Automatic Similarity Detection in LEGO Ducks

Mark Ferguson[1], Sebastian Deterding[1], Andreas Lieberoth[2], Marc Malmdorf Andersen[2], Sam Devlin[3], Daniel Kudenko[4] and James Alfred Walker[1]

[1] University of York, UK. maf541@york.ac.uk, sebastian.deterding@york.ac.uk, james.walker@york.ac.uk

[2] Aarhus University, Denmark. andreas@edu.au.dk, mana@cas.au.dk

[3] Microsoft Research Cambridge, UK. sam.devlin@microsoft.com

[4] Leibniz University Hannover, Germany. kudenko@l3s.de

Abstract

The automated evaluation of creative products promises both good-and-scalable creativity assessments and new forms of visual analysis of whole corpora. Where creative works are not 'born digital', such automated evaluation requires fast and frugal ways of transforming them into data representations that can be meaningfully assessed with common creativity metrics like novelty. In this paper, we report the results of training a Spatiotemporal DeepInfomax Variational Autoencoder (STDIM-VAE) on a digital photo pool of 162 LEGO ducks to generate a phenotypical landscape of clusters of similar ducks and dissimilarity scores for individual ducks. Visual inspection suggests that our system produces plausible results from image pixels alone. We conclude that under certain conditions, STDIM-VAEs may provide fast and frugal ways of automatically assessing corpora of creative works.

1 Introduction

How to evaluate creativity is a major and ongoing concern in research, the creative industries, education, and many other areas. Researchers have developed numerous methods to assess both human and computational creativity across the "four P's" – person, process, product, and press/environment (Kaufman, Plucker, and Baer 2008; Lamb, Brown, and Clarke 2018; Jordanous 2019; Plucker, Makel, and Qian 2019). In psychology, expert evaluations of creative products are often seen as the 'gold standard' (Plucker, Makel, and Qian 2019). In computational creativity evaluation, human judges are likewise frequently used (Lamb, Brown, and Clarke 2018). Expert evaluations feature comparatively high reliability, intersubjectivity, predictive value, and ecological and criterion validity: they are close to everyday practices around creative works like critiques, reviews, or prizes (Lamb, Brown, and Clarke 2018; Plucker, Makel, and Qian 2019). They also embody the sense-making within a social context that most contemporary definitions consider essential to creativity (Colton and Wiggins 2012; Plucker, Beghetto, and Dow 2004).

1.1 Automated Creativity Assessment

That said, expert evaluations are labour-intensive (Lamb, Brown, and Clarke 2018; Jordanous 2019); they don't scale to the volumes of creative products one may find with cultural archives, generative systems, or large-scale testing. This has led researchers to trial computational formalisations of creativity such as novelty to automate creativity evaluation (Lamb, Brown, and Clarke 2018). One unique opportunity of automated evaluations is that they can provide rich statistical and visual analyses of a whole corpus (Grace et al. 2015; Elgammal and Saleh 2015), in addition to individual products. This makes them potentially akin to cultural analytics (Manovich 2016), the use of computational and visualisation techniques to analyse massive cultural data sets. A good example are expressive range analyses of procedural content generators (Summerville 2018), which visualise the frequency space of creative outputs of a system as a heat map.

While there have been some attempts to apply automated creativity evaluation to human works (Grace et al. 2015; Karampiperis, Koukourikos, and Koliopoulou 2014; Elgammal and Saleh 2015; Zhu, Xu, and Khot 2009), it remains chiefly confined to computational creativity (Lamb, Brown, and Clarke 2018). Human creativity assessment continues to use either poor-but-scalable self-reports and test batteries or good-but-expensive human expert evaluations (Kaufman, Plucker, and Baer 2008). We are missing robust, usable, validated computational tools for automatically evaluating human creative works. Such tools could not only provide cheaper, more reliable creativity measurement at scale: they would also allow us to analyse whole corpora of creative products in the style of cultural analytics.

