On Narrative Information and the Distillation of Stories

Anonymous Author(s)
Affiliation
Address
email

Abstract

The act of telling stories is a fundamental part of what it means to be human. This work introduces the concept of narrative information, which we define to be the overlap in information space between a story and the items that compose the story. Using contrastive learning methods, we show how modern artificial neural networks can be leveraged to distill stories and extract a representation of the narrative information. We then demonstrate how evolutionary algorithms can leverage this to extract a set of narrative templates and how these templates—in tandem with a new novel curve-fitting algorithm we introduce—can reorder music albums to automatically induce stories in them. In the process of doing so, we give strong statistical evidence that these templates under narrative information are present in existing albums. While we experiment only with music albums here, the premises of our work extend to any form of (largely) independent media.

1 Introduction

Our ability to comprehend and devise narratives is a fundamental aspect of our nature, (see e.g., Ricoeur [17]), so much so that some dub humans the storytelling animal [6]. Work in cognitive psychology often relates our ability for story comprehension to Theory of Mind, showing overlap in their cognitive networks [13]. This suggests that one of the reasons we engage in story-telling is to hone our social cognition.

However, there are aspects of stories that go beyond the interaction of interpretation of cognitive agents (whether they be fiction or real). Stories are embedded even in collections of largely independent media that do not explicitly feature cognitive agents. For instance, we can talk about narrative choices curators made when arranging an art exhibition, or the overarching story that the track list of a concert induces. This work investigates a very general notion of *stories* as being collections of *atoms* (i.e., words, images, etc) and some meaningful ordering over the atoms which we refer to as the *narrative*. In line with this, we define here the *narrative information* of an atom as the mutual information between the atom and the story itself. In Section 2 we define the *narrative essence* of an atom as the low-dimensional learned representation of its narrative information and show how contrastive learning can be used to obtain it. In Section 3, we experimentally show that narrative essence can be used to extract prototypical narrative templates from music albums that partially explain the order in which the songs are arranged. We argue that qualitatively similar results should hold for most collections of (largely) independent media.

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Narrative Essence

A strong narrative is induced into any collection of media simply by putting it in a specific order. By doing that, we can start with our story with bang, or slowly build up excitement. We can place a climax at a certain position, and end on a high or a low note. Or something in between. There are intrinsic properties of each atom that determine its function—and thus also its placement—in a narrative. This insight leads to our formulation of *narrative essence*: it is a low-dimensional representation of the latent property of the items that is most informative about the narrative.

Formally, we define as narrative essence $f_E(x)$ of atom x, generated by a feature extractor f_E, as the feature which maximizes mutual information between the unordered set of features of the atoms in a collection c and the correct order $o(c)$ of c:

$$f_E = \arg\max_f I\big(\{f(x)|x \in c\}; o(c)\big). \tag{1}$$

In other words, narrative essence is the intrinsic feature of each atom that, if we know it for every atom in a collection, allows us to best predict the correct order of the collection.

Learning Narrative Essence from Data If we have a dataset of collections of a certain kind of media, we can use it to learn the narrative essence extractor f_E. For that, we use noise contrastive estimation [8]—specifically, a modification of InfoNCE [20]. The idea is the following: we give each item in a collection to a learnable feature extractor f_θ, a neural network with parameters θ. A second learnable function g_ϕ, a recurrent neural network with parameters ϕ, takes a sequence s of features as input and produces a scalar score $g_\phi(s)$. If g_ϕ receives a sequence in the correct ground-truth order, $s^* = (f_\theta(x_1), f_\theta(x_2), f_\theta(x_3), ...)$, it should produce a high score. For randomly ordered sequences, it should produce a low score. g_ϕ can only achieve this if (1) the correct orders of the collections in our dataset have some property that distinguishes them from random orders, and (2) f_θ learns some atom-wise feature that lets g_ϕ recognize this property. It can be achieved by training feature extractor f_θ and sequence model g_ϕ jointly to minimize the loss

$$\mathcal{L}_{\mathrm{N}} = -\mathbb{E}_{S\sim\mathcal{D}}\left[\log \frac{g_\phi(s^*)}{\sum_{s\in S} g_\phi(s)}\right], \tag{2}$$

where $\mathcal{D}$ is the dataset of collections with a ground truth order, and S is a set of N sequences that includes the correctly ordered sequence s^*. The other $N-1$ sequences in S are random permutations of s^*. The extracted features should be normalized across the sequence so that g_ϕ considers the relative, and not the absolute value the extracted feature.

