DeepCreativity: Measuring Creativity with Deep Learning Techniques

Giorgio Franceschelli [a,*] and Mirco Musolesi [b]

[a] *Alma Mater Studiorum Università di Bologna, Italy*
E-mail: giorgio.franceschelli@unibo.it
[b] *University College London, United Kingdom,*
The Alan Turing Institute, United Kingdom,
Alma Mater Studiorum Università di Bologna, Italy
E-mail: m.musolesi@ucl.ac.uk

Abstract. Measuring machine creativity is one of the most fascinating challenges in Artificial Intelligence. This paper explores the possibility of using generative learning techniques for automatic assessment of creativity. The proposed solution does not involve human judgement, it is modular and of general applicability. We introduce a new measure, namely *DeepCreativity*, based on Margaret Boden's definition of creativity as composed by *value*, *novelty* and *surprise*. We evaluate our methodology (and related measure) considering a case study, i.e., the generation of 19th century American poetry, showing its effectiveness and expressiveness.

Keywords: Computational Creativity, Deep Learning, Creativity Measure, American Poetry

1. Introduction

Evaluation is a crucial concern in Artificial Intelligence and in science more in general. Measures and metrics are fundamental not only to check the validity of a hypothesis, but also to understand if it is possible to use some given results with confidence as a starting point for future research. An example is Shannon's entropy, which plays a central role as a measure of information, choice and uncertainty [20] and underpins many results in Information Theory [39]. The role of measures is even more crucial in machine learning, where models need to be evaluated during and after training. For example, most of the machine learning algorithms are evaluated using metrics like accuracy.

Excellent progress has been made in benchmark tasks coupled with metrics used to estimate the performance of algorithms [34]: examples include [32] for machine translation or [36] for image generation. At the same time, there is also a need to derive new metrics for examining the behavior of algorithms in different environments and in relation with society [34]. Among the spectrum of behaviors that could be exhibited by a machine, creativity is certainly one of the most interesting [30]. In fact, we have witnessed the emergence of an entire new field of research, namely Computational Creativity, with a focus on the study of the behaviors exhibited by artificial systems that would be deemed as creative [7,48]. Indeed, one of the key goals of this field is the definition of computational techniques for measuring creativity.

In this paper, we present a novel methodology (and related measure) as a way to assess the creativity of an artifact. In particular, *DeepCreativity* is based on the very famous definition of creativity provided by Margaret Boden: "creativity is the ability to come up with ideas or artifacts that are new, surprising and valuable" [5]. Although it is not the only definition of creativity available (over one hundred of definitions have been proposed over the years [1,42]), it is central in the field of computational creativity. The three aspects that are

---
*Corresponding author. E-mail: giorgio.franceschelli@unibo.it.

at the basis of this metric have played a prominent role in the scholarly discourse around the definition of machine creativity. Our proposed measure uses deep learning techniques, avoiding the need of a human in the process, to measure how much an artifact is valuable, novel and surprising with respect to a given context, and, therefore, to quantify the ability of an agent to come up with creative results. To the best of our knowledge, this is one of the first attempts to define an evaluation method for assessing the overall creativity of an artifact that is automatic and of general applicability.

This work is structured as follows: a review of the literature about automatic methods to assess creativity is presented in Section 2; then, in Section 3 we discuss the proposed creativity measure. An evaluation of *DeepCreativity* is presented in Section 4, considering a case study of text generation in the context of 19th century American poetry; finally, we discuss limitations of the proposed approach and potential future work in Section 5.

2. Related Work

Over the years, several computational approaches have been proposed to automatically assess the creativity in products made by (human or artificial) agents, differing in the scope of evaluation or in the method. A complete survey can be found in [14]. All of them consider value and novelty as aspects of creativity, while only some of them also include surprise. In the following, we will consider the three factors separately.

2.1. Value

Value, sometimes referred as quality, expresses how an artifact compares to others in its class in terms of utility, performance or attractiveness. It is typically defined as a weighted sum of performance attributes or as a reflection of the acceptance of the artifact by society [25]. The authors of [11] follow the latter definition, which suggests to compute creativity by using an art graph where each vertex represents an artwork and each arc connecting an older to a newer work is labeled with the similarity between the two calculated by means of an appropriate similarity function. The higher the similarity with subsequent works, the higher the value (and the higher the creativity). However, this method does not allow to compute value for the most recent works, but only for the older ones. The former

definition is more common in the literature. For instance, in [25] the authors suggest to derive value as the weighted sum of pre-defined performance variables. In [26], value is defined using clusters of artifacts built on a performance space - with artifacts expressed as sets of attribute-value pairs. The authors of [15] define it as the synergy [8] between artifacts, expressed following the regent-dependent model. Also several domain-specific methods follow the definition of value as the sum of performance attributes or performance measures: for example, for poetry generation, the authors of [50] consider topic distribution (through LDA), fluency (through a neural language model) and coherence (through mutual information and TF-IDF) as components of value. In [52] coherence is used (through BLEU, originally proposed for machine translation in [32]) with quality (through perplexity), while the authors of [51] uses BLEU only. However, the definition of value as the weighted sum of sub-components has the limitation of requiring the correct identification of all the relevant factors and their relative weights, which is a complex and time-consuming task.

2.2. Novelty

Novelty is commonly defined as the measure of how much an artifact differs from known artifacts in its class [25]. For this reason, a classic technique to measure novelty consists in the calculation of the distance between a given artifact and the other artifacts on a descriptive space, as discussed in [25] and [26]. The descriptive space is usually identified by the attributes used to define the artifacts. Similarly, domain-specific methods consider novelty in terms of distance or dissimilarity: for instance, in case of text generation, the authors of [21] consider novelty as the average semantic distance between the dominant terms included in the textual representation of the story, compared to the average semantic distance of the dominant terms in all stories. In [50] diversity and innovation in poetry generation are measured by means of bigram-based average Jaccard similarity. As for value methods, the requirement of defining artifacts in terms of attributes appears to be as a rather strong limitation.

