CAEMSI : A Cross-Domain Analytic Evaluation Methodology for Style Imitation

Jeff Ens, Philippe Pasquier
School of Interactive Arts and Sciences
Simon Fraser University
{jeffe,pasquier}@sfu.ca

Abstract

We propose CAEMSI, a cross-domain analytic evaluation methodology for Style Imitation (SI) systems, based on a set of statistical significance tests that allow hypotheses comparing two corpora to be tested. Typically, SI systems are evaluated using human participants, however, this type of approach has several weaknesses. For humans to provide reliable assessments of an SI system, they must possess a sufficient degree of domain knowledge, which can place significant limitations on the pool of participants. Furthermore, both human bias against computer-generated artifacts, and the variability of participants' assessments call the reliability of the results into question. Most importantly, the use of human participants places limitations on the number of generated artifacts and SI systems which can be feasibly evaluated. Directly motivated by these shortcomings, CAEMSI provides a robust and scalable approach to the evaluation problem. Normalized Compression Distance, a domain-independent distance metric, is used to measure the distance between individual artifacts within a corpus. The difference between corpora is measured using test statistics derived from these inter-artifact distances, and permutation testing is used to determine the significance of the difference. We provide empirical evidence validating the statistical significance tests, using datasets from two distinct domains.

Introduction

There is growing demand for creative generative systems in the entertainment industry, which has prompted an abundance of research in the area of Style Imitation (SI). Given a corpus $C = \{c_1, ..., c_n\}$, SI systems aim to generate new artifacts that emulate the stylistic characteristics of C. Many of these SI systems generate some form of musical content, including; harmonic progressions, melodies (Yang, Chou, and Yang 2017), and polyphonic compositions (Liang et al. 2017). A more comprehensive overview of the work in the domain can be found elsewhere (Pasquier et al. 2017; Briot and Pachet 2017). In the visual art domain, the Creative Adversarial Network (CAN) is trained to generate visual art that deviates from the styles it has already learned (Elgammal et al. 2017). Moreover, many Natural Language Generation (NLG) systems have been developed that generate jokes, poetry, and narratives in a particular style (Gatt and Krahmer 2017). To accommodate the large influx of generative systems in recent years, we propose CAEMSI [1].

Ritchie mentions two conditions for determining if creativity has occurred: *novelty*, the degree to which an artifact is dissimilar to other examples within the corpus and *quality* (Ritchie 2007). He also emphasizes the notion of *typicality*, the degree to which a generated artifact is representative of the source corpus (C). In the context of style imitation, measuring typicality is of critical importance, as the performance of an SI system hinges on its ability to emulate the stylistic characteristics of the source corpus. As a result, CAEMSI focuses on measuring the typicality of a generated corpus, with respect to the source corpus. Although novelty is also an important indicator of the system's quality, as it is generally undesireable for an SI system to plagiarize large sections from the source corpus, we leave this aspect of evaluation for future work.

Traditionally, participants assess the capacity of a particular system to emulate a particular style, allowing researchers to make claims about the success of that system. Unfortunately, this is not a scalable solution, and can make it difficult to compare SI systems. With the long-term goal of creating highly capable SI systems, it is necessary to develop robust methods for the evaluation of these systems, as a lack of methodical evaluation can have a negative effect on research progress (Pearce, Meredith, and Wiggins 2002). The approach described in this paper is domain independent, harnessing the power of Normalized Compression Distance (NCD) (Cilibrasi and Vitányi 2005) and permutation testing to provide a scalable solution to the problem of SI system evaluation. In order to demonstrate the effectiveness of this approach, we conduct experiments on datasets in two different domains; the Wikiart image dataset[2] and the Classical Archives MIDI dataset[3].

Background

Evaluation Methodologies

Although many methodologies that evaluate the creative capacity of a generative system have been proposed, we will limit our discussion to those which have been used to measure typicality. In general, we can divide these methodolo-

[1] The code is available https://goo.gl/ejN1RM
[2] https://www.wikiart.org/
[3] https://www.classicalarchives.com/midi.html

gies into two categories, those which rely on human participants, and those based purely on computation. Unto our knowledge, the only statistical evaluation methodology for typicality was proposed by Gonzalez Thomas et al., however it is only capable of evaluating melodic composition systems (Gonzalez Thomas et al. 2013).

The Consensual Assessment Technique (CAT) (Amabile 1982) is based on the notion that experts are the most capable of distinguishing creative artifacts within their respective domain. To account for discrepancies, which arise given the subjective nature of these assessments, the CAT averages the assessments of several experts. Pearce and Wiggins employ the CAT to evaluate the success of melodic generation algorithms (Pearce and Wiggins 2007).

Another approach, inspired by the Turing-test, measures participants ability to discriminate between computer-generated artifacts and artifacts from the source corpus. This evaluation methodology has been used to evaluate many SI systems, including a Deep LSTM Network that generates Bach chorales (Liang et al. 2017), and a Generative Adversarial Network that generates images (Elgammal et al. 2017).

Related Work

To the best of the authors' knowledge, there are no domain independent metrics for typicality, however, several quantitative metrics for creative systems have been proposed. Maher has proposed two metrics for measuring creativity quantitatively. The first, equates *novelty* with distance from predominant clusters of artifacts, measures *surprise* using pattern matching algorithms, and calculates *value* using a fitness function (2010). However, it is not clear how the proposed metrics would be applied to an arbitrary domain, and no proof of concept is provided. The second, uses Bayesian inference to measure the novelty of an artifact, which is used to evaluate potential designs for laptop computers (Maher and Fisher 2012).