In analogy to van den Oord et al. [20], it can be shown that minimizing $\mathcal{L}_{\mathrm{N}}$ maximizes a lower bound on the mutual information between the atom-wise features extracted by f_θ and the order of the collection:

$$I\big(\{f_\theta(x)|x \in c\}; o(c)\big) \geq \log(N) - \mathcal{L}_{\mathrm{N}}. \tag{3}$$

Additional details are provided in Appendix F. Formulating narrative essence as we have done here enables a general and sound quantitative approach for determining whether collections of a particular type follow any narrative principles and the features of this narrative.

3 Experiments on Music Albums

In this section, we empirically investigate the notion of narrative essence using the example of music albums. We selected the FMA dataset [4] as it is—at the time of writing—the largest open dataset that includes raw audio files.

While in principle, very sophisticated and specialized feature extraction architectures could be used for f_θ, in our experiments, we restrict ourselves to relatively simple and computationally cheap methods. Each track is represented by features commonly used in music information retrieval that come pre-computed with the available dataset. These features form a sequence of 75 vectors of size 7 (for more details, see Appendix F.1). This

Table 1: Mutual Information (in bits) on the FMA validation set for different dimensionalities of narrative essence. Results are from five runs.

Features	Mutual Information
1	1.924 ± 0.0296
2	1.936 ± 0.0183
4	1.957 ± 0.0217
8	1.950 ± 0.0216
16	1.975 ± 0.0150

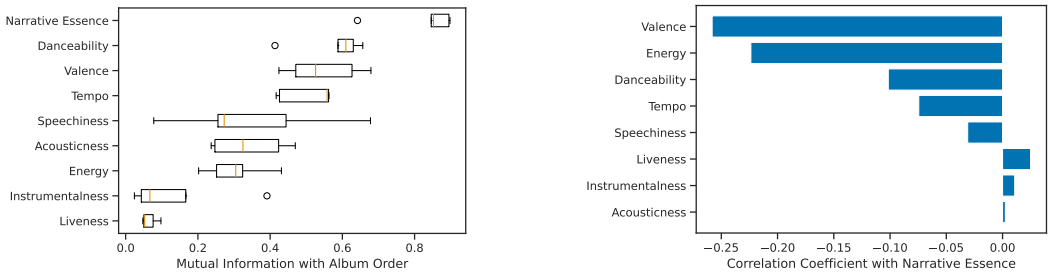

Figure 1: **(left)** The lower bound of the mutual information in bits between different features and the album order. Results are shown over five seeds. **(right)** The Pearson correlation coefficient of different features with the narrative essence.

sequence is the input to the feature extractor f_θ, for which we use a bidirectional LSTM model [7, 9]. We choose a recurrent feature encoder instead of a feed-forward architecture to give f_θ more powerful conditional processing abilities.

The output of f_θ is the narrative essence feature of the given song. In principle, the narrative essence could be a vector of any size. However, Table 1 shows that a higher dimensional narrative essence leads to only marginal improvements in terms of mutual information captured. This provides strong evidence that narrative essence, at least for songs in the context of an music album, can be represented as a scalar value. A benefit of this is that it is directly comparable to other available scalar features (see section 3.1). We model g_ϕ as a bidirectional LSTM as well. It takes a sequence of narrative essence features as input and computes a scalar score. In comparison to f_θ, g_ϕ has a lower capacity (fewer learnable parameters and more regularization) because there are much fewer full collections (albums) than individual items (songs). Thus g_ϕ is at a considerable risk of overfitting. A full description of the training setup can be found in Appendix F.2. When trained on the full FMA training set[1], the extracted narrative essence achieves a mutual information with the album ordering, as determined by equation 3, of ca. 1.924 bits on the validation set.