A different definition of novelty has been proposed in [3], namely as the degree an input differs from what an observer has experienced before. In [11] novelty is defined by considering the time dimension of personal experience: the lower the degree of similarity between an artifact and the previous works, the higher the novelty contribution of creativity. Even if not ex-

actly used as an evaluation technique, in [10] a novelty score is proposed to guide the training of the generative part of the Creative Adversarial Network. This can be considered as a *creative-oriented* variant of Generative Adversarial Networks (GANs) [16]. In addition to the classic adversarial loss provided by the discriminative model, the generator is trained to maximize a novelty loss that represents how much the generated artifact differs from previous works in terms of style. Although considering novelty as the deviation from style norms is somehow simplistic, it only requires a style classifier, automatically capturing an important aspect of novelty at the same time.

2.3. Surprise

In [3], surprise is defined as the degree of disagreement between the real input and what it was expected in its place. This classic definition of surprise based on unexpectedness is typically also referred to as surprisal [43]. In [25], unexpectedness is calculated considering whether or not the artifact follows the expected next artifact in the pattern recognized on recent artifacts. In [17], surprise is measured as the unlikelihood of observing a particular artifact according to the predictions about relationships between its attributes. In the specific domain of text generation, in [21], surprise is defined as the average semantic distance between consecutive fragments of each story. For sequential artifacts like texts or sounds, the authors of [6] adopt the expected maximum surprise as measure (as one minus the probability of the most unexpected token of the artifact) and the expected count of ψ-surprise (as the count of all the tokens for which predictability is lower than a given threshold $\frac{\psi}{K}$), where the expectations are provided by an *audience* neural network. In a similar way, [24] proposes to quantify surprise considering both the probability of the event X of interest and the probability of the most probable event Y, since the surprise of an event X also depends on the certainty of Y (e.g., ten equiprobable events have a very high unexpectedness, but they should have a very low surprise, since we are not surprised to see one of them occurring).

A quite different approach is adopted in [26], where the authors consider a new artifact as surprising if it creates a new cluster in the conceptual space (instead of perfectly fitting into an existing one). The idea of surprise as related with the difference between prior and posterior models is at the basis of Bayesian Surprise [2], used in [15] and [46]. It is a measure of sur-

prise in terms of the impact of a data point that changes a prior distribution into a posterior distribution, calculated applying Bayes' theorem (considering artifacts as a composition of attributes); here, surprise is the postobservation change rather than the prediction error.

3. Measuring Creativity using Deep Learning

We now present *DeepCreativity*, a new Deep Learning creativity measure. The goal is to define a measure of more general applicability. Deep Learning is used for avoiding the need of identifying the required attributes to describe the artifacts or the components of creativity [14]. This leads to a measure that allows for automatic evaluation of artifacts. As discussed in Section 1, *DeepCreativity* is based on the definition of creativity proposed by [5]. Therefore, the measure is based on three main factors, which will be explored in the next subsections separately: value (Subsection 3.1), novelty (Subsection 3.2) and surprise (Subsection 3.3). Finally, in Subsection 3.4, we will put everything together by providing a unified definition of creativity.

3.1. Value

We measure *value* by means of the discriminative part of a Generative Adversarial Network [16]. The GAN is trained by considering the real artifacts as the true ones; in this way, the discriminative model should learn a representation of real (and valuable) data, and its evaluation of a new artifact provides insights of its value in that context. Therefore, the value of an artifact a over the value discriminator D_v can be expressed as:

$$V(a, D_v) = D_v(a), \tag{1}$$

with $V(a, D_v)$ naturally constrained between 0 (not valuable at all) and 1 (highly valuable), since a sigmoid activation is applied to the output layer of D_v.

The choice of the real artifacts influence the value measure proposed above (acting as an Inspiring Set [35]). While it can be seen as a limitation of the approach, it is highly coherent with the nature of creativity itself. Creativity, and in particular value, are deeply *context-dependent*: the same work, proposed in two different moments of history or to two different social groups may be evaluated differently [5]. Under this lens, the need of real artifacts conceals the op-

portunity of representing, within the measure, a fundamental aspect of creativity. The real data used during GAN's training will therefore represent a specific context, well-defined in temporal and cultural terms.

To train the GAN, it is important to distinguish between continuous tasks (like image generation) and sequential tasks (like text or sound generation). With respect to continuous applications, a GAN can be trained using the following loss function [16]:

$$L = \min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)]$$
$$+ \mathbb{E}_{z \sim p_z(z)}[\log (1 - D(G(z)))], \quad (2)$$

with p_{data} as the real data distribution (representing the desired context), p_z as the input noise variable, and with discriminator D and generator G trained alternately; notice that several refinements have been proposed in the recent years (see [18] for potential variations).

As far as sequential applications are concerned, the impossibility of directly applying GAN to these tasks is a well-known problem [19]. A common way to solve it is by using SeqGAN [51]. SeqGAN considers the sequence generation process as a sequential decision making process, defining a reinforcement learning framework in which the generative model G_θ is the agent, the actual state $(y_1, ..., y_{t-1})$ is composed by the generated tokens so far, the next action y_t is the next token to be generated, and the reward is the evaluation provided by the discriminative model D_ϕ. The generative model is then seen as a stochastic parametrized policy; Monte Carlo search is used to approximate the state-action value and directly train the policy via policy gradient [51]. More specifically, the REINFORCE algorithm [49] for learning the policy (but other methods can be used as well [12]), which leads to the following update rule:

$$\theta \leftarrow \theta + \alpha Q_{D_\phi}^{G_\theta}(Y_{1:t-1}, y_t) \nabla_\theta \ln G_\theta(y_t | y_{1:t-1}), \quad (3)$$

where Q is the expected return obtained by the N-time Monte Carlo search.

