Towards Creativity Characterization of Generative Models via Group-based Subset Scanning

Celia Cintas1*, Payel Das2, Brian Quanz2, Girmaw Abebe Tadesse1, Skyler Speakman1 and Pin-Yu Chen2
1IBM Research Africa, Nairobi, Kenya
2IBM Research, Yorktown Heights, NY, USA
celia.cintas@ibm.com, {daspa,blquanz}@us.ibm.com, girmaw.abebe.tadesse@ibm.com, skyler@ke.ibm.com, pin-yu.chen@ibm.com

Abstract

Deep generative models, such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs), have been employed widely in computational creativity research. However, such models discourage out-of-distribution generation to avoid spurious sample generation, thereby limiting their creativity. Thus, incorporating research on human creativity into generative deep learning techniques presents an opportunity to make their outputs more compelling and human-like. As we see the emergence of generative models directed toward creativity research, a need for machine learning-based surrogate metrics to characterize creative output from these models is imperative. We propose group-based subset scanning to identify, quantify, and characterize creative processes by detecting a subset of anomalous node-activations in the hidden layers of the generative models. Our experiments on the standard image benchmarks, and their “creatively generated” variants, reveal that the proposed subset scores distribution is more useful for detecting novelty in creative processes in the activation space rather than the pixel space. Further, we found that creative samples generate larger subsets of anomalies than normal or non-creative samples across datasets. The node activations highlighted during the creative decoding process are different from those responsible for the normal sample generation. Lastly, we assess if the images from the subsets selected by our method were also found creative by human evaluators, presenting a link between creativity perception in humans and node activations within deep neural nets.

Novelty generation using powerful deep generative models, such as Variational Autoencoders (VAEs) [Kingma et al., 2016; Rezende and Mohamed, 2015] and Generative Adversarial Networks (GANs) [Goodfellow et al., 2014], have been attempted. However, such models discourage out-of-distribution generation to avoid instability and decrease spurious sample generation, limiting their creative generation potential. One of the most controversial issues is the measurement of creativity [Said-Metwaly et al., 2017]. Notably, in the machine learning domain, designing creativity evaluation schemes for creative processes and outputs is as essential as developing creative generative methods. Multiple aspects of creativity, such as expected type (transformational or combinatorial), criteria to assess both the process and the generated output (novelty, value, etc.) [Franceschelli and Musolesi, 2021] and quantitative metrics need to be better defined to allow the research community to develop and test hypotheses systematically [Cherti et al., 2017].

This paper proposes a method designed to detect and characterize when a generative deep neural net model produces a creative artifact as per a human evaluator. We employ group-based scanning to determine whether a given batch of generated processes contains creative samples by using an anomalous pattern detection method called group-based subset scanning [McFowland III et al., 2013].

In short, this work identifies which, of the exponentially-many, subset of samples has higher-than-expected activations at which, of the subset of nodes in a hidden layer of a generative model. For that purpose, we present experiments with a “Creative” variant [Das et al., 2020] of a VAE decoder and of an ArtGAN generator [Tan et al., 2019]. Furthermore, we provide a quantitative, as well as a qualitative, assessment under the WikiArt dataset and the creative samples guided selection process to evaluate the possibility of using group-based scanning as a proxy for the novelty component in creativity perception.

The main contributions of this paper are:

- We propose an approach to detect and characterize the creative processes and artifacts, within the framework of off-the-shelf generative models, by applying group-based subset scanning methods on node activations from internal layers.
- We present the unique ability to identify patterns of

1 Introduction

Creativity is a process that provides novel and meaningful ideas [Boden, 2004; Said-Metwaly et al., 2017]. Current deep learning approaches open a new direction enabling the study of creativity from a knowledge acquisition perspective.
anomalous activations across a group of artifacts.

- We evaluate the selected subgroups of artifacts by our proposed method with quantitative surrogate and detection power metrics. Further, we conduct a validation study with a group of human evaluators to qualitatively assess the “creative” quality of the selected samples with respect to human perception.

2 Related Work

There are multiple surrogate metrics for novelty [Wang et al., 2018; Ding et al., 2014; Kliger and Fleishman, 2018] in the literature that commonly used as a proxy for creativity; however, the ultimate test of creativity is done by human inspection. Human labelling has been used to evaluate deep generative models [Lopez and Tucker, 2018] or as a part of the generative pipeline [Lake et al., 2015; Salimans et al., 2016]. Although human judgment of creativity has numerous drawbacks, such as annotation is not feasible for large datasets due to its labor-intensive nature, operator fatigue, and intra/inter-observer variations related to subjectivity, it is still crucial to check how humans perceive and judge generated artifacts [Lamb et al., 2018].