Burns measures creativity as the combination of psychological arousal, which is computed using Shannon entropy, and appraisal, which is computed using Bayesian theory (Burns 2015). The Regent-Dependent Creativity (RDC) metric measures value and novelty. Artifacts are represented by a set of pairs $(P(\textit{regent}, \textit{dependent}))$, where *regent* is an action or attribute, and *dependent* is a state or target for an action (Rocha, Ribeiro, and El 2016). Using a graph, which includes associations between artifacts, they propose metrics to measure synergy, the value produced by various elements acting cooperatively, and Bayesian surprise, the degree to which an artifact is unexpected or novel. Although this metric seems to work well for the low dimensional problems presented in the paper, it is not clear that this approach could efficiently handle artifacts which require a large number of pairs for representation. Furthermore, it relies on the domain knowledge of synergy, which is difficult to determine in some domains.

Motivation

Although human-based evaluation methodologies are not without their strengths, the shortcomings of these methodologies directly motivated the development of the statistical tests proposed in this paper.

Domain Knowledge

Accurately assessing the typicality of an artifact with respect to a source corpus, requires a significant amount of domain knowledge, as the participant must be familiar with the stylistic characteristics of the source corpus. This issue is exacerbated when performing a CAT, since participants must have an expert level knowledge of the source corpus. Undoubtedly, this is one of the primary reasons an abundance of musical SI systems have focused on imitating Bach chorales, as there is a large pool of experts, and most people are familiar with Bach's work. Since a lack of domain knowledge undermines the reliability of the evaluation process, the types of scientific inquiries which have been explored are biased by restrictions on the source corpora, placing limitations on scientific progress in this area.

Bias Against Generative Systems

Previous research has shown that when participants were asked to distinguish between two folk melodies, some of which were human-composed and others which were recombinations of the human-composed melodies, participants attributed unusual or disagreeable human compositions to the computer (Dahlig and Schaffrath 1997). Norton, Heath, and Ventura found a significant bias against images labeled as being generated by a computer (2015). In contrast, several studies have demonstrated that the knowledge that a computer created a piece of music, does not significantly affect the participants' evaluation and enjoyment of the piece (Moffat and Kelly 2006; Friedman and Taylor 2014; Pasquier et al. 2016). Although Moffat and Kelly's study did not explicity test the same hypothesis as Dahlig and Schaffrath, their results corroborate the same conclusion, as participants attributed compositions they disliked to the computer, independent of their actual authorship.

When participants are tasked with making the distinction between human-generated and computer-generated artworks, they may in fact be searching for features which they expect to be generated by a computer, rather than focusing on the broader style of the composition (Ariza 2009). As a result, the test degenerates to one which is focused on counting perceived mistakes. This issue has been highlighted by Pearce in his discussion on the evaluation of musical composition systems (2005). Clearly, this type of bias is very problematic when attempting to evaluate an SI system that imitates artifacts that humans tend to find disagreeable, such as the atonal works of Arnold Schoenberg.

Variability

The subjective nature of creativity-based assessments poses problems for the systematic evaluation of creative systems in general. There is evidence that cultural background can have an effect on how an artifact is perceived. For example, Eerola et al. found that western and African listeners perceived musical attributes differently (2006). Furthermore, environmental factors will affect the reliability of these assessments, including the equipment used to observe the artifact, and the physical condition of the participant. Although those who design experiments take many steps to mitigate the effects of these factors, Schedl at al. (Schedl, Flexer, and Urbano 2013) provide evidence that inter-rater agreement is still limited in a practical setting. In one case, non-experts'

assessments of poetry were found to be negatively correlated with the assessments of experts (Lamb, Brown, and Clarke 2015). Similarly, Kaufman, Baer, and Cole found that experts were far more reliable than non-experts, when asked to judge the creativity of a short story, as measured by inter-rater reliability for both groups (2009).

Scalability

Unfortunately, using human participants places limitations on the total number of assessments that can be collected. Participants are only capable of making so many assessments before fatigue will begin to degrade the quality of their responses. Notably, this problem is exacerbated by the limited number of participants involved when conducting a *CAT*. Although crowdsourcing does make it easier to collect a large number of assessments, there are still monetary and time limitations that place restrictions on the the total number of assessments that can be feasibly collected. Clearly, the limited scalability of these evaluation methods is in direct conflict with the large number of artifacts which generative systems can produce.

In many cases, a small subset of the generated artifacts is used to evaluate the system, decreasing the number of assessments required. However, issues will naturally arise when the selected subset is not adequately representative of the system's output as a whole (Ariza 2009). Moreover, it is not trivial to determine if a subset of artifacts is representative of the systems output a priori. Most importantly, these limitations make it increasingly difficult to evaluate a large number of systems.

The Proposed Solution

In contrast to human-based evaluation methods, CAEMSI eschews the issues of domain knowledge, human bias, and variability. Admittedly, there are still limitations with respect to the size of corpora, which will be addressed in future work. However, computation based methods of evaluation are far more scalable than human-based solutions, as computers can process artifacts much faster than humans can.

Statistical Tests for Typicality

In what follows, $X = [x_i, i = 1, ..., n]$ denotes a vector X, containing n elements. $X \oplus Y$ denotes the concatenation of two vectors. We use the term *corpora* to denote a vector of binary strings. $\mu(X)$ denotes the mean of a vector X, while $\phi(X)$ denotes the median. p_{diff} and p_{eqv} denote the significance of the statistical test for difference and equivalence respectively.