3.2. Novelty

With respect to *novelty*, our definition is inspired by CAN [10] although the deviation from style norms

cannot be used directly to measure the difference between artifacts. Therefore, as additionally done by the CAN discriminator, a neural network D_n is trained to correctly recognize the style of real artifacts (from the given context). The neural network can just be a simple classifier (as in [31] for music or in [40] for paintings), outputting a probability vector of length N equal to the number of possible classes. Consequently, a novelty measure can be defined as:

$$N(a, D_n) = 1 - \frac{\sqrt{\sum_{i=1}^N \left(\frac{1}{N} - y_i\right)^2}}{UB}$$
$$\text{with} \quad UB = \frac{\sqrt{N(N-1)}}{N}, \quad (4)$$

where y is the output vector (of length N and sum 1) of D_n given in input artifact a. The formula computes the Euclidean Distance between y and the desired target vector of equiprobable values; in addition, it is constrained between 0 and 1, where it is equal to 1 when the distance is minimum (i.e., when the two vectors are equal) and it is equal to 0 when the distance is maximum (i.e., when a one-hot vector is considered). Please refer to Appendix A for the proof of this property.

3.3. Surprise

With respect to *surprise*, we follow the conceptual framework presented in [2]. Starting from a sequential generative model G_s trained to predict the next token given the previous ones on an appropriate training set (temporally and culturally defined, as stated for value), this allows for considering the impact of an artifact a over G_s. Its influence is calculated using a weight correction applied over G_s if G_s is trained to correctly predict a. In analogy with the Bayesian Surprise, surprise is measured as the distance between prior G_s (before training) and posterior G_s (after training on a). The difference is in how the posterior distribution is obtained, namely not by means of Bayes' theorem, but through backpropagation and gradient descent. Notice that this idea is very close to the intrinsic reward presented in [37], where a measure of surprise is derived by maximizing a distance function between prior and posterior distribution of a predictive model.

At inference time, only measuring surprise is relevant, while the model update is not actually required. It is only used to compute the weight correction Δw_{ji},

which expresses how much the posterior distribution will differ from the prior. Given an artifact $a = \{a_1, a_2, ..., a_N\}$, the mini-batch (of size N) gradient descent formula for Δw_{ji} can be used:

$$\Delta w_{ji} = -\eta \frac{1}{N} \sum_{k=1}^{N} \frac{\partial J_k}{\partial w_{ji}}, \tag{5}$$

where η is the learning rate and J_k is the loss function considering token k[1].

We can now define the surprise measure more formally. Given a sequential generative model G_s, an artifact a has a surprise over G_s equal to:

$$S(a, G_s) = avg_{j,i} \left| \frac{\Delta w_{ji}}{w_{ji}} \right|. \tag{6}$$

We note that the correction is divided by the weight to represent the degree of correction, i.e., the influence of the artifact. Then, the learning rate in Eq. (5) is not the learning rate used during training, but a parameter to adjust the magnitude of correction for the surprise measure. Even a value of 1 can be reasonable in certain problems. Finally, this approach requires G_s in order to consider artifacts as sequential data, even if they are continuous. In case of image, G_s may be, for instance, an autoregressive model [33,44,45].

3.4. Putting All Together

Given the definition of $V(a, D_v), N(a, D_n), S(a, G_s)$ in the previous subsections, the *DeepCreativity* measure (indicated with DC) is obtained by computing the creativity of a generative agent producing artifact a over a temporal and cultural context TCC as:

$$DC(a, TCC) = \alpha_1 V(a, D_v) +$$
$$\alpha_2 N(a, D_n) + \tag{7}$$
$$\alpha_3 S(a, G_s),$$

where $\alpha_1, \alpha_2, \alpha_3 \in [0, 1]$ and $\alpha_1 + \alpha_2 + \alpha_3 = 1$. D_v, D_n and G_s are trained over TCC, which is a set of ex-

amples $(x_1, y_1), ..., (x_n, y_n)$ where $x_1, ..., x_n$ are the real artifacts, and $y_1, ..., y_n$ are their labels representing the class (assuming N different values). $\alpha_1, \alpha_2, \alpha_3$ weight the three single components of creativity; the immediate setting is to consider them as equal, as we will do in the following experiments. Nonetheless, it is possible to change them according to the specific domain, if some of the properties are found as more relevant in creativity assessment.

4. Experiments

There is no common agreement about how to evaluate creativity measures. All the methodologies discussed in Section 2 have not been evaluated against a ground truth; on the contrary, they have just been tested over a generative system, in comparison with human judgements (always considering the products of a generative system) or they have not been tested at all. This can be attributed to the difficulty of finding a common definition of creativity, which is reflected in the lack of correct evaluation of creative productions.

However, a ground truth about this creativity process exists in this case: art history. The fact that in a certain moment of history, in a certain place, an artwork was appreciated or at least considered of sufficient quality to be "printable" may be used as useful information for evaluating a creative agent. Inspired by considerations done in [30] about CAN and its ability of intercepting the historical trajectory of art, a meta-evaluation test is defined, based on historical trajectories, to study if and how the proposed measure is able to correctly capture the changes of creativity over time in a fixed culture. In particular, the following experiment will concern the context of American Poetry. American poetry has been chosen because of the importance of this artistic production and its heterogeneity in terms of movements and styles, while still referring to a specific cultural context. Other cultures or artistic fields could have been selected as well; while the study presented in this work aims at demonstrating the effectiveness of *DeepCreativity* in principle, we are aware that further work is necessary to demonstrate its generality (see Section 5).