3.1 Narrative Essence in Comparison with other Features

When treating the feature extractor f_θ as fixed and only learning the scoring model g_ϕ to minimize $\mathcal{L}_{\mathrm{N}}$, we can use Equation 3 to approximate the mutual information between any available feature and the collection orders. Here, we compare the narrative essence feature extracted using f_θ learned on the FMA dataset with some of the other features available in the dataset. To do so, we learn a dedicated scoring model g_ϕ for each available feature—including energy, tempo and valence, which is a feature designed to capture the mood of a song, roughly ranging from sad to happy [18]. As can be seen in the left side of Figure 1, narrative essence has more mutual information with the album order than any of the other features. Note that the mutual information lower bounds seen in Figure 1 are significantly lower than the ones achieved when training on the full dataset (compare Table 1). This is a consequence of the fact that we are limited to the subset of the FMA dataset includes the listed pre-computed features when learning the scoring models. Nevertheless, these results show that our formulation of narrative essence in combination with the very general processing techniques (global track feature statistics and an LSTM encoder) robustly outperforms highly engineered features as a candidate for the narrative.

The right side of Figure 1 shows the correlation of the learned narrative essence feature with eight other features. Here we see a high correlation with the features we would expect to be associated with the narrative (e.g. valence, energy) and a low correlation with more technical or applied features like acousticness, instrumentalness, and liveness. Note that the orientation (sign) of the narrative essence is simply a random product of the initialization and has no further meaning; the negative of the narrative essence would have exactly the same amount of mutual information with the collection order.

[1]Note that here and everywhere else we have excluded albums with less than 3 or more than 20 tracks.

3.2 Narrative Essence and Story Templates

The concept of prototypical narratives has been explored in the dramatic arts—such as novels, plays or operas—for a long time [3, 5, 21]. Other kind of media, like music albums, may follow different types of narratives. This section uses the genetic algorithm described in Appendix D to learn a set of k template curves from albums in the FMA dataset using the learned narrative essence feature.

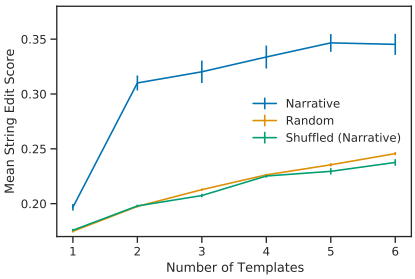

Figure 2: The performance on the FMA validation set of the templates learned with narrative essence.

To evaluate the quality of a set of templates, we find the orderings of the albums that come closest to our learned templates using a novel curve fitting algorithm (described in Appendix E). Then we encode each song as an alphanumeric character and compare the encoded string representing the original ordering with the likewise encoded string representing the fitted album. We borrow from computational linguistics and use a string edit score[2] to determine the similarity of the fitting to the original order and thus the quality of the fitting.[3] To account for the assumption that an album could follow one of several different story curves, we always assign albums to the best fitting curve. We compare the maximal mean string edit score when fitting the narrative essence to the learned templates with (1) the maximal mean string edit score of k random orderings and (2) the maximal mean string edit score when the reported valances of the tracks in each album are randomly shuffled. The latter baseline captures the gain from fitting the narrative essence as opposed to fitting noise here. We report these baselines and the mean of the fittings on the validation split as a function of k in Figure 2. Our results show with statistical significance ($p < 0.05$) that the album ordering can be partially explained by the narrative essence following our learned templates (details in Appendix C). See Appendix A for a closer look at the template curves found for the FMA dataset. In Appendix B, we showcase a potential use of these techniques.

4 Related Work

The use of machine learning to derive narrative arcs has previously been explored by Reagan et al. [15]—who used machine learning to explicitly derive the emotional arc of stories from a large corpus of English texts. More recently, Mathewson et al. [14] used an information-theoretic approach to design a narrative arc and applied this to dialogue generation. For music playlist ordering, work remains sparse. However, a considerable body of work has recently emerged in music playlist continuation (e.g., Bonnin and Jannach [2], Maillet et al. [12], and Vall et al. [19]). For a detailed overview of spatial representation of musical form, see Bonds [1].