1.2 The Context: The LEGO Duck Task

In response, we have been exploring the automated evaluation of a creativity exercise, the LEGO Duck Task (henceforth 'Duck Task'). In this task, participants are instructed to make a duck from a standard set of six LEGO bricks. Task instructions can vary from e.g. making 'the most creative duck' to making as many different ducks as possible in a given time. The Duck Task has many attractive features for (automated) creativity assessment: It is easily understood across cultures. It is repeatable, unlike other task-based assessments where knowing the solution biases subsequent runs. Recombining the six bricks opens a vast phenotypic landscape of possible ducks and non-ducks. And yet the six bricks present a small set of simple, low-dimensional shapes that are relatively easy to formalise in terms of their (dis)similarity or other dimensions of interest.

One challenge we discovered early on is transforming physical LEGO ducks into computational representations. Standard methods of image recognition ran into interesting issues that are beyond the scope of this paper. We also considered but early on discarded the use of digital LEGO construction tools. Not only are physical LEGO bricks more accessible and familiar: research suggests that physical tools afford forms of embodied creative cognition that their digital remediations can constrain (Dove et al. 2017). We reasoned that this physical-to-data transformation poses a general challenge for in-the-wild automated evaluation of 'born analog' creative works. Hence, we began exploring potential ways of transforming large sets of physical works – LEGO ducks – into data representations that lend themselves to automated creativity assessment.

1.3 Contribution & Structure of this Paper

In this paper, we present one fast and frugal method for transforming a small corpus of physical creative works into a phenotypical landscape and individual novelty metrics using STDIM-VAEs, formalising novelty as corpus-relative dissimilarity. Our method generated plausible scores and clusters of human-meaningful similarity for our LEGO duck pool from raw pixels of simple mobile phone photos. This is particularly surprising, as existing methods for automatically assessing novelty rely on well-structured data sets in which human algorithm designers pre-specified likely meaningful dimensions (Grace et al. 2015; Pérez Y Pérez et al. 2011; Karampiperis, Koukourikos, and Koliopoulou 2014; Elgammal and Saleh 2015; Zhu, Xu, and Khot 2009; Correia et al. 2019).

We will first present the Duck Task and how we generated a diverse set of human-made LEGO ducks and photographed them. We will then present the computational architecture and methods we used to pre-process images, evaluate duck novelty, and generated a phenotypical landscape of the total corpus. Finally, we present our results and discuss their ramifications in light of the existing literature.

2 Lego Duck Data Set

A corpus of trial ducks was generated by passing members of the public during an open science event at Aarhus University. Sealed packs of the six bricks were piled on a tabletop under a large sign saying 'Build a duck for science!', displaying a yellow rubber duck but no finished LEGO ducks to avoid constraint by example. Participants who approached the booth were given a brief verbal introduction to the purpose and asked to build a duck in whatever way they saw fit, with various tweaks to the patter throughout the day: sometimes encouraging builders to come up with something new, and many told to 'just have fun'. Only after building a duck were participants invited to proceed to the back of the booth, where previous ducks were on display.

3D photography was conducted using an iPhone with the *Foldio 360 app* (orangemonkey.com/app) attached in place on a *Foldio 2* photography light tent, with a *Foldio 360* turntable for rotating ducks. Each duck photographed was stored as 36 individual jpg files, and as a 3D rotation video in .GIF and .mp4 formats. Later in the process, this was reduced to 24 pictures. 518 ducks were 3D photographed in the initial stages of the project, in groups of 169, 162 and 187 observations. The second group (162) was chosen for the study in this paper.[1]

3 Methodology

3.1 STDIM-VAE

To assess the (dis)similarity of LEGO duck models, we compressed the high-dimensional image data into low-dimensional representations with the STDIM-VAE hybrid encoder (Ferguson et al. 2020). This combines features of a Spatiotemporal DeepInfomax (ST-DIM) (Anand et al. 2019) and a Variational Autoencoder (VAE) (Kingma and Welling 2013). The architecture for the hybrid encoder is shown in Figure 1.

