

---

# TOWARDS CREATIVITY CHARACTERIZATION OF GENERATIVE MODELS VIA GROUP-BASED SUBSET SCANNING

**Celia Cintas**  
IBM Research Africa  
Nairobi, Kenya

**Payel Das**  
IBM Research  
Yorktown Heights, NY, USA

**Brian Quanz**  
IBM Research  
Yorktown Heights, NY, USA

**Skyler Speakman**  
IBM Research Africa  
Nairobi, Kenya

**Victor Akinwande**  
IBM Research Africa  
Nairobi, Kenya

**Pin-Yu Chen**  
IBM Research  
Yorktown Heights, NY, USA

## ABSTRACT

Deep generative models, such as Variational Autoencoders (VAEs), have been employed widely in computational creativity research. However, such models discourage out-of-distribution generation to avoid spurious sample generation, 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 to 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 quantify, detect, and characterize creative processes by detecting a subset of anomalous node-activations in the hidden layers of generative models. Our experiments on original, typically decoded, and “creatively decoded” (Das et al., 2020) image datasets reveal that the proposed subset scores distribution is more useful for detecting 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 normal sample generation.

## 1 INTRODUCTION

Creativity is a process that provides novel and meaningful ideas (Boden, 2004). Current deep learning approaches open a new direction enabling the study of creativity from a knowledge acquisition perspective. 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. However, such models discourage out-of-distribution generation to avoid instability and decrease spurious sample generation, limiting their creative generation potential. Novelty of the generated samples is often used as a proxy for human perception of creativity in those studies. Therefore, earlier studies mostly focus on estimating the novelty of generated samples, without explicitly considering to the creativity aspect of human perception. Further, those novelty measures do not connect with the generative model features in a quantitative manner, which can provide explanation of the creative generation process. The design of creativity evaluation schemes is as essential as developing creative generative methods. Multiple aspects of creativity need to be better defined to allow the research community to develop and test hypotheses systematically (Cherti et al., 2017). There are multiple surrogate metrics for novelty (Wang et al., 2018; Ding et al., 2014; Kliger & Fleishman, 2018) in the literature; however, the ultimate test of creativity is done by human inspection. Human labelling 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). Although human judgement 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. This paper proposes a method designed to detect and characterize when the generative model produces a creative artefact

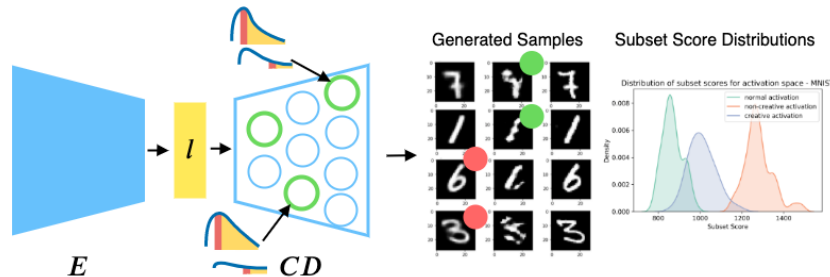


Figure 1: Overview of the proposed approach. First, we analyze the distribution of the activation space of the Creative Decoder  $CD$ . After we extracted the activations from the model for a set of latent vectors  $l$ , we compute the empirical  $p$ -values followed by the maximization of non-parametric scan statistics (NPSS). Finally, distributions of subset scores for creative, non-creative processes are estimated, a subset of samples and the corresponding anomalous subset of nodes in the network are identified.

as per a human evaluator. We employ group-based scanning to determine whether a given batch of generated processes contains creative samples using an anomalous pattern detection method called group-based subset scanning (Neill, 2012; McFowland III et al., 2013).

## 2 PROPOSED APPROACH: 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 characterisation problem as a search for the *most anomalous* subset of observations in the data. This 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  $CD$ . Where  $CD$  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:

$$S^* = \arg \max_S F(S) \quad (1)$$

where the score function  $F(S)$  defines the anomalousness of a subset of samples from the latent space and node activations. 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.