5 Conclusion and Future Work

Inspired by their seeming importance of in human cognition, this work set out to use machine learning to distill the narrative information from stories. In doing so, we showed how to produce the narrative essence of their compositional atoms. We then used evolutionary algorithms alongside narrative essence to extract a series of templates from the music albums in the FMA dataset. We went on to give statistical evidence that these templates described the ordering of music albums and showed that even one-dimensional narrative essence explains them better than other commonly used music track features. We hope that this work can be applied to partially automate the induction of stories in existing collections of works (e.g., photo galleries, music playlists, etc.). While we only experimented with music albums here, our work extends to any collection of (largely) independent media.

[2]We define the string edit score here as $f(T, o) = \max_{\hat{o} \in T} \frac{1}{1+g(\hat{o},o)}$ where T is a set of k proposed orders, o is the ground-truth order, and g is the string edit distance (i.e., Levenshtein distance) function.

[3]Preliminary results suggest that similar metrics, such as Spearman's rank correlation coefficient, produce qualitatively similar results.

Acknowledgements

REDACTED

References

Bonds, M. E. (2010). The spatial representation of musical form. *Journal of Musicology*, *27*(3), 265–303. https://doi.org/10.1525/jm.2010.27.3.265

Bonnin, G., & Jannach, D. (2013). A comparison of playlist generation strategies for music recommendation and a new baseline scheme. *Papers from the 2013 AAAI Workshop*, 16–23. https://www.aaai.org/ocs/index.php/WS/AAAIW13/paper/view/7078

Campbell, J. (2008). *The hero with a thousand faces* (3rd ed.). New World Library.

Defferrard, M., Benzi, K., Vandergheynst, P., & Bresson, X. (2017). FMA: A dataset for music analysis, 316–323. https://ismir2017.smcnus.org/wp-content/uploads/2017/10/75%5C_Paper.pdf

Freytag, G. (1894). *Die technik des dramas*. S. Hirzel.

Gottschall, J. (2012). *The storytelling animal: How stories make us human*. Houghton Mifflin Harcourt.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional lstm networks. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, *4*, 2047–2052.

Gutmann, M., & Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopcroft, J. E., & Karp, R. M. (1973). An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, *2*(4), 225–231. https://doi.org/10.1137/0202019

Jonker, R., & Volgenant, A. (1987). A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, *38*(4), 325–340. https://doi.org/10.1007/BF02278710

Maillet, F., Eck, D., Desjardins, G., & Lamere, P. (2009). Steerable playlist generation by learning song similarity from radio station playlists. *Proceedings of the 10th International Society for Music Information Retrieval Conference*, 345–350. https://archives.ismir.net/ismir2009/paper/000012.pdf

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual review of psychology*, *62*, 103–34.

Mathewson, K. W., Castro, P. S., Cherry, C., Foster, G. F., & Bellemare, M. G. (2020). Shaping the narrative arc: Information-theoretic collaborative dialogue. *Proceedings of the 11th International Conference on Computational Creativity*, 9–16. http://computationalcreativity.net/iccc20/papers/010-iccc20.pdf

Reagan, A. J., Mitchell, L., Kiley, D., Danforth, C. M., & Dodds, P. S. (2016). The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, *5*(1), 31–42. https://doi.org/10.1140/epjds/s13688-016-0093-1

RIAA. (2021). *Gold & platinum*. Recording Industry Association of America. https://www.riaa.com/gold-platinum/?tab_active=awards_by_album

Ricoeur, P. (1991). Narrative identity. *Philosophy today*, *35*(1), 73–81.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*(3), 154–165.

Vall, A., Quadrana, M., Schedl, M., & Widmer, G. (2019). Order, context and popularity bias in next-song recommendations. *International Journal of Multimedia Information Retrieval*, *8*(2), 101–113. https://doi.org/10.1007/s13735-019-00169-8

van den Oord, A., Li, Y., & Vinyals, O. (2018). *Representation learning with contrastive predictive coding*. arXiv. http://arxiv.org/abs/1807.03748

Vonnegut, K. (1981). *Palm sunday*. Delacorte Press.