Given two corpora, $A = [a_i, i = 1, ..., n]$ and $B = [b_i, i = 1, ..., m]$, we test the null hypothesis $H_{D0} : A = B$ ($p_{\text{diff}} > \alpha$) against $H_{D1} : A \neq B$ ($p_{\text{diff}} \leq \alpha$) and the null hypothesis $H_{E0} : A \neq B$ ($p_{\text{eqv}} > \alpha$) against $H_{E1} : A = B$ ($p_{\text{eqv}} \leq \alpha$). When the result of a statistical test is insignificant, we accept the null hypothesis, which only indicates that there was insufficient evidence to support the alternate hypothesis, and does not validate or invalidate the null hypothesis. As a result, accepting the null hypothesis $H_{D0} : A = B$ is not the same as rejecting the null hypothesis $H_{E0} : A \neq B$ and accepting the alternative hypothesis $H_{E1} : A = B$, as only the latter indicates that $A = B$. Consequently, we can determine if $A = B$ using p_{eqv} and if $A \neq B$ using p_{diff}.

Normalized Compression Distance

Put simply, the *Kolmogorov complexity* ($K(x)$) of a finite length binary string x is the minimum number of bits required to store x without any loss of information. More formally, $K(x)$ denotes the length of the shortest Universal Turing Machine that prints x and stops (Solomonoff 1964). Intuitively, the minimum number of bits required to store a random string would be close to the number of bits used to represent the original string. As a result, a random string would have a high Kolmogorov complexity. In contrast, a string with a large number of repeated subsequences, would have a low Kolmogorov complexity. Although Kolmogorov complexity provides an absolute lower bound on the compression of a string, $K(x)$ is non-computable (Li et al. 2004), so a real-world compressor is used to approximate $K(x)$ in practice.

The *conditional Kolmogorov complexity* ($K(x|y)$) of a string x relative to a string y, denotes the length of the shortest program that prints x and stops, with y provided as additional input to the computational process. For example, if $x \simeq y$, $K(x|y)$ would be very small, as the program could reproduce x from y without requiring much additional information. In contrast, if x and y are highly dissimilar, $K(x|y)$ would be quite large.

Information distance is the length of the shortest binary program that can compute x from y and y from x. As a result, when x and y have a lot of mutual information, the length of this program will be fairly short. Li et al. propose the *normalized information distance* (1).

$$d(x, y) = \frac{\max(K(x|y), K(y|x))}{\max(K(x), K(y))} \tag{1}$$

Since $K(x|y) \simeq K(xy) - K(x)$ (Li et al. 2004), where xy denotes the concatenation of strings x and y, we can reformulate (1) to arrive at a computable *normalized compression distance* (NCD) (2). In practice, $K(x)$ is the length of string produced by a real-world compression algorithm, such as zlib. Although we tested several compression algorithms, we did not notice significant variation in terms of performance.

$$D(x, y) = \frac{K(xy) - \min(K(x), K(y))}{\max(K(x), K(y))} \tag{2}$$

Li et al. demonstrate that *NCD* is a universal distance metric, satisfying the following constraints.

1. $D(x, y) = 0$ *iff* $x = y$ (Identity)
2. $D(x, y) + D(y, z) \geq D(x, z)$ (Triangle Equality)
3. $D(x, y) = D(y, x)$ (Symmetry)

Notably, *NCD* has been applied to problems in a variety of domains, including music classification (Cilibrasi, Vitányi, and De Wolf 2004; Li and Sleep 2005), protein sequence classification (Kocsor et al. 2006), image registration (Bardera et al. 2010), and document classification (Axelsson 2010). Väyrynen and Tapiovaara use *NCD* to evaluate *machine translation* (MT) by measuring the distance between the predicted translation and the ground truth translation (Väyrynen and Tapiovaara 2010).

Distance Matrix Construction

Given a valid distance metric D and two corpora ($A = [a_i, i = 1, ..., n]$ and $B = [b_i, i = 1, ..., m]$), we can construct a pairwise distance matrix M, where $M_{ij} = D(c_i, c_j)$, and $C = A \oplus B = [c_i, i = 1, ..., n+m]$. We use several subsets of M to perform the proposed statistical tests. In the formula below, w_A and w_B are vectors containing all distinct within group distances for corpora A, and B respectively, while $b_{A,B}$ contains all between group distances. Notably, $l = n + m$ in the equations below.

$$w_A = [M_{ij}, i = 1, ..., n; j = 1, ..., n; j > i] \quad (3)$$

$$w_B = [M_{ij}, i = n+1, ..., l; j = n+1, ..., l; j > i] \quad (4)$$

$$b_{A,B} = [M_{ij}, i = 1, ..., n; j = n+1, ..., l] \quad (5)$$

Permutation Testing

A *permutation test* is a statistical significance test which requires no prior knowledge about the distribution of the test statistic under the null hypothesis, as this distribution is generated by calculating the test statistic for each possible labelling of the data. For example, consider the vector $C = A \oplus B$, which is comprised of two corpora delineated by the labels $\mathbf{L} = [l_i, i = 1, ..., n + m; l_{i \leq n} = 0, l_{i > n} = 1]$, and a test statistic $S = \mu(C_0) - \mu(C_1)$, where $C_j = \{c_i \mid l_i = j\}$. First, compute S using $\mathbf{L}$. Then compute S for each possible permutation of $\mathbf{L}$ to construct the distribution under the null hypothesis. Since the number of permutations grows exponentially, as comparing two corpora of size 50 would require $\binom{100}{50} \simeq 10^{29}$ distinct permutations, we approximate this procedure by randomly selecting m permutations. This procedure accommodates complex test statistics, for which it would be intractable, or overly difficult, to compute the distribution of the test statistic under the null hypothesis.