The goal of this experiment is to measure the creativity of poems from different moments of history, but training the neural networks for the computation of *DeepCreativity* on a specific historical context. *Deep-Creativity* can therefore be considered an appropriate creativity measure if the resulting creativity is

---

[1]It is worth noting that the loss function represents in a way the expectation error, i.e., the surprisal.

higher for the artworks which really come after the context, because these are the ones been considered as highly creative in that moment. Consequently, it should also recognize the other works as less creative: later works should be judged more novel and surprising, but less valuable and understandable; while contemporary works should be judged more valuable but less novel and surprising.

Two separated experiments are conducted to verify if the measure is able to *capture* creativity for a certain period of time. Both of them involve poems from the 19th century (i.e., from American Renaissance (Brahmins and Romantics), Local Color, Naturalism, and Neogothic (or Protodecadentism)) as the context. The first experiment involves poems from the 20th century (i.e., a selection of poems from Imagism, Harlem Renaissance, Objectivism, Beat Generation, and Confessional Movement). The second one also considers poems from the 17th and the 18th century (i.e., a selection of poems from Puritanism, African-American Poetry, and American Revolution). A sample of poems from the training set is always considered for a complete comparison; full details are reported in Appendix C. As far as details about the implementation are concerned, two types of NNs have been used: a LSTM-based RNN for the generative models, and a CNN for the discriminative models. Full details about architectures and training processes can be found in Appendix B.
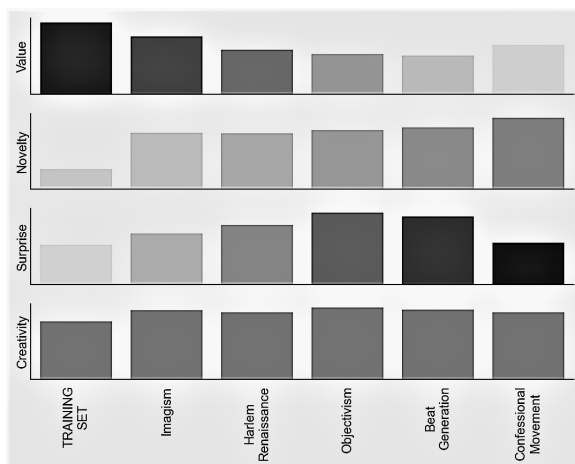

Fig. 1. The average of value, novelty, surprise and creativity computed on a sample from the training set and on 20th century American poems.

With respect to the first experiment, Figure 1 shows the average of the creativity components during move-ments and the final creativity measure. It is interesting to note that the higher the novelty the further from the training set. This correctly captures the fact that a movement, which immediately follows a certain period has to be novel with respect to it. Moreover, the next movement has to be novel with respect to both the works produced in that period and the first one. The surprise curve generally also shows a similar behavior: temporally distant artifacts are the result of different contexts and different situations and they are more difficult to be predicted only considering a *past* version of the same culture. The last movement, the Confessional one, could be considered as an exception. This can be explained by considering how surprise is measured: it is calculated as the degree of change that the work causes over a 19th century American poems model, which is strictly related to a semantic view of the context, because it is based on the content. Indeed, temporally far movements might have a lower surprise measure if their themes (e.g., love) are semantically closer to those in the training set. The same consideration can be done to explain the value curve. For the first four movements, it tends to decrease with time, as expected. On the other side, Confessional Movement has a higher value; since its semantic content is closer to the one from the context, it results in a more similar and therefore comprehensible and admirable style, with a higher value.

In general, it is possible to observe that creativity tends to decrease further in time from the period of reference of the training set, while it is higher for the central movement, which is able to conciliate a high degree of surprise without a consistent loss in value.

With respect to the second experiment, Figure 2 shows the three components also considering previous movements. This should help study the appropriateness of the three measures over the time dimension. It is therefore interesting to note that two over three curves follow the same trends observed for the subsequent century. Novelty is the only one behaving differently, since the closest movement has more or less the same novelty than the furthest. On the contrary, value decreases further away from the period of reference of the training set, as desirable; in the same way, surprise increases. In addition, it is interesting to note that surprise is smaller than for the 20th century on average. This can be explained by observing that the 19th century poems include some knowledge about the previous poems, making them more predictable.
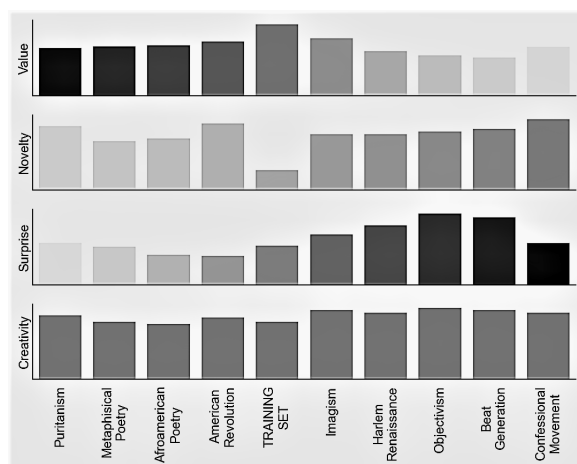
Fig. 2. The average of value, novelty, surprise and creativity computed on a sample from the training set and on both 18th and 20th century American poems.

5. Conclusion and Future Work

In this work, we have introduced *DeepCreativity*, a new creativity measure based on three components with the objective of measuring the value, novelty and surprise of a generative process or algorithm, in terms of their products. This general approach overcomes the limits of having measures applicable only to certain domains; in addition, the use of deep learning techniques overcomes the limits of having to manually define the attributes or the components which characterize creativity. Moreover, the automation of evaluation allows for embedding *DeepCreativity* in the creative generation itself; for instance, a generative model can be trained to learn to maximize such a measure [4] or it can be used inside a generate-and-test iterative process [41]. Finally, the need of a training set allows for the definition of a specific context of evaluation, which has been found to be a fundamental constraint of creativity. However, few limitations can also be found: novelty only considers the style or the genre, while it might lie in other traits of a work; surprise requires a sequential generator, which could be not optimal for (supposedly simpler) continuous tasks.

