.137	434.923	1.263	3.284	0.273	0.169	415.610
L4	1.157	4.171	0.650	691.493	1.309	5.197	0.289	524.912	1.453	5.453	0.463	0.228	541.208

Table 2: Comparison between human judgment and novelty metrics: Novelty score $NS(z)$ (considering top 5 nearest neighbors), reconstruction distance D_r, in-domain MNIST classifier entropy ICE_M, in-domain FMNIST classifier entropy ICE_F, in-domain score (IS) obtained using one-class SVM classifier. L1 (Creative), L2 (Novel but not creative), L3 (Not novel or creative), L4 (Inconclusive).

For example, a trained L1-regularized logistic regression classifier for predicting consensus "creative" or "not creative" on FMNIST evaluations yields a 10-fold CV mean accuracy of 71.0% whereas the best single-metric accuracy was 60.7%.

We also analyzed metric results for other generation methods (e.g., our cluster- and correlation-based creative decoding, and latent-space linear interpolation), for 1 to 25 neurons turned on (results for one metric shown in Figure 4). This revealed a general increase of novelty scores as flipped neurons increases, with proposed neuro-inspired methods generally dominating. In accordance with human evaluation results, the number of flipped neurons required to induce changes in surrogate metrics that are indicative of creativity was minimal for neuro-inspired low-active method compared to other methods. Activating multiple “off" neuronal clusters together was found effective as well, emphasizing need for atypical neuronal coordination for novelty generation. Linear interpolation in the z space produced the opposite trend in terms of surrogate metrics (in the context of creativity) and visual inspection confirmed resulting samples were not novel.

We also performed additional experiments by turning off active neurons (not shown). To yield any significant effect required turning a larger number of neurons off. Typically, results quickly become less coherent or meaningful (e.g. become blurry), as we turn more active neurons off. This is likely due to a combination of redundancy and task- focused nature of highly active neurons.

Applications to more complex datasets. Figure 3 shows random generations of low-active decoding applied to ArtGAN trained on WikiArt dataset and a larger VAE trained on CelebA, which demonstrate generality and effectiveness of the proposed creative decoding approach. The method works as well, if not better, on bigger networks for larger, more complex, colored images and also with GANs. Full analysis and human evaluation of these results are left to future work. Figure 3 indicates that the proposed creative decoding causes changes in color, style, texture, and even shape - sometimes seeming to add components that fit into the scene. *E.g.*, in Figure 3 the first WikiArt example image seems to have an added sunrise effect, in the fourth the middle portion is converted into a waterfall, and in the fifth a circular shape is changed to an arm and Elvis-style hair. For CelebA, we see sunglasses being added, shape structure/ethnicity changed, etc.

Comparison with recent creative generation methods. To our knowledge, the proposed method is the first approach for generating creative modifications that is fully unsupervised, model-agnostic (can be applied to any trained generator network) and does not require any special model training. Earlier reports [Cherti and Kégl, 2016; Kégl *et al.*, 2018]

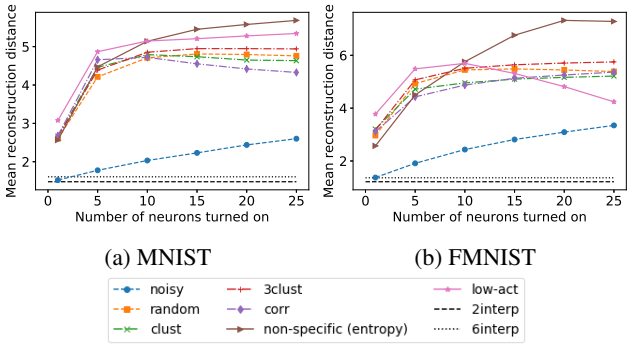

Figure 4: Mean reconstruction distance per approach, for increasing number of neurons turned on (1 to 25). "low-act" is the proposed low-active method. "clust" and "3clust" are the proposed clustering method selecting 1 and 3 clusters, respectively, out of 11 total clusters. "corr" is the proposed correlation method, and "non-specific (entropy)" is the non-specific variant of the low-active method. "2interp" and "6interp" are interpolation between random latent space points in different classes, using 2 and 6 points, respectively. Turning on more neurons generally increases novelty metrics up to some saturation point; when low-active neurons run out, more regular concept neurons can be turned on leading to reduced novelty with too many neurons.

used either specific training or supervised learning to generate novel images. Since the code or model from those studies is not publicly available to our knowledge, it is not possible to compare our results with those directly, although visual inspection reveals high similarity with some of the "symbols" generated from MNIST in [Cherti and Kégl, 2016; Kégl *et al.*, 2018]. We also estimated Inception Score [Salimans *et al.*,] using the MNIST classifier and found that the creatively decoded samples have lower scores (within 7-8 range) than test samples (9.9) and significantly higher scores than random samples (2.9), consistent with [Kégl *et al.*, 2018].

7 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]. In fact, high capacity decoders in a VAE-setting are known to ignore latent information while generating, which has been widely addressed [Kingma *et al.*, 2016; Rezende and Mohamed, 2015].

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 and Carbin, 2018]; effects of pruning them will be investigated in the 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

[Anticevic *et al.*, 2012] 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.

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Beaty *et al.*, 2016] 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.

[Beaty *et al.*, 2017] Roger E Beaty, Alexander P Christensen, Mathias Benedek, Paul J Silvia, and Daniel L Schacter. Creative constraints: Brain activity and network dynamics underlying semantic interference during idea production. *Neuroimage*, 148:189–196, 2017.

[Beaty *et al.*, 2018] 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*, pages 1087–1092, 2018.

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

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

[Cherti *et al.*,] Mehdi Cherti, Balázs Kégl, and Akin Kazakçı. Out-of-class novelty generation: an experimental foundation. In *Tools with Artificial Intelligence (ICTAI), 2017 IEEE 29th International Conference on*.

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

[Diedrich *et al.*, 2015] Jennifer Diedrich, Mathias Benedek, Emanuel Jauk, and Aljoscha C Neubauer. Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9(1):35, 2015.

[Ding *et al.*, 2014] Xuemei Ding, Yuhua Li, Ammar Belatreche, and Liam P Maguire. An experimental evaluation of novelty detection methods. *Neurocomputing*, 135:313–327, 2014.

[DiPaola and Gabora, 2009] 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.

[Dosovitskiy *et al.*, 2016] 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.

[Elgammal *et al.*, 2017] 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.

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

[Gao *et al.*, 2017] 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.

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

[Goodfellow *et al.*, 2014] 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*, pages 2672–2680, 2014.

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

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

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

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

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

[Kingma *et al.*, 2016] 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*, pages 4743–4751, 2016.

[Kliger and Fleishman, 2018] Mark Kliger and Shachar Fleishman. Novelty detection with gan. *arXiv preprint arXiv:1802.10560*, 2018.

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

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

[Nguyen *et al.*, 2015] 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*, pages 959–966. ACM, 2015.

[Nichol *et al.*, 2018] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

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

[Rezende *et al.*, 2016] 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.

[Salimans *et al.*,] 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*.

[Sbai *et al.*, 2018] Othman Sbai, Mohamed Elhoseiny, Antoine Bordes, Yann LeCun, and Camille Couprie. Design: Design inspiration from generative networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[Shi *et al.*, 2018] 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.

[Spreng *et al.*, 2015] 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.

[Szegedy *et al.*, 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*, pages 1–9, 2015.

[Wang *et al.*, 2018] 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.