Testing for Difference

To test the hypothesis that two corpora are different, we adapt a permutation testing framework that was used to compare two groups of brain networks (Simpson et al. 2013). Simpson et al. create a pairwise distance matrix M using the Kolmogorov-Smirnov statistic, however, we use *NCD* instead.

$$R(M) = \frac{\mu(b_{A,B})}{\mu(w_A \oplus w_B)} \quad (6)$$

When R is greater than 1, the average between group distance is greater than the within group distance. Therefore, $R > 1$ suggests that the two corpora are likely distinct. In contrast, when $R \simeq 1$, there is likely no difference between the two corpora. The proposed test is detailed in the steps below, where $\mathbf{I}(\cdot) = 1$ if $(\cdot)$ is true and 0 otherwise.

1. Given two corpora $A = [a_i, i = 1, ..., n]$ and $B = [b_i, i = 1, ..., m]$, create a pairwise distance matrix M using (2).
2. Calculate the test statistic $T = R(M)$ using (6).
3. Take a random permutation (u^*) of the ordering $u = (1, ..., n+m)$ and reorder the columns and rows using this ordering to create M^*.
4. Calculate the test statistic $T^* = R(M^*)$ using (6).
5. Repeat steps 3 and 4 N times, producing the output $[T^*_n, n = 1, ..., N]$.
6. Calculate the p-value, $p_{\text{diff}} = \sum_{n=1}^{N} \mathbf{I}(T^*_n \geq T)/N$.

Testing for Equivalence

The proposed test for the equivalence of two corpora, is based on the following assumption.

$$(w_A = b_{A,B}) \wedge (w_B = b_{A,B}) \implies A = B \quad (7)$$

The intuition behind this assumption is shown in Figure 1 and 2, which show the cumulative distributions of w_A, w_B, and $b_{A,B}$ for an intra-artist comparison and an inter-artist comparison respectively. When two distinct corpora are compared, $b_{A,B} \neq w_A$ and $b_{A,B} \neq w_B$, as shown in Figure 1. In contrast, when two similar corpora are compared, $b_{A,B} \simeq w_A \simeq w_B$, as shown in Figure 2. In practice, the distributions of w_A, w_B, and $b_{A,B}$ are frequently skewed, and sometimes multi-modal, which necessitates a non-parametric test for equivalence.

As a result, we employ a permutation testing framework (Pesarin et al. 2016), which is based on Roy's Union-Intersection approach (1953), to test for the equivalence of two distributions. First, it is necessary to define an equivalence interval on which the two distributions will be considered equal. ε_I and ε_S denote the inferior and superior margins, respectively. Then we test two hypotheses; $H_{I0} : \delta \geq -\varepsilon_I$ against $H_{I1} : \delta < -\varepsilon_I$ and $H_{S0} : \delta \leq \varepsilon_S$ against $H_{S1} : \delta > \varepsilon_S$, where δ is the divergence between the two distributions being compared. In some cases, this is measured as the difference between the means (μ), however we use the difference between the medians (ϕ), as it is more robust to outliers. As a result, the global null hypothesis (H_{E0}) is true if both one-sided null hypotheses (H_{I0}, H_{S0}) are true, and the global alternative hypothesis (H_{E1}) is true if at least one of H_{I1} and H_{S1} is true. The following algorithm is used to test for the equivalence of two distributions.

1. Given two vectors $F = [f_i, i = 1, ..., n]$ and $G = [g_i, i = 1, ..., m]$, compute the rank transform of $F \oplus G$ to derive a rank transformed version F and G.
2. Given the superior and inferior equivalence margins $(\varepsilon_I, \varepsilon_S)$, we create two vectors $X_I = F \oplus (G + \varepsilon_I)$ and $X_S = F \oplus (G - \varepsilon_S)$, and an ordering $u = (1, ..., n+m)$.
3. Compute the test statistic for both hypothesis $T_I = \phi(X_{IF}) - \phi(X_{IG})$ and $T_S = \phi(X_{SG}) - \phi(X_{SF})$ where

$$X_{IF} = [X_I(u_i), i = 1, ..., n]$$
$$X_{IG} = [X_I(u_i), i = n+1, ..., n+m]$$
$$X_{SF} = [X_S(u_i), i = 1, ..., n]$$
$$X_{SG} = [X_S(u_i), i = n+1, ..., n+m]$$

   and $X(j)$ denotes the j_{th} element in X.
4. Take a random permutation (u^*) of the ordering u.
5. Compute the test statistics using the ordering u^*. $T^*_I = \phi(X_{IF}) - \phi(X_{IG})$ and $T^*_S = \phi(X_{SG}) - \phi(X_{SF})$.
6. Repeat steps 3 and 4 N times to simulate the distribution of the two partial test statistics, producing the output $[(T^*_{In}, T^*_{Sn}), n = 1, ..., N]$.