The experiments conducted in the context of generative learning of 19th century American poetry have demonstrated that the measure is able to capture the historic trajectory of creativity over time, either only considering future poems or also previous ones, showing its effectiveness. Additional tests should be carried out in order to confirm the correctness and the general applicability of the measure, ideally in different domains; contemporary contexts should be considered too, in order to have the evaluation of *DeepCreativity* validated against human judges (as in [10]). This is part of our future research agenda.

References

[1] A. G. Aleinikov, S. Kackmeister, and R. Koenig. *Creating Creativity: 101 Definitions (what Webster Never Told You)*. Alden B. Dow Creativity Center Press, 2000.

[2] P. Baldi and L. Itti. Of Bits and Wows: A Bayesian Theory of Surprise with Applications to Attention. *Neural networks: the official journal of the International Neural Network Society*, 23:649–666, 2010.

[3] D. E. Berlyne. *Aesthetics and psychobiology*. Appleton-Century-Crofts, 1971.

[4] S. Berns and S. Colton. Bridging Generative Deep Learning and Computational Creativity. In *ICCC 2020*, pages 406–409, 2020.

[5] M. A. Boden. *The Creative Mind: Myths and Mechanisms*. Routledge, 2003.

[6] R. C. Bunescu and O. O. Uduehi. Learning to Surprise: A Composer-Audience Architecture. In *ICCC 2019*, pages 41–48. Association for Computational Creativity (ACC), 2019.

[7] S. Colton and G. A. Wiggins. Computational Creativity: The Final Frontier? In *ECAI 2012*, 2012.

[8] P. Corning. *Nature's Magic: Synergy in Evolution and the Fate of Humankind*. Cambridge University Press, 2003.

[9] J. Duchi, E. Hazan, and Y. Singer. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[10] A. Elgammal, B. Liu, M. Elhoseiny, and M. Mazzone. CAN: Creative Adversarial Networks, Generating "Art" by Learning About Styles and Deviating from Style Norms. In *ICCC 2017*, 2017.

[11] A. Elgammal and B. Saleh. Quantifying Creativity in Art Networks. In *ICCC 2015*, pages 39–46, 2015.

[12] W. Fedus, I. Goodfellow, and A. M. Dai. MaskGAN: Better Text Generation via Filling in the _____. In *ICLR 2018*, 2018.

[13] D. Foster. *Generative Deep Learning*. O'Reilly, 2019.

[14] G. Franceschelli and M. Musolesi. Creativity and Machine Learning: A Survey, 2021. arXiv:2104.02726 [cs.LG].

[15] C. França, L. F. W. Góes, A. Amorim, R. C. O. Rocha, and A. R. Da Silva. Regent-Dependent Creativity: A Domain Independent Metric for the Assessment of Creative Artifacts. In *ICCC 2016*, 2016.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *NeurIPS 2014*, pages 2672–2680. Curran Associates, Inc., 2014.

[17] K. Grace and M. L. Maher. What to expect when you're expecting: The role of unexpectedness in computationally evaluating creativity. In *ICCC 2014*, 2014.

[18] J. Gui, Z. Sun, Y. Wen, D. Tao, and Y. Jie-ping. A Review on Generative Adversarial Networks: Algorithms, Theory, and Applications, 2020. arXiv:2001.06937 [cs.LG].

[19] F. Huszár. How (not) to Train your Generative Model: Scheduled Sampling, Likelihood, Adversary?, 2015. arXiv:1511.05101 [stat.ML].

[20] A. Jha. Without Claude Shannon's information theory there would have been no internet. _The Guardian_, 2016. https://www.theguardian.com/science/2014/jun/22/shannon-information-theory.

[21] P. Karampiperis, A. Koukourikos, and E. Koliopoulou. Towards Machines for Measuring Creativity: The Use of Computational Tools in Storytelling Activities. In _ICALT 2014_. IEEE, 2014.

[22] Y. Kim. Convolutional Neural Networks for Sentence Classification. In _EMNLP 2014_, pages 1746–1751, 2014.

[23] D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, 2014. arXiv:1412.6980 [cs.LG].

[24] L. Macedo, R. Reisenzein, and A. Cardoso. Modeling forms of surprise in artificial agents: empirical and theoretical study of surprise functions. In _CogSci 2004_, pages 873–878, 2004.

[25] M. Maher. Evaluating creativity in humans, computers, and collectively intelligent systems. In _DESIRE 2010_, pages 22–28, 2010.

[26] M. Maher and D. Fisher. Using AI to Evaluate Creative Designs. _ICDC 2012_, 1, 2012.

[27] T. Margoni. Artificial Intelligence, Machine Learning and EU Copyright Law: Who Owns AI? _CREATe Working Paper_, 2018.

[28] S. Meemulla Kandi. _Language Modelling for Handling Out-of-Vocabulary Words in Natural Language Processing_. PhD thesis, London School of Economics and Political Science, 2018.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space, 2013. arXiv:1301.3781 [cs.CL].

[30] A. I. Miller. _The Artist in the Machine_. The MIT Press, 2019.

[31] J. Nam, K. Choi, J. Lee, S.-Y. Chou, and Y.-H. Yang. Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from Bach. _IEEE Signal Processing Magazine_, 36(1):41–51, 2019.

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In _ACL 2002_, pages 311—-318. Association for Computational Linguistics, 2002.

[33] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image Transformer. In _PMLR 2018_, volume 80, pages 4055–4064, 2018.

[34] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N. A. Christakis, I. D. Couzin, M. O. Jackson, N. R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D. C. Parkes, A. Pentland, M. E. Roberts, A. Shariff, J. B. Tenenbaum, and M. Wellman. Machine behaviour. _Nature_, 568:477–486, 2019.

[35] G. Ritchie. Some Empirical Criteria for Attributing Creativity to a Computer Program. _Minds and Machines_, 17:67–99, 2007.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. _International Journal of Computer Vision_, 115:211–252, 2015.