Among the surrogate metrics of novelty, for example, In-domain Scores (IS) from a classic novelty detector, such as a One-class SVM or Isolation Forests [Liu et al., 2008], have been a popular choice. Commonly, novel classes are often absent during training, poorly sampled, or not well defined. We can classify each generated image to get an in-domain score (IS)- the signed difference from the normal-classifying hyperplane in the case of One-class SVM or the mean anomaly score of the trees in the forest. The lower this value is, the stronger the outlierness of the image. Similarly, reconstruction distance of any novel image using trained autoencoder models will be larger compared to known samples.

Earlier studies mainly focused on estimating generated samples’ novelty without explicitly considering the creativity aspect from a human perception perspective. [Grace et al., 2015; Grace and Maher, 2019] argue that the novelty component of creativity should be considered relative to observers’ expectations rather than based solely on distance in the latent space of a generative model. Further, those novelty measures do not connect with the internal features of the generative model in a quantitative manner, which can provide explanation of the creative generation process. We propose a method to quantify the link between anomalous node activations in inner layers of generative processes and human perception of creativity.

3 Group-based Subset Scanning Over The Creative Decoder Activation Space

A visual overview of the proposed approach is shown in Figure 1. Subset scanning treats the creative quantification and characterization problem as a search for the most anomalous subset of observations in the data, this could be pixels or activations from a given layer of the generator. The corresponding exponentially large search space is efficiently explored by exploiting mathematical properties of our measure of anomalousness. Consider a set of samples from the latent space $X = \{X_1 \cdots X_M\}$ and nodes $O = \{O_1 \cdots O_J\}$ within the creative decoder $CD$ or $G$ (See Figure 1 for a graphical reference). Where $CD$ and $G$ is a generative neural network capable of producing creative outputs [Das et al., 2020]. Let $X_S \subseteq X$ and $O_S \subseteq O$, we then define the subsets $S$ under consideration to be $S = X_S \times O_S$. The goal is to find the most anomalous subset:

$$S^* = \arg \max_S F(S)$$

(1)

where the score function $F(S)$ defines the anomalousness of a subset of samples from the node activations of a given layer.
from $CD$ or $G$. Group-based subset scanning uses an iterative ascent procedure that alternates between two steps: a step identifying the most anomalous subset of samples for a fixed subset of nodes, or a step that identifies the converse. There are $2^M$ possible subsets of samples, $X_S$, to consider at these steps. However, the Linear-time Subset Scanning property (LTSS) [Neill, 2012; Speakman et al., 2016] reduces this space to only $M$ possible subsets while still guaranteeing that the highest scoring subset will be identified. This drastic reduction in the search space is the key feature that enables subset scanning to scale to large networks and sets of samples.

### 3.1 Non-parametric Scan Statistics (NPSS)

Group-based subset scanning uses NPSS that has been used in other pattern detection methods [McFowland III et al., 2013; Chen and Neill, 2014; Cintas et al., 2020]. Given that NPSS makes minimal assumptions on the underlying distribution of node activations, our approach has the ability to scan across different types of layers and activation functions. There are three steps to use the non-parametric scan statistics on the model’s activities. The first is to form a distribution of “expected” activations at each node ($H_0$). We generate this distribution by letting the generative process create samples that are known to be from the training data (sometimes referred to as “background” samples) and record the activations at each node. The second step involves scoring a group of samples in a test set that may contain creative or normal artifacts. We record the activations induced by the group of test samples and compare them to the baseline activations created in the first step. This comparison results in a $p$-value for each sample in the test set at each node. Lastly, we quantify the anomalousness of the resulting $p$-values by finding $X_S$ and $O_S$ that maximize the NPSS, which estimates how much an observed distribution of $p$-values deviates from the uniform distribution.

Let $A_{H_0}^{ij}$ be the matrix of activations from $l$ latent vectors of training samples at each of $J$ nodes in a creative decoder layer. Let $A_{ij}$ be the matrix of activations induced by $M$ latent vectors in the test set, that may or may not be novel. Group-based subset scanning computes an empirical $p$-value for each $A_{ij}$, as a measurement for how anomalous the activation value of a potentially novel sample $X_i$ is at node $O_j$. This $p$-value $p_{ij}$ estimates the proportion of activations from the $Z$ background samples, $A_{H_0}^{ij}$, that are larger or equal to the activation from an evaluation sample at node $O_j$.

$$p_{ij} = \frac{1 + \sum_{z=1}^{\vert Z \vert} I(A_{ij}^{H_0} \geq A_{ij})}{\vert Z \vert + 1} \quad (2)$$

Where $I(\cdot)$ is the indicator function. A shift is added to the numerator and denominator so that a test activation that is larger than all activations from the background at that node is given a non-zero $p$-value. Any test activation smaller than or tied with the smallest background activation at that node is given a $p$-value of 1.0.