**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; McFowland et al., 2018; Chen & Neill, 2014; Cintas et al., 2020; Akinwande 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 type of layers and activation functions. There are three steps to use non-parametric scan statistics on model’s activation data. The first is to form a distribution of “expected” activations at each node ( $H_0$ ). We generate the distribution by letting the regular decoder process 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 records the activations induced by the group of test samples and compares them to the baseline activations created in the first step. This comparison results in a  $p$ -value at each node, for each sample from the latent space in the test set. Lastly, we quantify the anomalousness of the resulting  $p$ -values by finding  $X_S$  and  $O_S$  that maximize the NPSS, which quantify how much an observed distribution of  $p$ -values deviates from the uniform distribution.

Let  $A_{zj}^{H_0}$  be the matrix of activations from  $l$  latent vectors from 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  $p_{ij}$  is the proportion of activations from the  $Z$  background samples,  $A_{zj}^{H_0}$ , that are larger or equal to the activation from an evaluation sample at node  $O_j$ .

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

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  $N(S)$  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 (Berk & Jones, 1979) is defined as:

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

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  $[0, 1]$ .

### 3 EXPERIMENTAL SETUP AND RESULTS

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 *groups of samples* that may appear anomalous when analyzed together. We apply our approach to the Creative Decoder and scan the both the pixel/input and activation space Das et al. (2020). We used images from MNIST (LeCun et al., 1998) and Fashion-MNIST (FMNIST) (Xiao et al., 2017) datasets. 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 AUC.

**Datasets and Creative Labelling** For human evaluation, 9 evaluators annotated a pool of 500 samples per dataset (we used agreement amongst  $> 3$  annotators as consensus), generated from either using the Creative Decoder, and regular decoding. Following Das et al. (2020), we used four labels - ‘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’.

**Subset Scanning Setup** We run individual and group-based scanning on node activations extracted from the Creative Decoder. 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 ( $A^{H_0}$ ) for experiments with both datasets. For evaluation, each test set had samples were 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).

Table 1: **Detection Power** (AUC) for group-based and individual subset scanning over pixel and activation space for the Creative Decoder.

Space	Dataset	Subset Scanning		
		50%	10%	Indv.
Pixel Space	MNIST	0.971	0.791	0.255
Activation Space	MNIST	<b>0.991</b>	<b>0.972</b>	0.531
Pixel Space	FMNIST	0.952	0.743	0.381
Activation Space	FMNIST	<b>0.990</b>	<b>0.962</b>	0.596