[37] J. Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990–2010). _IEEE Transactions on Autonomous Mental Development_, 2(3):230–247, 2010.

[38] R. M. Schmidt, F. Schneider, and P. Hennig. Descending through a Crowded Valley - Benchmarking Deep Learning Optimizers. In _ICML 2021_, volume 139, pages 9367–9376, 2021.

[39] C. E. Shannon. A mathematical theory of communication. _The Bell System Technical Journal_, 27(3):379–423, 1948.

[40] W. R. Tan, C. S. Chan, H. E. Aguirre, and K. Tanaka. Ceci n'est pas une Pipe: A Deep Convolutional Network for Fine-Art Paintings Classification. In _ICIP 2016_, pages 3703–3707, 2016.

[41] H. Toivonen and O. Gross. Data mining and machine learning in computational creativity. _WIREs Data Mining and Knowledge Discovery_, 5(6):265–275, 2015.

[42] D. J. Treffinger. _Creativity, Creative Thinking, and Critical Thinking: In Search of Definitions_. Center for Creative Learning, 1996.

[43] M. Tribus. _Thermodynamics and Thermostatics: An Introduction to Energy, Information and States of Matter, with Engineering Applications_. Van Nostrand, 1961.

[44] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel Recurrent Neural Networks. In _ICML 2016_, page 1747–1756. JMLR.org, 2016.

[45] A. Van Den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu. Conditional Image Generation with PixelCNN Decoders. In _NeurIPS 2020_, page 4797–4805. Curran Associates Inc., 2016.

[46] L. R. Varshney, F. Pinel, K. R. Varshney, D. Bhattacharjya, A. Schoergendorfer, and Y.-M. Chee. A big data approach to computational creativity: The curious case of Chef Watson. _IBM Journal of Research and Development_, 63(1):7:1–7:18, 2019.

[47] L. Weaver and N. Tao. The Optimal Reward Baseline for Gradient-Based Reinforcement Learning. In _UAI 2001_, page 538–545, 2001.

[48] G. A. Wiggins. Searching for Computational Creativity. _New Generation Computing_, 24:209–222, 2006.

[49] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. _Machine Learning_, 8:229–256, 1992.

[50] X. Yi, M. Sun, R. Li, and W. Li. Automatic Poetry Generation with Mutual Reinforcement Learning. In _EMNLP 2018_, pages 3143–3153. Association for Computational Linguistics, 2018.

[51] L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In _AAAI 2017_, page 2852–2858. AAAI Press, 2017.

[52] X. Zhang and M. Lapata. Chinese Poetry Generation with Recurrent Neural Networks. In _EMNLP 2014_, pages 670–680. Association for Computational Linguistics, 2014.

Appendix

A. Euclidean Distance Bounds with Proof

The advantage of using Euclidean Distance is that it is bounded - both an upper and a lower bound can be derived. More formally:

$$\sqrt{\sum_{i=1}^{N} \left(\frac{1}{N} - y_i \right)^2} \geq 0, \tag{8}$$

where

$$\sqrt{\sum_{i=1}^{N} \left(\frac{1}{N} - y_i \right)^2} = 0 \tag{9}$$

with $\quad y_i = \frac{1}{N}, i = 1, .., N,$

and

$$\sqrt{\sum_{i=1}^{N} \left(\frac{1}{N} - y_i \right)^2} \leq \frac{\sqrt{N(N-1)}}{N}, \tag{10}$$

where

$$\sqrt{\sum_{i=1}^{N} \left(\frac{1}{N} - y_i \right)^2} = \frac{\sqrt{N(N-1)}}{N} \tag{11}$$

with $\quad y_i = \begin{cases} 0, & \text{if } i = 1, ..., N, i \neq j \\ 1, & \text{if } i = j. \end{cases}$

The derivation of Equation (9) is trivial: the lower bound of Euclidean Distance is 0. It is reached if and only if the two vectors are equal. In the following we present the derivation of Equations (10) and (11).

The squared sum of the difference can be decomposed as:

$$\sum (x - y)^2 = \sum x^2 + \sum y^2 - 2 \sum xy. \tag{12}$$

Here, x is a constant vector of N values, each of them equal to $\frac{1}{N}$; for this reason,

$$\sum x^2 = \sum_{i=1}^{N} \left(\frac{1}{N} \right)^2 = \left(\frac{1}{N} \right)^2 \sum_{i=1}^{N} 1$$
$$= \frac{1}{N^2} N = \frac{1}{N}. \tag{13}$$

Since x is constant, it is possible to say that:

$$2 \sum xy = 2 \sum_{i=1}^{N} \frac{1}{N} y = \frac{2}{N} \sum_{i=1}^{N} y, \tag{14}$$

but y is a vector of probabilities, therefore $\sum y = 1$, so:

$$2 \sum xy = \frac{2}{N} \sum_{i=1}^{N} y = \frac{2}{N}. \tag{15}$$

Two of the three terms are constant; the last one depends on the variables of y. For a vector y that has the property of $\sum y = 1$, the theoretical maximum of $\sum y^2$ is attained when all its entries are 0 except one, which is 1, obtaining that:

$$\sum y^2 = \sum_{i=1}^{N} y_i^2 \leq 1, \tag{16}$$

where

$$\sum_{i=1}^{N} y_i^2 = 1 \tag{17}$$

with $\quad y_i = \begin{cases} 0, & \text{if } i = 1, ..., N, i \neq j \\ 1, & \text{if } i = j. \end{cases}$

The upper bound of the Euclidean Distance can be found rewriting:

$$\sqrt{\sum (x - y)^2} = \sqrt{\sum x^2 + \sum y^2 - 2 \sum xy}$$

$$= \sqrt{\frac{1}{N} + 1 - \frac{2}{N}} = \sqrt{1 - \frac{1}{N}} \qquad (18)$$

$$= \sqrt{\frac{N - 1}{N}} = \frac{\sqrt{N (N - 1)}}{N},$$

$$\text{with} \quad y_i = \begin{cases} 0, & \text{if } i = 1, ..., N, i \neq j \\ 1, & \text{if } i = j. \end{cases}$$

Equation (4) can be finally obtained by setting $N = \frac{N - N_{min}}{N_{max} - N_{min}}$ and then $N = 1 - N$, which ensures that the measure is bounded between 0 and 1 and satisfies the desired properties listed in Subsection 3.2.