A FMA Template Curves

During our experiments, we noted that when $k = 3$, a relatively flat curve usually appears in the set of templates. We thus hypothesize that, similar to when $k = 1$ the presence of this curve is likely a hybrid of many different kinds of story curve. From $k = 4$ onward, this curve is no longer prominently. Thus we argue that $k = 4$ is a good minimal choice for the number of dominant narrative structures in the FMA dataset. The best-performing learned templates curves for $k = 4$, along with the best and worst fitting albums assigned to these curves, are shown in Figure 3.

Figure 4 takes a closer look at the narrative templates found in the FMA dataset. By comparing the learned templates for different k, we can draw a genealogy of prototypical narrative arcs—starting from the almost wholly flat average (blue) we for $k = 1$, to an increasing diversity of story curves.

While it is beyond the scope of this work to analyze the genealogy of these curves in detail, we note that the curves here seem to display some notable differences with story curves found elsewhere: where normally there would be a climax, story curves of albums in the FMA dataset seem to instead become more neutral as the album progresses. Further analysis of this phenomenon is left as future work.

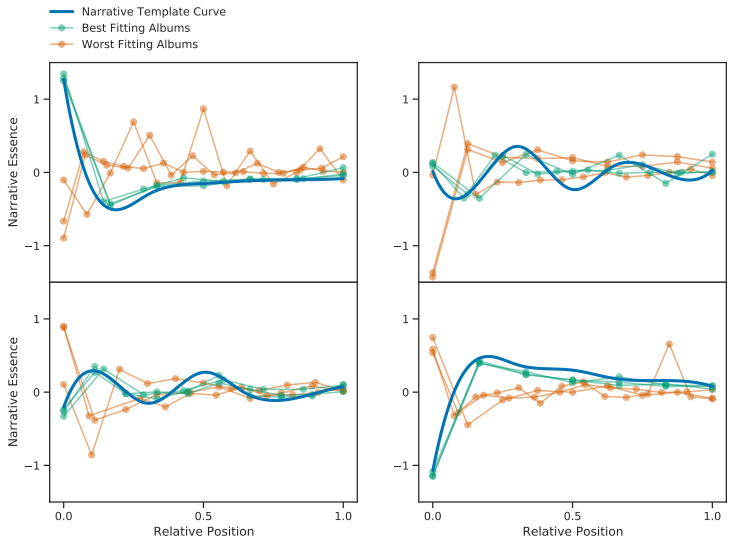

Figure 3: Four narrative template curves, as well as the three best and the three worst fitting albums (in terms of mean squared distance) assigned to each curve. Only albums of typical length (between 7 and 15 tracks) are considered in this plot.

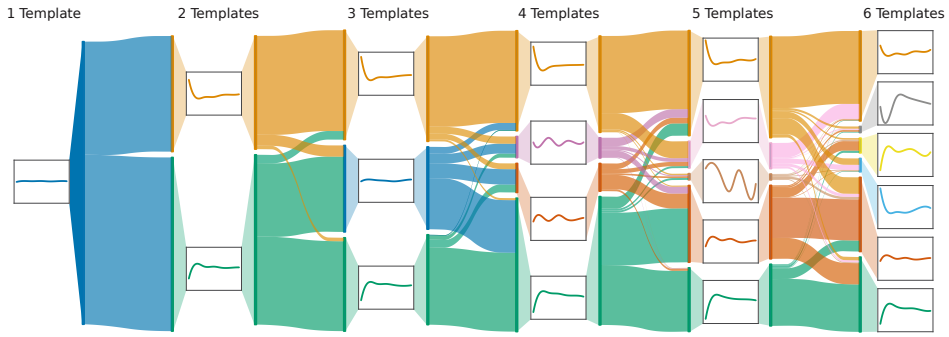

Figure 4: How the assignment of individual songs to template curves learned with narrative essence progress as the number of templates increases. Colors show post-hoc analysis to try and find similar prototypical curves.

B Demonstration

Figure 5 in gives a demonstration of the practical application of this work. Here, we used the narrative essence feature of the tracks in Michael Jackson's *Thriller* to fit the album to the set of four distinct narrative template curves given in Figure 4. By doing so, we have induced several different stories in the album. The methods presented here can trivially carry out the same task with any album.