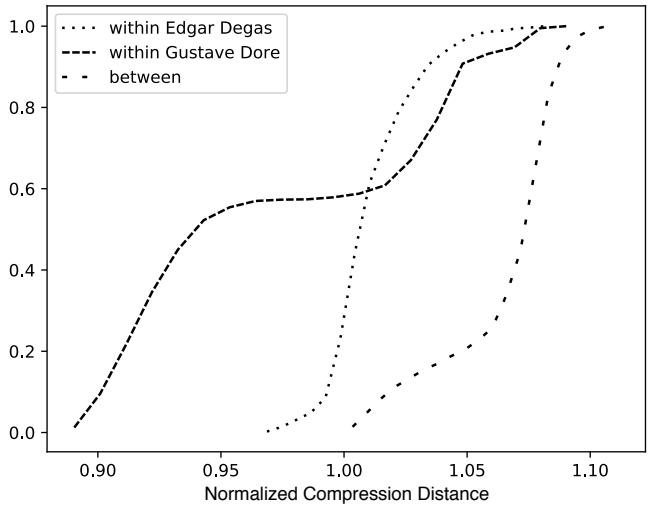

Figure 1: The cumulative *NCD* distributions (w_A, w_B, and $b_{A,B}$) used to compare 50 of Edgar Degas' (A) artworks and 50 of Gustave Dor's (B) artworks.

7. Compute the two partial test statistics $\lambda_h = \sum_{n=1}^{N} \mathbf{I}(T^*_{hn} \geq T_h)/N$ for $h = I, S$. Then the global test statistic is $\lambda(F, G) = \max(1 - \lambda_I, 1 - \lambda_S)$.

To test for the equivalence of two corpora, we compute the distance matrix M using *NCD*, then we compute (8). As a result, if both $\lambda(w_A, b_{A,B})$ and $\lambda(w_B, b_{A,B})$ are significant, then we consider the two corpora equivalent.

$$p_{\text{eqv}} = \max(\lambda(w_A, b_{A,B}), \lambda(w_B, b_{A,B})) \tag{8}$$

Experiment

Methodology

To evaluate the proposed statistical tests, we use datasets from two different domains; the classical archives MIDI dataset, which consists of 14,724 compositions by 843 distinct composers, and the Wikiart dataset, which consists of 19,052 paintings by 23 artists. There are two conditions, one where both corpora (A, B) have the same class (they are created by the same composer or artist), and another where the corpora have a different class. Therefore, the ground truth is calculated using (9), and the condition predicted by each statistical test is calculated using (10), with the standard significance level ($\alpha = 0.05$). To create corpora of different sizes, we randomly select artifacts without replacement belonging to the same class.

$$g(a, b) = \begin{cases} 0, & \text{if class}(a) \neq \text{class}(b) \\ 1, & \text{else} \end{cases} \tag{9}$$

$$\hat{g}(a, b) = \begin{cases} 0, & \text{if } p_{\text{eqv}} \geq \alpha \text{ or } p_{\text{diff}} < \alpha \\ 1, & \text{if } p_{\text{eqv}} < \alpha \text{ or } p_{\text{diff}} \geq \alpha \end{cases} \tag{10}$$

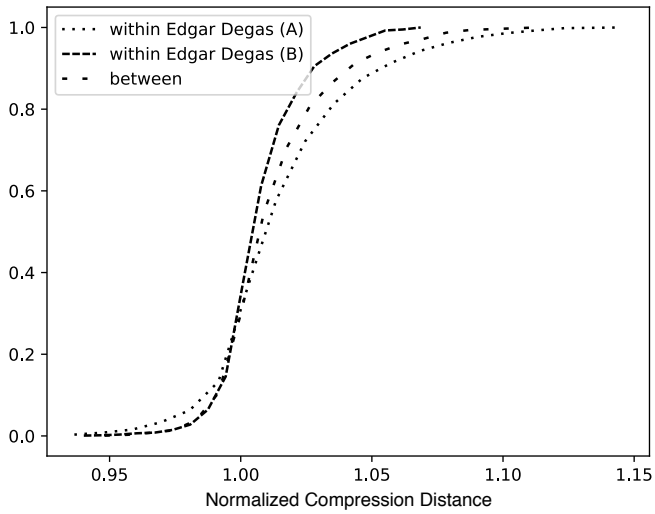

Figure 2: The cumulative *NCD* distributions (w_A, w_B, and $b_{A,B}$) used to compare two disjoint subsets of Edgar Degas' artwork, both of size 50.

Preprocessing Step	0	1	2
Wikiart	19052	19052	18874
Classical MIDI Archives	14724	12117	11943

Table 1: The corpus size after each preprocessing step

Data Pre-Processing

Since our test statistic takes the pairwise distance of all items within a corpora into consideration, having a number of duplicate items would artificially decrease values in w_A and w_B. As a result, we took the following steps to remove duplicate items in each dataset.

1. Remove all artifacts which belong to the same class and have the same title.
2. Remove all artifacts which belong to the same class and have a similarity greater than a threshold (t_{sim}).

We measure the similarity between two images using the structural similarity index (Wang et al. 2004), which takes structural information into account, rather than quantifying visible differences. To measure the similarity of two MIDI files, we extract a list of the pitches in the MIDI file ordered by onset time. Given compute time constraints, we only take the first 1000 notes into consideration. The following equation is used to quantify similarity, where $E(a, b)$ denotes the edit distance between two pitch sequences.

$$s = 1 - \frac{E(a, b)}{1000} \tag{11}$$

We set the similarity threshold (t_{sim}) at 0.75. Although this is quite conservative, we found that this did not eliminate too many artifacts, while providing confidence that duplicate artifacts are not included in the dataset. Table 1 lists the size of each corpus after each preprocessing step.