B. Details about Training Process and Implementation

G_v and G_s are LSTM-based RNNs composed of an embedding layer of size 300 (with input length of 20), a LSTM layer with 256 units and a dropout of 0.2 rate, a dense layer with softmax activation [13]. Following [38], Adagrad [9] with a learning rate equal to 0.01 has been used as optimiser. D_v and D_n are CNNs, implemented using an embedding layer of size 300 (with input length equal to the maximum poem length), three convolutional layers of 1 dimension (with tanh activation) with kernel sizes of, respectively, 3, 4 and 5, a max pooling over each output of the last convolutional layer, a dropout of 0.5 rate and finally a dense layer with softmax activation for D_n and sigmoid activation for D_v [22]. In this case, Adam [23] with a learning rate equal to 0.0001 has been adopted as optimiser. Word2Vec [29] has been used for the embedding, pre-trained on Google News, and then fine-tuned for 100 epochs on the specific dataset.

G_v and D_v have been trained following the Seq-GAN algorithm [51] except for the update rule followed, where REINFORCE with Baseline [47] has been used in place of REINFORCE (with only positive rewards, it avoids to remain stuck in poor situations); G_v has been pre-trained for 50 epochs on TCC, and D_v for 5 epochs on batches of 32 outputs of G_v, with $N = 1$ in the computation of the expected return of the Monte Carlo search, with $g_{steps} = 8$ and $d_{steps} = 4$, with a discriminator batch size of 32, and in total 550 epochs (with more, the discriminator becomes overfitted). D_n has been trained for 56 epochs using Categorical Cross-Entropy; G_s has been trained for 136 epochs using Sparse Categorical Cross-Entropy. In both cases, the number of epochs has been found as the one which minimizes the validation loss.

Finally, out-of-vocabulary (OOV) words of test set poems have been treated by substituting each OOV word with the one that has the most similar embedding to the one predicted by a sequence generator [28] - here, G_s.

C. American Poetry Dataset

The training set is composed by 2676 poems from the 19th century, divided into five groups: American Renaissance (Brahmins and Romantics); Local Color; Naturalism; Neogothic (or Protodecadentism). For Brahmins, we included poems written by Henry Wadsworth Longfellow (extracted by *The complete poetical works of Henry Wadsworth Longfellow*), Oliver Wendell Holmes (extracted by *Songs in many keys*), James Russell Lowell (extracted by *Poems*) and John Greenleaf Whittier (extracted by *Poems of nature plus poems subjective and reminiscent and religious poems*, *Anti-slavery poems and songs of labor and reform*, *Personal poems*). For Romantics, we included poems written by Emily Dickinson (extracted by *Poems*), Walt Whitman (extracted by *Leaves of grass*) and Ralph Waldo Emerson (extracted by *Poems*). For Local Color, we included poems written by Bret Harte (extracted by *East and west: poems*), Frances Harper (extracted by *Poems*) and Rose Terry Cooke (extracted by *Poems*). For Naturalism, we included poems written by Stephen Crane (extracted by *The black riders and other lines* and *War is kind*) and Hamlin Garland (extracted by *Prairie songs*). For the last group, we included poems written by Edgar Allan Poe (extracted by *The complete poetical works of Edgar Allan Poe*). Since all of these works are in public domain, we downloaded them from Project Gutenberg[2] without any risk of copyright infringement.

The first test set, instead, is divided into five poetical movements of the 20th century. These movements are temporally consecutive, and they model a sort of timeline of American poetry. Of course, many other movements could be considered beside the five presented here, but they overlap the chosen five in time, making it difficult to interpret the results; in addition, no representativeness has been lost in our opinion, since the goal is not to retrace every single historically relevant moment in American poetry. Instead, the goal is to check if the defined measure is able to capture the concept of creativity for a certain period of time. We took into consideration a timeline covering the following movements: Imagism; Harlem Renaissance; Objectivism; Beat Generation; and Confessional Movement. For each movement, 23 poems were considered. For Imagism, the poems are *Autumn*, *The embankment*, *Above the dock*, *Conversion* (by T.E.