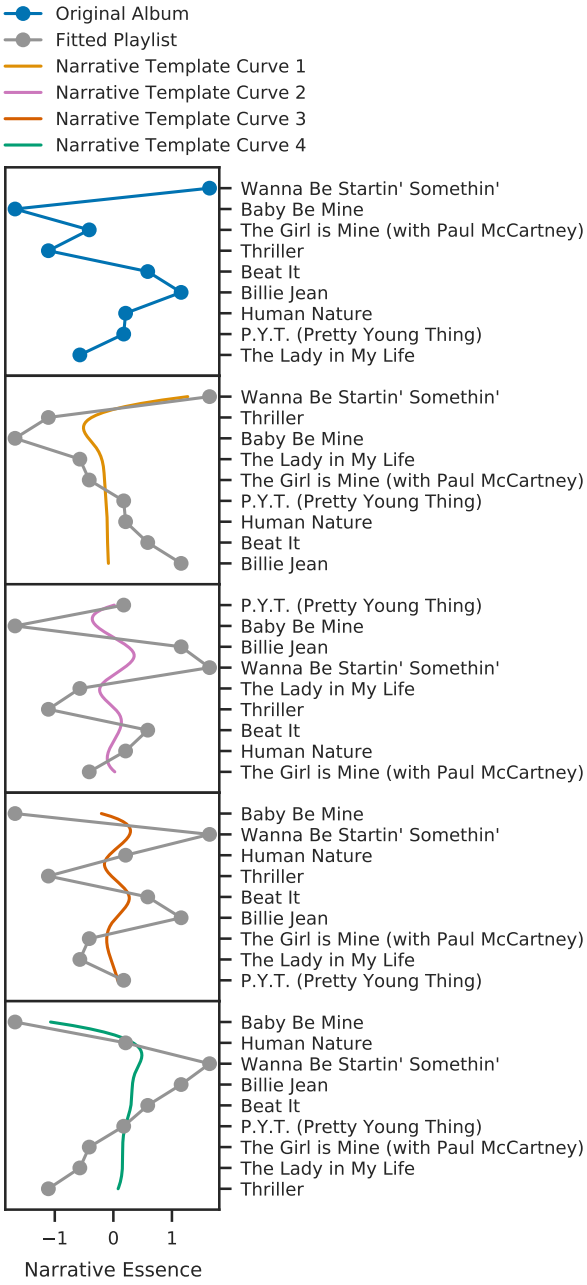

Figure 5: Narrative Essence of the album *Thriller* by Michael Jackson—the best-selling original album of all time [16]—in the original order, and fitted to the four narrative template curves found using the method described in Section D

C Evidence for the Existence of Story Templates

An important question we must address while looking at Figure 2 is whether or not the improvement of the learned curves over the baselines is significant. This is equivalent to asking if the order of the albums is partially explained by the narrative essence and thus if the narrative structures discovered by our algorithm exist within music albums. To answer this, we compare the mean string edit score for the selected $k = 4$ templates with the mean string edit score for both baselines on the test set. We find that the difference observed is significant with a family-wise error rate of $p < 0.05$ using t-tests with Holm-Bonferroni corrections.

D Story Template Extraction Algorithm

Extracting a set of narrative arc templates from a collection of albums can be done using Algorithm 1. Note that this algorithm is general and can make use of any collection of media if the narrative essence is replaced by a semantically similar metric. In our experiments, we always used $\mathbf{x} = [0.0, 0.2, 0.3, 0.5, 0.65, 0.8, 1.0]^T$. To derive the value of a template at a given x, cubic-spline interpolation is recommended; for the cost of fitting an album to a template, using the mean-squared error is recommended.