Data Representation

In order to avoid taking metadata, such as the title, composer, and author into consideration when computing the *NCD*, we do not use a binary representation of the MIDI files. Instead we create a representation which excludes irrelevant data. Since the velocity of MIDI note onsets is primarily based on the performer's interpretation of the composition, and in some cases may be set to a constant value if the MIDI file was created in a notation editor, we ignore this information. As a result, we represent a MIDI file as a sequence of onsets, offsets and time deltas. We represent onsets on the range $[0, 127]$, offsets on the range $[128, 255]$, and time deltas on the range $[256-)$. This results in a sequence of integers, which is then converted to a binary string before measuring the *NCD*. The representation used for images is much simpler. Each image is resized to have the shape 64×64, with three color channels (RGB), where each pixel is represented as an integer on the range $[0, 255]$.

Results

In Table 2 we present the results of 1000 trials, half of which have a ground truth of 0, and half which have a ground truth of 1, for a variety of corpora sizes. The *accuracy* (ACC), *true positive rate* (TPR), *false positive rate* (FPR), *true negative rate* (TNR), and *false negative rate* (FNR), are reported, using the formulas shown below, where n is the number of trials. *True positive* indicates trials in which the statistical test predicts 1 and the ground truth is also 1 ($\hat{g}(a,b) = 1 \wedge g(a,b) = 1$). Similarly, *true negative* indicates trials in which the statistical test predicts 0 and the ground truth is also 0 ($\hat{g}(a,b) = 0 \wedge g(a,b) = 0$). $\varepsilon = \varepsilon_I = \varepsilon_S$ denotes the equivalence range, which is normalized with respect to the length of $F = [f_i, i = 1, ..., n]$ and $G = [g_i, i = 1, ..., m]$ in Table 2. For example, if $\varepsilon = 0.1$ denotes an equivalence range of $(m + n) * 0.1$.

$$ACC = \frac{\sum \text{True positive} + \sum \text{True negative}}{n} \tag{12}$$

$$TPR = \frac{2 \sum \text{True positive}}{n} \tag{13}$$

$$TNR = \frac{2 \sum \text{True negative}}{n} \tag{14}$$

$$PPV = \frac{\sum \text{True positive}}{\sum \text{Predicted positive}} \tag{15}$$

$$NPV = \frac{\sum \text{True positive}}{\sum \text{Predicted negative}} \tag{16}$$

A robust statistical test, will minimize the probability of type I error (α), incorrectly rejecting a true null hypothesis, and type II error (β), incorrectly rejecting a true alternative hypothesis. The power of a statistical test is $1 - \beta$, which is equivalent to the *TNR* with respect to the test for difference (p_{diff}), and the *TPR* with respect to the test for equivalence (p_{eqv}). Since we also must verify that the tests minimize type I error, we provide the *TPR* and *TNR* which are equivalent to statistical sensitivity, for p_{diff} and p_{eqv} respectively. For each trial, we perform 1000 permutations, as this is what Marozzi suggests when estimating the power of a permutation test (2004).

Discussion

Given the degree of intra-corpus variation, and inter-corpus similarity, it is difficult to establish a ground truth for corpus comparison. In many cases, an artist or composer may explore several different sub-styles over the span of their career. Furthermore, artists and composers are often inspired by their colleagues, creating works that exhibit a greater than average degree of similarity. As a result, it would be unreasonable to expect extremely high values of accuracy. Nevertheless, according to Cohen, 0.80 is an adequate level for statistical power (1988), which most of the tests surpass. Overall, the results of the experiment demonstrate that the proposed tests provide a robust measurement of the stylistic difference between two corpora.

We used different values for ε to account for the decrease in variability of w_A, w_B and $b_{A,B}$ as the size of the corpora increases. For example, if two paintings are randomly selected from the work of a single artist, in some cases, given the variability of that artist's work, the mutual information between these two paintings will be fairly low. In other cases, when both paintings are part of the same sub-style, the mutual information may be fairly high. However, as we increase the number of paintings selected, stylistic tendencies will start to emerge, and the amount of mutual information amongst the selected paintings will converge. As a result, w_A and w_B will decrease in variability as the size of the corpora is increased, which allows us to decrease the size of the equivalence interval by decreasing ε.

The results in Table 2 show two trends. On average the statistical tests performed better on MIDI than on images. There are two possible explanations for this; composers may have a more consistent style than artists, or the representation we used for images is not optimized for comparison. However, the fact that images were not preprocessed, as we simply resized each image and extracted the raw pixel values, demonstrates that *NCD* is capable of finding commonalities in the raw data. Secondly, the statistical tests perform better on larger corpora than smaller corpora, which is primarily the result of decreasing stylistic variability as the size increases.

Since the strings that are being compared were quite long, the *NCD* between two items was heavily skewed towards 1, as shown in Figure 1 and 2. Consequently, we do not suggest interpreting these values as interval, but rather as ordinal values. Despite the skew of these values, discrepancies between w_A, w_B, and $b_{A,B}$ can be quite pronounced.