Group-based subset scanning processes the matrix of $p$-values ($P$) from test samples with a NPSS to identify a submatrix $S = X_S \times O_S$ that maximizes $F(S)$, as this is the subset with the most statistical evidence for having been affected by an anomalous pattern. The general form of the NPSS score function is

$$F(S) = \max_{\alpha} F_{\alpha}(S) = \max_{\alpha} \phi(\alpha, N_{\alpha}(S), N(S)) \quad (3)$$

where $N(S)$ is the number of empirical $p$-values contained in subset $S$ and $N_{\alpha}(S)$ is the number of $p$-values less than (significance level) $\alpha$ contained in subset $S$. It has been shown that for a subset $S$ consisting of $N$ (empirical) $p$-values, $E[N_{\alpha}(S)] = N(S)\alpha$ [McFowland III et al., 2013]. Group-based subset scanning attempts to find the subset $S$ that shows the most evidence of an observed significance higher than an expected significance, $N_{\alpha}(S) > N(S)\alpha$, for some significance level $\alpha$.

In this work, we use the Berk-Jones (BJ) test statistic as our scan statistic. BJ test statistic is defined as:

$$\phi_{BJ}(\alpha, N, N) = N \ast KL \left( \frac{N_{\alpha}}{N}, \alpha \right) \quad (4)$$

where $KL$ refers to the Kullback-Liebler divergence, $KL(x, y) = x \log \frac{x}{y} + (1 - x) \log \frac{1-x}{1-y}$, between the observed and expected proportions of significant $p$-values. We can interpret BJ as the log-likelihood ratio for testing whether the $p$-values are uniformly distributed on $[0,1]$.

### 4 Experimental Setup

We hypothesize that creative content leaves a subtle but systematic trace in the activation space that can be identified by looking across multiple creative samples. Further, we assume that not all generative models will have the same throughput of creative samples in a batch. Thus, we need to evaluate our method under different proportions to see if even models that generate a small percentage of creative samples can be detected by our method. We test this hypothesis through group-based subset scanning over the activation space that encodes articles of artifacts that may appear anomalous when analyzed together. Given that the proposed approach is model agnostic, we test it under two generative models, a creative VAE Decoder and a Creative Generator variant of ArtGAN architectures [Das et al., 2020; Tan et al., 2019]. We scan both the pixel and activation space. We also validate the proposed approach across multiple datasets, including images from MNIST, Fashion-MNIST (FMNIST) [Xiao et al., 2017], Combo [Das et al., 2020], which is a combination of the previous two datasets, and WikiArt (with ArtGAN) [Tan et al., 2019].

In our experiments, we quantify detection power, that is the method’s ability to distinguish between test sets that contain some proportion of creative samples and test sets containing only normal content, using Area Under the Receiver Operating Characteristic Curve (AUROC) (See Figure 3B). Furthermore, we compute multiple existing surrogate metrics of novelty (e.g., in-domain score and reconstruction error, see Related Work) in the groups of artifacts found by our method to compare the novelty score of the found samples under group-based subset scanning. Further, for the WikiArt dataset, we conducted a questionnaire across multiple human evaluators to understand what makes a sample creative under this dataset and see if there is a correspondence between the anomalous
subset selected by our method and the creativity perception of the evaluator.

4.1 Datasets and Creative Labelling

For human evaluation under FMNIST, MNIST, and Combo datasets, we use the labeled artifacts from [Das et al., 2020] to test the performance of our proposed method. Nine evaluators did the human annotation over a pool of 500 samples per dataset (we used agreement amongst > 3 annotators as consensus), generated from either the creative generative process or regular decoding.

To evaluate the creative perception and our approach for a guided “creative” sampling process, we follow [Amabile, 1982], this technique measures creativity using an set of evaluators, who assess creative works individually and in isolation. The evaluators’ feedback is then collected and aggregated to establish an overall rating or measure. In our preliminary study, ten voluntary evaluators completed a questionnaire evaluating three sets of eleven images under the WikiArt dataset with ArtGAN, an example of one set can be found in Figure 3A. We show the baseline generated by standard ArtGAN and the sampled subset of artifacts selected by our method in the online form. These samples can belong to the ArtGAN output or the creative variant. We ask if the sample is creative compared to the initial image (See Figure 3C).