Algorithm 1 Story Template Extraction

Input: $\mathbf{x} = [x_0, x_1, ..., x_q]^T$ where x_i is the relative position of the i-th point in the desired templates and a set of albums $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_n\}$ with each $\mathbf{a}_i = \{(u_0, v_0), (u_1, v_1), ..., (u_m, v_m)\}^T$ where u_j is the relative position of track j in the album, and v_j is the normalized narrative essence of track j

Output: set of templates $\{\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p\}$ with each $\mathbf{t}_i = [y, y_1, ..., y_q]^T$ where y_j is the normalized narrative essence of the j-th point in the template

$s \leftarrow$ population size
$b \leftarrow$ number of children for each generation
for $i \in \{1..s\}$ **do**
  for $j \in \{1..p\}$ **do**
    for $k \in \{1..q\}$ **do**
      $P[i, j, k] \leftarrow \mathcal{N}(0, 1)$
    end for
  end for
end for

while not done **do**
  $\sigma = \mathcal{N}(0, 1)$
  for $i \in \{1..b\}$ **do**
    $father \leftarrow$ random integer in $\{0, 1, ..., s\}$
    $mother \leftarrow$ random integer in $\{0, 1, ..., s\} - \{father\}$
    for $j \in \{1..p\}$ **do**
      for $k \in \{1..q\}$ **do**
        $P[s + i, j, k] \leftarrow P[father, j, k]$ with probability p and $P[mother, j, k]$ with probability $1 - p$
        $P[s + i, j, k] \leftarrow P[s + i, j, k] + \mathcal{N}(0, \sigma)$
      end for
    end for
  end for
  for $i \in \{1..(b + s)\}$ **do**
    $\mathbf{c}_i \leftarrow$ minimum cost as defined in Equation 4 for fitting albums using the templates $P[i, :, :]$
  end for
  order P in increasing order of corresponding $\mathbf{c}$
  $P \leftarrow P[1 : s, :, :]$
end while
return P

While many different cost functions for a set of templates could be used here, we uses the following:

$$\mathbf{c} = \sum_{i=1}^{n} \min_{p} \frac{1}{l_i} \sum_{j=1}^{l_i} \left(v_i(j) - t_p(j_r)\right)^2, \tag{4}$$

where n is the number of albums in the training set, l_i the number of tracks in album i, $v_i(j)$ the normalized narrative essence value of the jth track of album i, and $t_p(j_r)$ is the value of template p at the relative position $j_r = (j - 1)/(l_i - 1)$. We learn these templates using the training split provided by the FMA dataset and evaluate them on the validation split by fitting the narrative essence of each album to the templates using the algorithm given in Appendix E.

E Template Curve Fitting Algorithm

Deriving an ordering of the media such that their respective values fit a narrative template can be done using Algorithm 2. The ordering Algorithm 2 finds will be minimal first in the maximum deviation of a value from the template curve and minimal second in the average deviation of values from the template curve. For n items, the worst-case time complexity of this algorithm—provided efficient bipartite matching algorithms such as Hopcroft-Karp [10] and LAPJVsp [11] are used—is in $O(n^3)$. In most applications of this work—and for all but the largest collections of independent media—extracting the values that will be fitted will consume vastly more time than the fitting itself.

Algorithm 2 Template Curve Fitting

Input: normalized values to fit $\mathbf{y}$ and template curve function f with domain and range $[0, 1]$
Output: ordering $\mathbf{x}$ over values $\mathbf{y}$ such that the i-th value in the ordering is $\mathbf{x}[i]$

$\mathbf{z} \leftarrow \left[f(\frac{0}{|\mathbf{y}|-1}), f(\frac{1}{|\mathbf{y}|-1}), ..., f(\frac{|\mathbf{y}|-1}{|\mathbf{y}|-1})\right]^T$
$\mathbf{d} \leftarrow \mathbf{y}\mathbf{z}^T$

$a \leftarrow 1$
$b \leftarrow |\mathbf{d}|$
while $a \neq b$ **do**
    $p \leftarrow a + \lfloor (b-a)/2 \rfloor$
    $L, R \leftarrow \{1..|\mathbf{y}|\}$
    $E \leftarrow \{(i \in L, j \in R) \mid \|\mathbf{y}[i] - \mathbf{z}[j]\| \leq \mathbf{d}[p]\}$
    if $\exists$ perfect matching for bipartite graph (L, R, E) **then**
        $b \leftarrow p$
    else
        $a \leftarrow p + 1$
    end if
end while