Application

There are several ways in which the proposed tests could be used. In the most basic sense, the tests could be used to compare the source corpus (C) with a corpus of artifacts generated by the SI system (G). The magnitude of p_{eqv} can indicate how similar the two corpora are. In the case that $p_{\text{eqv}} >= \alpha$, the test for difference can be used to determine if there is a significant difference between the two corpora. In addition, it may be of particular interest to measure the similarity of $\hat{C}_s$ and $\hat{G}_s$, which denotes the projection of C and G into a lower dimensional feature space s. For example, in the music domain, one could use a representation that only contains rhythmic information, and another that

	Test	Corpus A		Corpus B		ACC	TPR	TNR	PPV	NPV	ε
		size	classes	size	classes						
WikiArt	p_{diff}	25	23	25	23	0.85	0.96	0.75	0.79	0.95	-
	p_{eqv}	25	23	25	23	0.78	0.78	0.77	0.78	0.77	0.15
	p_{diff}	50	23	50	23	0.92	0.97	0.87	0.88	0.96	-
	p_{eqv}	50	23	50	23	0.86	0.85	0.87	0.87	0.85	0.1
	p_{diff}	100	23	100	23	0.94	0.98	0.90	0.90	0.98	-
	p_{eqv}	100	23	100	23	0.92	0.94	0.90	0.90	0.93	0.075
	p_{diff}	50	23	100	23	0.82	0.98	0.67	0.75	0.97	-
	p_{eqv}	50	23	100	23	0.88	0.87	0.88	0.88	0.88	0.0875
Classical Archvies	p_{diff}	25	74	25	74	0.98	0.99	0.97	0.97	0.99	-
	p_{eqv}	25	74	25	74	0.92	0.88	0.95	0.94	0.89	0.15
	p_{diff}	50	37	50	37	0.99	1.00	0.99	0.99	1.00	-
	p_{eqv}	50	37	50	37	0.91	0.92	0.90	0.90	0.92	0.1
	p_{diff}	100	20	100	20	0.99	1.00	0.99	0.99	1.00	-
	p_{eqv}	100	20	100	20	0.93	0.94	0.91	0.92	0.94	0.075
	p_{diff}	50	37	100	20	0.85	1.00	0.75	0.77	0.99	-
	p_{eqv}	50	37	100	20	0.89	0.87	0.91	0.91	0.87	0.0875

Table 2: The results of 1000 randomized trials for each statistical test $(p_{\text{eqv}}, p_{\text{diff}})$ using a variety of corpora sizes.

only contains information about the harmonic progression, to gauge the degree to which the SI emulates the rhythm, and harmonic progressions which characterize C.

These tests could also be used to assess the CAN (Elgammal et al. 2017), which attempts to produce visual art in a style that is distinct from those it is trained on. In this scenario, we would have a set of corpora on which the CAN is trained $(S = C_i : i = 1, ..., n)$, and for each $C_i \in S$ we would need to verify that $p_{\text{diff}} < \alpha$, using corrections for multiple hypothesis testing. Most importantly, since *NCD* operates on binary strings, these statistical tests are domain independent, as any digital data can be represented as a binary string.

Conclusion

Scientific progress is hindered in the absence of robust evaluation methodologies. This is an issue of particular contention in the field of computational creativity, as the subjective nature of assessments on creative artifacts can be problematic. In addition to issues of adequate domain knowledge, bias, and inter-rater reliability, the finite capacity of human participants limits the scalability of many evaluation approaches. This is a particular issue for SI systems, where the source corpus is often large, and the generated corpus is infinite. To address this issue, we propose CAEMSI for the evaluation of SI systems, providing compelling evidence that the statistical tests are reliable in two distinct domains. Future work involves further experimention with datasets from other domains, and the evaluation of generative systems with CAEMSI.

Acknowledgments

Thank you to Pierre R. Schwob for providing the Classical Archives MIDI dataset.

References

Amabile, T. M. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology* 43(5):997–1013.

Ariza, C. 2009. The Interrogator as Critic : The Turing Test and the Evaluation of Generative Music Systems. *Computer Music Journal* 33(2):48–70.

Axelsson, S. 2010. Using normalized compression distance for classifying file fragments. In *Proceedings of the 5th International Conference on Availability, Reliability, and Security*, 641–646.

Bardera, A.; Feixas, M.; Boada, I.; and Sbert, M. 2010. Image registration by compression. *Information Sciences* 180(7):1121–1133.

Briot, J.-P., and Pachet, F. 2017. Music Generation by Deep Learning - Challenges and Directions. *arXiv preprint arXiv:1712.04371*.

Burns, K. 2015. Computing the creativeness of amusing advertisements: A Bayesian model of Burma-Shave's muse. *Artificial Intelligence for Engineering Design, Analysis and Manufacturing (AIEDAM)* 29(1):109–128.

Cilibrasi, R., and Vitányi, P. M. B. 2005. Clustering by compression. *IEEE Transactions on Information Theory* 51(4):1523–1545.

Cilibrasi, R.; Vitányi, P.; and De Wolf, R. 2004. Algorithmic clustering of music. *Computer Music Journal* 28(4):49–67.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences*. New York: Lawrence Erlbaum Associates.

Dahlig, E., and Schaffrath, H. 1997. Judgements of human and machine authorship in real and artificial folksongs. *Computing in Musicology* 11:211–219.

Eerola, T.; Himberg, T.; Toiviainen, P.; and Louhivuori, J. 2006. Perceived complexity of western and African folk melodies by western and African listeners. *Psychology of Music* 34(3):337–371.