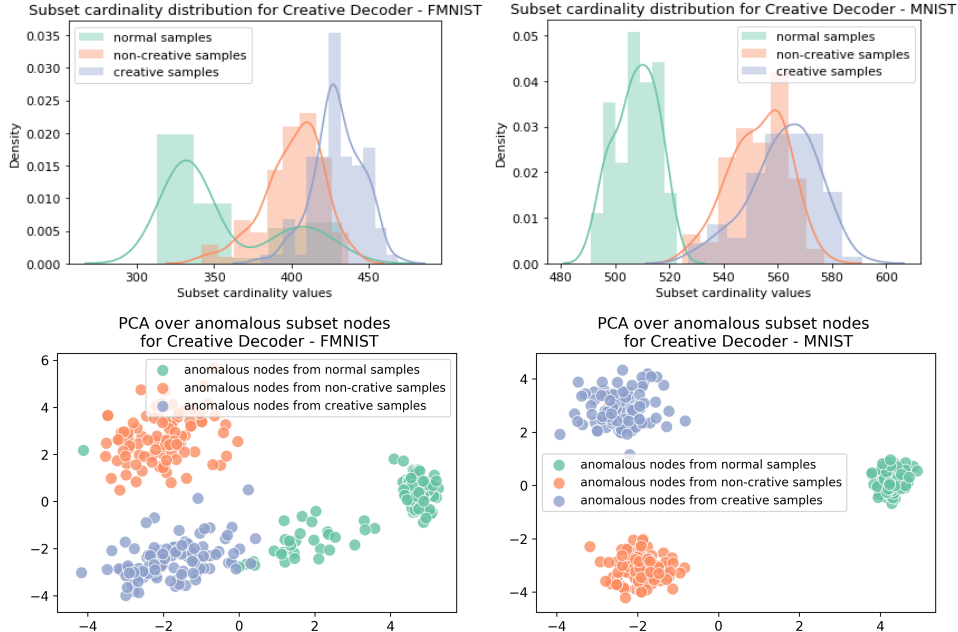


Figure 2: Activations characterization. **(top)** Subset Cardinality distributions for anomalous subsets for different type of generated samples. **(bottom)** PCA over anomalous subset nodes for the creative decoder activations under generation of normal, non-creative and creative samples.

**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. Additionally, in Figure 2 for both datasets we observe a larger extent of anomalous nodes during creative generation compared to normal and non-creative. This observation is consistent with the basic principle of the creative decoding process (Das et al., 2020). 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. Notably, for FMNIST 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.

#### 4 CONCLUSION AND FUTURE WORK

Our proposed method for creativity detection in machine-generated images works by analyzing the activation space for an off-the-shelf Creative Decoder. We provide both the subset of the input samples identified as creative and the corresponding nodes in the network’s activations that identified those samples as creative. In future, we will compare the proposed creativity quantification with other surrogate metrics for novelty (Wang et al., 2018; Ding et al., 2014; Kliger & Fleishman, 2018). Additionally, we will test across larger datasets and other generative models to understand how we can better capture the human perception of creativity under more complex domains. The final goal is to use this creativity quantification approach as a control for more efficient generation of artefacts that are consistent with human perception of creativity.

---

## REFERENCES

- Victor Akinwande, Celia Cintas, Skyler Speakman, and Srihari Sridharan. Identifying audio adversarial examples via anomalous pattern detection. In *Workshop on Adversarial Learning Methods for Machine Learning and Data Mining, KDD'20*, 2020.
- Robert H. Berk and Douglas. H. Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 47:47–59, 1979.
- Margaret A Boden. *The creative mind: Myths and mechanisms*. Routledge, 2004.
- Feng Chen and Daniel B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *KDD '14*, pp. 1166–1175, 2014.
- 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*, 2017.
- Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan, and Edward McFowland. In Christian Bessiere (ed.), *IJCAI 2020*, pp. 876–882, 7 2020. doi: 10.24963/ijcai.2020/122. URL <https://doi.org/10.24963/ijcai.2020/122>. Main track.
- Payel Das, Brian Quanz, Pin-Yu Chen, Jae-wook Ahn, and Dhruv Shah. Toward a neuro-inspired creative decoder. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 2746–2753. International Joint Conferences on Artificial Intelligence Organization, 7 2020. Main track.
- Xuemei Ding, Yuhua Li, Ammar Belatreche, and Liam P Maguire. An experimental evaluation of novelty detection methods. *Neurocomputing*, 135:313–327, 2014.
- 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.
- 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.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Mark Kliger and Shachar Fleishman. Novelty detection with gan. *arXiv preprint arXiv:1802.10560*, 2018.
- Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- CS Lopez and CE Tucker. Human validation of computer vs human generated design sketches. *ASME Paper No. DETC2018-85698*, 2018.
- E. McFowland, III, S. Somanchi, and D. B. Neill. Efficient Discovery of Heterogeneous Treatment Effects in Randomized Experiments via Anomalous Pattern Detection. *ArXiv e-prints*, March 2018.
- Edward McFowland III, Skyler D Speakman, and Daniel B Neill. Fast generalized subset scan for anomalous pattern detection. *The Journal of Machine Learning Research*, 14(1):1533–1561, Jun 2013.
- Daniel B. Neill. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)*, 74(2):337–360, 2012.

- 
- Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, pp. 2234–2242, 2016.
- Skyler Speakman, Sriram Somanchi, Edward McFowland III, and Daniel B. Neill. Penalized fast subset scanning. *Journal of Computational and Graphical Statistics*, 25(2):382–404, 2016. doi: 10.1080/10618600.2015.1029578. URL <https://doi.org/10.1080/10618600.2015.1029578>.
- 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.
- Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.