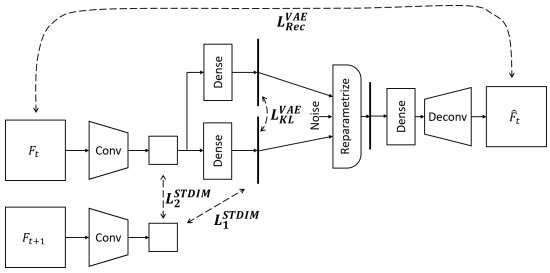

Figure 1: STDIM-VAE Network Architecture

As this encoder is designed for video data, the duck rotation videos were used, which stitch together the available images of each duck at different angles to generate a video with the camera 'circling' the duck. This is key since while the VAE section aims to purely encode visual information frame by frame, the ST-DIM section aims to learn a representation that maintains high mutual information between local features in the same spatial location in sequential frames, as well as between the global features and all the next sequential frame's local features.

3.2 Image Pre-processing

Before feeding video frames into the network to train the encoder, we pre-processed them to generate a mask for each duck that would align images in each frame and reduce their size. Smaller images reduced the compute requirement, while alignment ensured dissimilarity is not reported due to shifts in the model placement. To generate masks, we first extracted edges using color-based edge detection (Chen and Chen 2010) and thresholding. We then applied Connected Component Labeling (CCL) (Wu, Otoo, and Shoshani 2005) to label each region. As the duck is fully contained within the image, any region that touches the side of the image can be discarded. A smoothing filter was then passed over the image along with a second round of thresholding. Next, we relabelled the resulting image. We then took the largest labelled region as the mask and: calculated its centre of mass,

[1]Data set available at https://osf.io/73kv2/.

recorded the distance from the centre to each mask edge, and determined a bounding box for the whole image set based on the maximum distance in each direction. This box was then used to crop each image around their centre point.

3.3 Evaluating (Dis)similarity

We combined three approaches to evaluate the (dis)similarity of LEGO duck models. Firstly, we used Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018) to project representations down into a 2D space. We then used hierarchical clustering with cosine distance over the concatenated representation of each frame. The extracted clusters were visually examined to access the similarity of ducks within the same cluster. Finally, each duck was ranked based on the average distance and cluster size (from hierarchical clustering). This allowed us to examine a small set of models that the system believes are either common or novel.

4 Results & Discussion

The projection generated by UMAP can be seen in Figure 2, where the point color represents the cluster label generated by hierarchical clustering.[2] Seven regions have been highlighted. While not strictly abiding by the labels from the hierarchical clustering, they visually appear to group. Visual inspection suggested that models within the same region indeed shared common traits to the human eye.