Elgammal, A.; Liu, B.; Elhoseiny, M.; and Mazzone, M. 2017. CAN: Creative Adversarial Networks, Generating Art by Learning About Styles and Deviating from Style Norms. In *Proceedings of the 8th International Conference on Computational Creativity (ICCC)*, 96–103.

Friedman, R. S., and Taylor, C. L. 2014. Exploring emotional responses to computationally created music. *Psychology of Aesthetics, Creativity, and the Arts* 8(1):87–95.

Gatt, A., and Krahmer, E. 2017. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research (JAIR)* 61(1):65–170.

Gonzalez Thomas, N.; Pasquier, P.; Eigenfeldt, A.; and Maxwell, J. 2013. Meta-Melo: A system and methodology for the comparison of melodic generation models. In *Proceedings of the 14th International Symposium on Music Information Retrieval (ISMIR)*, 561–566.

Kaufman, J. C.; Baer, J.; and Cole, J. C. 2009. Expertise, domains, and the consensual assessment technique. *Journal of Creative Behavior* 43(4):223–233.

Kocsor, A.; Kertész-Farkas, A.; Kaján, L.; and Pongor, S. 2006. Application of compression-based distance measures to protein sequence classification: a methodological study. *Bioinformatics* 22(4):407–412.

Lamb, C.; Brown, D. G.; and Clarke, C. L. A. 2015. Human Competence in Creativity Evaluation. In *Proceedings of the 6th International Conference on Computational Creativity*, 102–109.

Li, M., and Sleep, R. 2005. Genre Classification via an LZ78-Based String Kernel. In *Proceedings of the 6th International Symposium on Music Information Retrieval*, 252–259.

Li, M.; Chen, X.; Li, X.; Ma, B.; and Vitányi, P. M. B. 2004. The Similarity Metric. *IEEE Transactions on Information Theory* 50(12):3250–3264.

Liang, F.; Gotham, M.; Johnson, M.; and Shotton, J. 2017. Automatic Stylistic Composition of Bach Chorales with Deep LSTM. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 449–456.

Maher, M., and Fisher, D. 2012. Using AI to evaluate creative designs. In *Proceedings of the 2nd International Conference on Design Creativity*, 45–54.

Maher, M. L. 2010. Evaluating creativity in humans, computers, and collectively intelligent systems. In *Proceedings of the 1st DESIRE Network Conference on Creativity and Innovation in Design*, 22–28.

Marozzi, M. 2004. Some remarks about the number of permutations one should consider to perform a permutation test. *Statistica* 64(1):193–201.

Moffat, D. C., and Kelly, M. 2006. An investigation into people's bias against computational creativity in music composition. In *Proceedings of the 3rd International Joint Workshop on Computational Creativity*.

Norton, D.; Heath, D.; and Ventura, D. 2015. Accounting for Bias in the Evaluation of Creative Computational Systems: An Assessment of DARCI. In *Proceedings of the 6th International Conference on Computational Creativity*, 31–39.

Pasquier, P.; Burnett, A.; Thomas, N. G.; Technology, A.; Maxwell, J. B.; and Loughin, T. 2016. Investigating Listener Bias Against Musical Metacreativity Introduction : Computational Creativity. In *Proceedings of the 7th International Conference on Computational Creativity*, 47–56.

Pasquier, P.; Eigenfeldt, A.; Bown, O.; and Dubnov, S. 2017. An Introduction to Musical Metacreation. *Computers in Entertainment* 14(2):1–14.

Pearce, M. T., and Wiggins, G. a. 2007. Evaluating Cognitive Models of Musical Composition. In *Proceedings of the 4th International Joint Workshop on Computational creativity*, 73–80.

Pearce, M.; Meredith, D.; and Wiggins, G. 2002. Motivations and Methodologies for Automation of the Compositional Process. *Musicae Scientiae* 6(2):119–147.

Pearce, M. T. 2005. *The Construction and Evaluation of Statistical Models of Melodic Structure In Music Perception and Composition*. Ph.D. Dissertation, City, University of London.

Pesarin, F.; Salmaso, L.; Carrozzo, E.; and Arboretti, R. 2016. Unionintersection permutation solution for two-sample equivalence testing. *Statistics and Computing* 26(3):693–701.

Ritchie, G. 2007. Some empirical criteria for attributing creativity to a computer program. *Minds and Machines* 17(1):67–99.

Rocha, R.; Ribeiro, A.; and El, E. 2016. Regent-Dependent Creativity : A Domain Independent Metric for the Assessment of Creative Artifacts. In *Proceedings of the 7th International Conference on Computational Creativity*, 68–75.

Roy, S. 1953. On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics* 24(2):220–238.

Schedl, M.; Flexer, A.; and Urbano, J. 2013. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems* 41(3):523–539.

Simpson, S. L.; Lyday, R. G.; Hayasaka, S.; Marsh, A. P.; and Laurienti, P. J. 2013. A permutation testing framework to compare groups of brain networks. *Frontiers in Computational Neuroscience* 7(11):1–13.

Solomonoff, R. 1964. A formal theory of inductive inference. Part I. *Information and Control* 7(1):1–22.

Väyrynen, J., and Tapiovaara, T. 2010. Normalized compression distance as an automatic MT evaluation metric. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, 343–348.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4):600–612.

Yang, L.-C.; Chou, S.-Y.; and Yang, Y.-H. 2017. MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, 324–331.