Hulme); *Comraderie*, *Piazza San Marco*, *Ballad for gloom*, *A song of the virgin mother*, *Grace before song* (by Ezra Pound); *Leda*, *Pursuit*, *Gift*, *The shepherd*, *All mountains* (by Hilda Doolittle); *Sex and trust*, *Furniture*, *Censors*, *After school*, *Autumn sunshine* (by D.H. Lawrence); *After all*, *A night piece*, *Modern love*, *The feather* (by Ford Madox Ford). For Harlem Renaissance, the poems are *Living earth*, *Skyline*, *Words for a hymn to the sun*, *Banking coal*, *Men*, *Peers* (by Jean Toomer); *God*, *I look at the world*, *Madam and the movies*, *Silly animals*, *Blues fantasy*, *God to hungry child* (by Langston Hughes); *The expulsion of hagar*, *In memory of Arthur Clement Williams*, *An offering*, *Judith*, *Ode to the sun*, *Early spring* (by Eloise A. Bibb); *Idolatry*, *My heart has known its winter*, *A note of humility*, *Nocturne of the wharves*, *Miracles* (by Arna Wendell Bontemps). For Objectivism, the poems are *Brilliant sad sun*, *The aftermath*, *The dish of fruit*, *Spring*, *Flowers by the sea*, *A goodnight* (by William Carlos Williams); *A clerk tiptoeing the office floor*, *Death of an insect*, *Episode in Iceland*, *The doctor's wife*, *Lesson of job*, *Hardly a breath of wind* (by Charles Reznikoff); *Ode to the commonplace*, *Aubade*, *In the first circle of limbo*, *Museum*, *Eye to eye*, *Ode on arrival* (by Carl Rakosi); *I'm the worse for drinking again*, *Poor soul! softly, whisperer*, *Gin the goodwife stint*, *Darling of gods and men*, *Isn't it poetical a chap's mind?* (by Basil Bunting). For the Beat Generation movement, the poems are *Homeless compleynt*, *Ego confession*, *G.S. reading poesy at Princeton* (by Allen Ginsberg); *The leaves danced to Mozart*, *Roma*, *I saw great neptune*, *So much depends upon*, *I saw two lovers* (by Lawrence Ferlinghetti); *Doctors will be protected*, *Vows*, *Out west*, *Why California will never be like Tuscany*, *Lodgepole* (by Gary Snyder); *Storm at low tide*, *The gash*, *No sound*, *Good morning*, *The sacred distillate* (by William Everson); *Getting to the poem*, *Inter and outer rhyme*, *Daydream*, *Destiny*, *Sunrise* (by Gregory Corso). For the Confessional Movement, the poems are *Stopped dead*, *The night dances*, *A secret*, *Cut*, *Amnesiac*, *You're*, *The moon and the yew tree* (by Sylvia Plath); *Buying the whore*, *The red dance*, *Old*, *Not so, not so*, *The fury of jewels and coal*, *The fallen angels*, *June bug* (by Anne Sexton); *The moth chorale*, *Parents*, *Phone message*, *Lasting*, *Cherry saplings*, *Leavings* (by W.D. Snodgrass); *The cage*, *Letter to his brother*, *Rock-study with wanderer* (by John Berryman). Apart from T.E. Hulme's poems, which are in the public domain, the others are retrieved through ProQuest's Literature Online database, by means of the license agreement of the University of

---

Bologna. The poems were copied only for the amount of time required for their use in the experiments, and then deleted; in this way, their usage is in compliance with the current legislation, as explained in [27].

The second test set is divided into four periods: Puritanism (Colonial age), in the 17th century; Metaphisical Poetry, at the beginning of the 18th century; the birth of African-American Poetry, at the half of the 18th century; and the American Revolution, at the end of the 18th century. For each period, one poet is considered: for the first one, Anne Bradstreet with the poems *In memory of my dear grandchild Elizabeth Bradstreet*, *In thankful remembrance*, *For deliverance from a feaver*, *Davids lamentation for Saul and Jonathan*, *What God is like to him I serve*, *In my solitary hours in my dear husband his absence*, *An apology*, *Upon some distemper of body*, *Spirit*, *The vanity of all wordly things*, *Deliverance from another sore fit*, *Meditations divine and moral*, *Upon a fit of sickness*, *My thankfull heart with glorying tongue*, *As spring the winter doth succeed*, *The author to her book*, *As weary pilgrim, now at rest*, *Another*, *To my dear children*, *The four elements*, *Here follows some verses upon the burning of our house*, *The flesh and the spirit*, *We may live together*; for the second one, Edward Taylor, with the poems *1*, *6*, *29*, *32*, *38* and *39* from *Preparatory Meditations - First Series*, *7*, *12*, *62*, *143* and *146* from *Preparatory Meditations - Second Series*, *The experience*, *The Souls Groan to Christ for Succour*, *Upon wedlock, and death of children*, *Head of a white woman winking*, *The wrong way home*, *Upon a wasp chilled with cold*, *Ebb and flow*, *The joy of church fellowship rightly attended*, *Huswifery*, *Upon a spider catching a fly*, *The souls admiration hereupon*, *The souls address to Christ against these assaults*; for the third one, Phillis Wheatley, with the poems *On virtue*, *To the University of Cambridge, in New-England*, *On the death of a young lady of five years of age*, *On the death of a young gentleman*, *To a lady on the death of her husband*, *Thoughts on the works of Providence*, *To a lady on the death of three relations*, *To a clergyman on the death of his lady*, *An hymn to the morning*, *An hymn to the evening*, *On recollection*, *On imagination*, *To a lady on her coming to North-America with her son, for the recovery of her health*, *To a lady on her remarkable preservation in an hurricane in North-Carolina*, *To a lady and her children, on the death of her son and their brother*, *On the death of J. C. an infant*, *To S. M. a young African painter, on seeing his works*, *A farewell to America*, *On the death of Dr. Samuel Marshall*, *To a gentleman and lady on the death of the lady's brother and sister, and a child of the name of Avis, aged one year*, *A funeral poem on the death of C. E. an infant of twelve months*, *On the death of the Rev. Dr. Sewell*, *On being brought from Africa to America*; for the fourth one, Philip Freneau, with the poems *A New-York tory*, *To lord Cornwallis*, *The vanity of existence*, *To the memory of the brave americans*, *The royal adventurer*, *A speech - that should have been spoken by the King of the Island of Britain to his Parliament*, *Lines - occasioned by Mr. Rivington's new titular types to his Royal Gazette*, *A prophecy*, *The argonaut - or, lost adventurer*, *Barney's invitation*, *Sir Guy Carleton's address to the americans*, *Scandinavian war song*, *The projectors*, *A picture of the times*, *Satan's remonstrance*, *The refugee's petition to Sir Guy Carleton*, *To a concealed royalist*, *The prophecy of king Tammany*, *Stanzas - occasioned by the departure of the British from Charleston*, *On the british king's speech*, *Manhattan city*, *A news-man's address*, *The happy prospect*. Since all of these works are in public domain without risks of copyright infringement, they were downloaded from Project Gutenberg.