4.2 Subset Scanning and Base Models Setup

We use a VAE [Kingma et al., 2016] and ArtGAN [Tan et al., 2019] as the base generative models, and the “low-active” method proposed in [Das et al., 2020] as the creative process for our experiments. The “low-active” method identifies neurons that typically have low activation across all the training data and turns some number of them on at decoding time.

We run individual and group-based scanning on node activations extracted from the Creative VAE Decoder and the Creative ArtGAN Generator. We tested group-based scanning across several proportions of creative content in a group, ranging from 10% to 50%. We used $Z = 250$ latent vectors to obtain the background activation distribution ($A_{H0}$) for experiments with both datasets. For evaluation, each test set had samples drawn from a set of 100 normal samples from the regular decoder (separate from $Z$) and 100 samples labeled as creative and 100 non-creative samples (not novel or creative label).

5 Results

In Table 1 we present results showing the creative detection capabilities of both activation and pixel spaces. We see that the characterization improves when detecting the creative samples in the activation space, than when we scan over the pixel space. One can also observe that, even when the creative
The activations for different types of samples are distinctive. Notably, for FMNIST and Combo we start noticing some overlap for normal and creative samples. Based on this observation, we hypothesize that as more complex datasets are subject to creative decoding, we will see appearance of more overlapping nodes.

5.1 Creativity Art Perception in Anomalous Subgroups

Due to the lack of labeled samples in term of creativity in the generated WikiArt, we performed an unsupervised experiment (See Table 2) that employed extraction of the anomalous subsets. The model generates a low throughput (10%) of creative samples, the proposed method is capable of detecting these samples.

Additionally, as shown in Figure 2, for the creative-labeled datasets we observe a higher cardinality of anomalous nodes during creative generation, when compared to normal and non-creative in FMNIST and MNIST. This observation is consistent with the basic principle of the creative decoding process [Das et al., 2020], which induces a neuro-inspired atypical neuronal activation in the neural net generator to promote creativity. To further inspect the activations, we visualize the principal component projections of the anomalous subset of nodes for different sets of samples. As we can see, the activations for different types of samples are distinctive.
The word clouds containing all responses for each question to tag a sample as creative or not, we requested a brief description of what makes the sample creative and non-creative. Across the three example batches, on average 78% of the images were consistently found as creative by evaluators can be seen in Figure 3D. Non-creative models, to understand how we can better capture the human perception of creativity under more complex domains, such as the WikiArt dataset. We assess the anomaly-guided generations with existing surrogate metrics for novelty, as well as with a human evaluation study. Results consistently show the ability of nodes with anomalous activations within a deep generative model for identifying the novelty aspect of creative samples.

Additionally, the anomaly-guided generations were evaluated with respect to creativity by a group of human observers (Figure 3C). Across the three example batches, on average 78% of the images were consistently found as creative by evaluators’ perception. Apart from requesting the evaluator to tag a sample as creative or not, we requested a brief description of what makes the sample creative and non-creative. The word clouds containing all responses for each question from evaluators can be seen in Figure 3D. Non-creative modifications were usually described by evaluators as, similar to the base image, blurred artifacts, or not holding any structure with the initial image, and less variation of shapes and colors. While creative samples were described with changing one property (different color palette or variation on the structure).

### 6 Conclusion and Future Work

Our proposed method for creativity detection in machine-generated images works by analyzing the activation space for off-the-shelf generative models (such as Creative VAE Decoder and Creative ArtGAN Generator). We provide both the subset of the generated artifacts identified as creative and the corresponding nodes in the network’s activations that identified those samples as creative. We evaluate the proposed approach across computer vision datasets and different generative models, to understand how we can better capture the human perception of creativity under more complex domains, such as the WikiArt dataset. We assess the anomaly-guided generations with existing surrogate metrics for novelty, as well as with a human evaluation study. Results consistently show the ability of nodes with anomalous activations within a deep generative model for identifying the novelty aspect of creative samples.

Currently, the study faces limitations given the binary creative definition for the artifacts. Future research will consider questions to evaluators to provide a disentanglement of creativity properties such as novelty, surprise, and value. Hence, providing a more refined granularity will help better understanding of the connections between human perception of creativity and the activation space. Further, we plan to leverage the proposed creativity quantification approach as a control for more efficient and trusted generation of artifacts that are consistent with human perception of the novelty component of creativity.
References