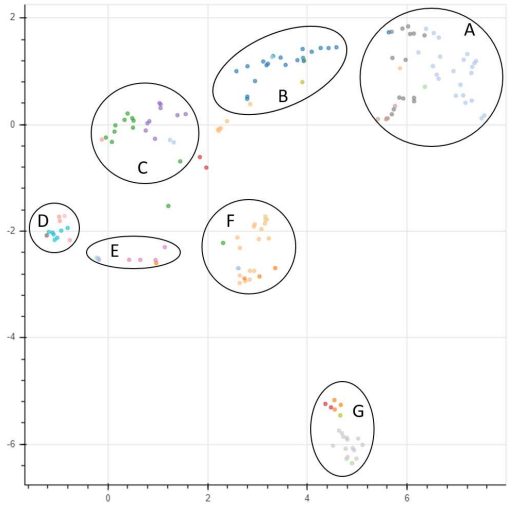

Figure 2: UMAP of Lego duck representations

For ducks in region A (e.g. Duck 0,1), it appears that generally ducks have attempted to model the whole duck. Additionally, four parts of a duck's anatomy are commonly created in the same manner. The head is commonly created by combining a 2×2 yellow brick on top of a 2×3 red plate. The tail is generally modeled by placing a 1×2 brick on the back of a 2×4 brick. The legs are commonly constructed by attaching a 2×2 brick under a 2×4 brick, although the position this attachment occurs changes within the region. Finally, a red 2×3 plate was often used to create the duck's feet. In comparison, models within region C (e.g. Duck 97,109) appear to focus on modelling the ducks head, where it is common to use two 2×3 red plates for the duck's bill. For region B, there appears to be no one common feature. Instead, the region seems to act as a transition between regions A and C. In region D (e.g. Duck 135,147), a duck's wings are often represented by two 2×3 red plates on top of a 2×4 brick, although slight variations exist in how the duck's head and tail are created. In the final two regions, F (e.g. Duck 34,57) and G (e.g. Duck 45,75), the duck's head and feet are created in a similar manner to region A. However, models in regions F and G often differ based on the height of ducks. In the F region, the models normally have a maximum height of 3 bricks and 2 plates, whereas models in region G generally have a maximum height of 4 bricks and 1 or 2 plates.

When we visually examined the twenty clusters identified from hierarchical clustering, many of the clusters could indeed fit into one of seven regions previously discussed. Whenever a cluster did not clearly fit into one region, the cluster was normally small and contained models that visually looked novel.

Finally, the ducks were ranked based on average distance to all other ducks and the size of the cluster. Common ducks should have a low average distance and large cluster size, whereas novel ducks should have a high average distance and small cluster size. The top two models for each of these are shown in Figure 3.

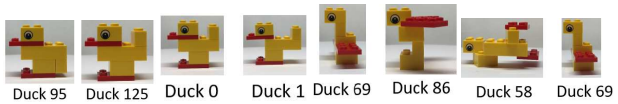

Figure 3: Common ducks identified with (a) small average distances and (b) large cluster sizes. Novel ducks identified with (c) large average distances and (d) small cluster sizes

We assume our method produced plausible results at least partially because the possibility space of creative works (LEGO ducks) and resulting pixel distributions is well-constrained and features inherent segmentation in the shape of LEGO blocks. That is, the informational properties of our raw data offloaded some of the 'heavy lifting' of recognising meaningful (dis)similarities, which human algorithm designers may need to do with more diverse corpora by pre-specifying semantically rich features or dimensions. This was further aided by our image pre-processing, cropping and aligning all image data. Any remaining dissimilarities were likely to be 'inherently' meaningful and structured. Hence, we would not expect our method to easily generalise to more inherently diverse corpora like e.g. 'construct any entity from any kind and number of LEGO bricks'. However, for our purposes of developing an automated creativity evaluator of a scalable human creativity assessment – the Duck Task – the initial results are encouraging.

[2]Find an interactive version at https://osf.io/kyzum/.

5 Conclusions & Future Work

The main aim of this work was to test the viability of a STDIM-VAE on photo imagery to create a representation that allows the easy creativity evaluation of products of the Duck Task human creativity assessment. The UMAP projection generated from our data shows that the representation indeed encapsulates human-legible feature differences in duck models, such as using particular bricks/plates to create particular parts of duck anatomy, or modelling the whole duck vs. just the head. Ranking ducks on two dissimilarity metrics generated a ranking topped by ducks that appeared on first inspection to be novel.

To validate our findings, future work will compare these to rankings by human expert evaluators. Further experiments on unused datasets would allow to test the replicability of our method, including evaluating the trained encoder on duck models unseen during training. Additionally, by sweeping all starting points to find the best match, the assumption of temporal alignment can be removed. Finally, exploring the generalisability of our approach to more diverse corpora is an interesting area for future work.

Acknowledgments

This work was supported by Microsoft Research PhD Scholarship Programme, the Digital Creativity Labs funded by EPSRC/AHRC/Innovate UK (EP/M023265/1), and the PLAYTrack project, supported by a research grant from the LEGO Foundation.

References

Anand, A.; Racah, E.; Ozair, S.; Bengio, Y.; Côté, M.-A.; and Hjelm, R. D. 2019. Unsupervised state representation learning in atari. *arXiv preprint arXiv:1906.08226*.