$L, R \leftarrow \{1..|\mathbf{y}|\}$
$E \leftarrow \{(i \in L, j \in R, \|\mathbf{y}[i] - \mathbf{z}[j]\|) \mid \|\mathbf{y}[i] - \mathbf{z}[j]\| \leq \mathbf{d}[a]\}$
$M \leftarrow$ minimum-cost perfect matching for weighted bipartite graph (L, R, E)
for $i \in \{1..|\mathbf{y}|\}$ **do**
    for $j \in \{1..|\mathbf{y}|\}$ **do**
        if $(i, j) \in M$ **then**
            $\mathbf{x}[j] = i$
        end if
    end for
end for

return $\mathbf{x}$

F Narrative Essence as Mutual Information with the Collection Order

Recall that c is an unordered collection of items x, $o(c)$ is its correct order. S is a set of N sequences of the encoded items, containing the correct sequence $s^* = (f_\theta(x_1), f_\theta(x_2), f_\theta(x_3), ...)$, i.e., the one adhering to $o(c)$, and $N-1$ random permutations of s^*. The probability that a particular sequence s_i from S is the correct sequence s^* is

$$
\begin{aligned}
p(s_i = s^* | S, o(c)) &= \frac{p(S|p(s_i = s^*, o(c))}{\sum_j p(S|p(s_j = s^*, o(c))} \\
&= \frac{p(s_i|o(c)) \prod_{l \neq i} p(s_l)}{\sum_j p(s_j|o(c)) \prod_{l \neq j} p(s_l)} \\
&= \frac{\frac{p(s_i|o(c))}{p(s_i)}}{\sum_j \frac{p(s_j|o(c))}{p(s_j)}}.
\end{aligned}
$$

With Equation 2, $g_\phi(s)$ is trained to estimate the density ratio $\frac{p(s|o(c))}{p(s)}$. This means that that we can write (following the steps from [20])

$$
\begin{aligned}
\mathcal{L}_{\mathrm{N}}^{\mathrm{opt}} &= -\mathbb{E}_{S \sim \mathcal{D}} \log \left[\frac{\frac{p(s^*|o(c))}{p(s^*)}}{\frac{p(s^*|o(c))}{p(s^*)} + \sum_{s \in S_{\mathrm{neg}}} \frac{p(s|o(c))}{p(s)}} \right] \\
&\approx \mathbb{E}_{S \sim \mathcal{D}} \log \left[1 + \frac{p(s^*|o(c))}{p(s^*)} (N-1) \right] \\
&\geq \mathbb{E}_{S \sim \mathcal{D}} \log \left[\frac{p(s^*, o(c))}{p(s^*) p(o(c))} N \right] \\
&= -I(s^*; o(c)) + \log(N) \\
&= -I(f_E(x_1), f_E(x_2), f_E(x_3), ...; o(c)) + \log(N).
\end{aligned}
$$

F.1 Track Input Features

The FMA dataset provides the following track features: 12 Chroma features, 6 Tonnetz features, 20 MFCC features, Spectral centroid, Spectral bandwidth, 7 Spectral contrast features, Spectral rolloff, RMS energy and Zero-crossing rate.

For every feature, 7 global statistical properties are given: mean, standard deviation, skew, kurtosis, median, minimum and maximum. We treat these statistical properties as a vector and construct a sequence of these vectors from the 75 features, which constitutes the input for the narrative essence extractor f_θ. The length of this feature sequence is constant, independent of the track's length.

For tracks that are not included in the FMA data, these features can easily be computed directly from audio standard MIR techniques (the implementation is provided by [4]). Before giving the sequence to g_ϕ, learnable start- and end-of-sequence tokens are added.

F.2 Model Hyperparameters

All models described in Section 3 have the same hyperparameters. The batch size is 16, N is 32 (that means we have 31 negative samples for each example in the batch).

The feature encoder f_θ is a bidirectional LSTM with 2 layers, 7 input features, 128 hidden units and a sigmoid output nonlinearity. For regularization we use dropout of 0.1 and no weight decay.

The sequence scoring model g_ϕ is also bidirectional LSTM with 2 layers. It has 32 hidden units and no output nonlinearity. For regularization we use no dropout and a weight decay of 10^{-5}.

For both models, we use the Adam optimizer with a learning rate of 10^{-4}, and early stopping based on the validation loss.

G Source Code

The source code used to generate the results presented in this paper is available at REDACTED