Chen, X., and Chen, H. 2010. A novel color edge detection algorithm in rgb color space. In *IEEE 10th International Conference On Signal Processing Proceedings*, 793–796. IEEE.

Colton, S., and Wiggins, G. A. 2012. Computational creativity: The final frontier? *Frontiers in Artificial Intelligence and Applications* 242:21–26.

Correia, J.; Machado, P.; Romero, J.; Martins, P.; and Amílcar Cardoso, F. 2019. Breaking the Mould: An Evolutionary Quest for Innovation Through Style Change. In Veale, T., and Cardoso, F. A., eds., *Computational Creativity*. Springer. 353–398.

Dove, G.; Biskjaer, M. M.; Lundqvist, C.; Olesen, J. F.; and Halskov, K. 2017. Constraints and ambiguity: Some design strategies for supporting small-scale creativity in the classroom. In *Proceedings of European Conference on Cognitive Ergonomics 2017*, 69–76. ACM.

Elgammal, A., and Saleh, B. 2015. Quantifying Creativity in Art Networks. In *Proceedings of the Sixth International Conference on Computational Creativity*, 39–46.

Ferguson, M.; Devlin, S.; Kudenko, D.; and Walker, J. 2020. Player style clustering without game variables. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*.

Grace, K.; Maher, M. L.; Fisher, D.; and Brady, K. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3(3-4):125–147.

Jordanous, A. 2019. Evaluating evaluation: Assessing progress in computational creativity research. In Veale, T., and Cardoso, F. A., eds., *Computational Creativity*. Springer. 102–107.

Karampiperis, P.; Koukourikos, A.; and Koliopoulou, E. 2014. Towards machines for measuring creativity: The use of computational tools in storytelling activities. *Proceedings - IEEE 14th International Conference on Advanced Learning Technologies, ICALT 2014* 508–512.

Kaufman, J. C.; Plucker, J. A.; and Baer, J. 2008. *Essentials of creativity assessment*. Hoboken, NJ: John Wiley & Sons.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lamb, C.; Brown, D. G.; and Clarke, C. L. 2018. Evaluating computational creativity: An interdisciplinary tutorial. *ACM Computing Surveys* 51(2).

Manovich, L. 2016. The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. *Journal of Cultural Analytics* 1(1):1–15.

McInnes, L.; Healy, J.; Saul, N.; and Grossberger, L. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software* 3(29):861.

Pérez Y Pérez, R.; Ortiz, O.; Luna, W.; Negrete, S.; Castellanos, V.; Peñalosa, E.; and Ávila, R. 2011. A system for evaluating novelty in computer generated narratives. In *Proceedings of the 2nd International Conference on Computational Creativity, ICCC 2011*, 63–68.

Plucker, J. A.; Beghetto, G. T.; and Dow, R. A. 2004. Why Isn't Creativity More Important to Educational Psychologists? Potentials, Pitfalls, and Future Directions in Creativity Research. *Educational Psychologist* 39(2):83–96.

Plucker, J. A.; Makel, M. C.; and Qian, M. 2019. Assessment of Creativity. In Kaufman, J. C., and Sternberg, R., eds., *The Cambridge Handbook of Creativity*. Cambridge: Cambridge University Press, 2nd edition. 44–68.

Summerville, A. 2018. Expanding expressive range: Evaluation methodologies for procedural content generation. In *Proceedings of the 14th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE 2018*, 116–122.

Wu, K.; Otoo, E.; and Shoshani, A. 2005. Optimizing connected component labeling algorithms. In *Medical Imaging 2005: Image Processing*, volume 5747, 1965–1976. International Society for Optics and Photonics.

Zhu, X.; Xu, Z.; and Khot, T. 2009. How creative is your writing? A linguistic creativity measure from computer science and cognitive psychology perspectives. In *Proceedings ofthe NAACL HLT Workshop on Computational Approaches to Linguistic Creativity*, 87–93.