Towards Co-Creative Drawing Based on Contrastive Language-Image Models

Francisco Ibarrola
School of Architecture, Design and Planning
The University of Sydney
Sydney, Australia
francisco.ibarrola@sydney.edu.au

Oliver Bown
Interactive Media Lab
University of New South Wales
Sydney, Australia
o.bown@unsw.edu.au

Kazjon Grace
School of Architecture, Design and Planning
The University of Sydney
Sydney, Australia
kazjon.grace@sydney.edu.au

Abstract

Recent advances in generative machine learning, particularly in the area of text-to-image synthesis, have created huge potential for co-creative systems. It is non-trivial, however, to adapt algorithms intended to generate images that match a given prompt to suit the task of effective collaboration with humans. This paper presents initial experimentation towards developing an agent that can work cooperatively with a human designer in the task of drawing. We do so by utilizing Contrastive Language Image Pretraining (CLIP) to guide the drawing's semantic meaning on a drawing completion process, and fidelity terms to enforce geometric alignment (with what would be the user's in-progress sketch). Preliminary results are presented as a proof of concept, attesting that drawing outputs are both diverse and identifiable as matching the provided prompt, which we interpret as steps towards co-creativity.

Introduction

The traditional conception of the role of a computer within a creative process is that of a tool: precise, effective, and unobtrusive. But today's AI-driven capabilities have started to bend the barrier between tools and collaborators, as reflected on recent studies in Human-Computer Co-Creative Processes (Kantosalo et al. 2020). In this work, we seek to test that barrier further, exploring how generative AI models can be applied to develop co-creative systems that can help designers sketch. There have been many amazing sketching systems developed in the last decade (Davis et al. 2016; Karimi et al. 2018), but several questions remain unanswered before those systems could be applied in practice: Can a co-creative sketching system work towards a user-specified goal? Can it both respect the user's progress and propose modifications when needed? Can it propose small, diverse steps towards completion rather than one-shot autocomplete a drawing? We tackle the task of building a co-creative agent that can answer some of these questions in the affirmative.

For a co-creative drawing agent to be able to be truly cooperative in this context, it should not only be able to pick up on a partial design made by the user, but also to somewhat grasp a sense of its semantic meaning, and produce an output consistent with the user's end goal. Until recently, image generation models were only capable of producing outputs based on simple, specifically trained conditioning labels (Mirza and Osindero 2014). But there has been rapid recent progress in context-agnostic models trained in huge datasets, that have an ability to translate the meaning of complete sentences into images, such as CLIPDraw (Frans, Soros, and Witkowski 2021), and Dall-E (Ramesh et al. 2021).

Our goal in this work is to make progress towards systems with which users can engage in dialogue during creative collaboration, fluidly discussing both strategic and tactical aspects of their emerging design (Bown et al. 2020).

We present a co-creative drawing model that can complete a user's design by taking advantage of CLIPDraw's ability to produce drawings aligned with the semantic meaning of the desired output's description. To this we add loss terms for building on a user's partial sketch and penalising drawing outside a (user-specified) region (see Figure 1).

Related Work

Our co-drawing model builds on CLIPDraw (Frans, Soros, and Witkowski 2021), which built on CLIP (Ramesh et al. 2021), which in turn built on ConVIRT (Zhang et al. 2020).

CLIP and ConVIRT

Contrastive training is based on the following idea: let us consider a batch of (text, image) pairs, $\{(t_n, I_n)\}_{n=1,\dots N}$, paired in the sense that the text t_n describes the image I_n. Then, two functions g and f mapping text and images (respectively) to a latent space $\mathbb{R}^D$ are built using appropriately chosen Neural Network (NN) architectures. These functions are trained to minimize a loss function based on the cosine distance between the image and the text (and vice versa).

$$L_c(t_k, I_k) \doteq -\log \frac{\exp \langle g(t_k), f(I_k) \rangle / \tau}{\sum_{n=1}^{N} \exp \langle g(t_k), f(I_n) \rangle / \tau},$$

where $\langle \cdot, \cdot \rangle$ denotes the cosine similarity and $\tau > 0$ is a scale parameter. Finding g and f minimizing L_c essentially means we are fitting g and f so that t_k is mapped closer to I_k than any other image on the batch. The same is done, using a complementary loss function, to ensure $f(I_k)$ is closer to $g(t_k)$ than to the mapping of any other text within the batch. The result is a shared embedding of both images and text prompts into a space where similarity can be measured between any combination of either. As soon as it was released,

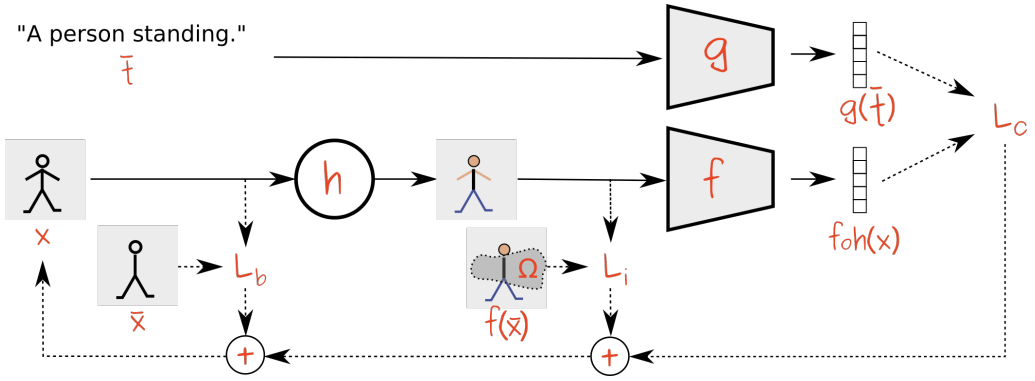

Figure 1: Co-Drawing Model schema: Three different losses are computed on three instances of the pipeline, and the set of Bézier curves x that defines the drawing is optimized with respect to the sum. This ensures parametric similarity with the given curve set $\bar{x}$, consistency with the partial drawing $h(\bar{x})$ and compliance with the semantic meaning $g(\bar{t})$.

CLIP became the focus of a vital and diverse community of online artistic exploration of its capabilities. Much of this exploration was based around generating images from a GAN that match a particular prompt (Liu and Chilton 2021).

CLIPDraw

Of most interest to our goal of co-creative drawing is the recent coupling of the CLIP-based semantic loss (i.e. match to a provided prompt) with a differentiable rasteriser (Li et al. 2020). The resulting system, CLIPDraw, generates an image that fits a provided prompt by manipulating a set of Bézier curves (Frans, Soros, and Witkowski 2021).

Let us denote by $\mathcal{B}$ the space of Bézier curves and let $h : \mathcal{X} \to \mathbb{I}$ be the aforementioned differentiable function that maps the set of finite subsets of $\mathcal{B}$ to the image space. Then, given a set x of Bézier curves, it is possible to build a gradient descent optimization method as:

$$x \leftarrow x + \eta \nabla_x \langle g(t), f \circ h(x) \rangle, \tag{1}$$

where $\eta > 0$ is the learning step.

Put simply, this lets us find a vector drawing that matches a given text prompt, thus enforcing semantic meaning to our model's outputs.

Co-Creative Drawing

Key to co-creative drawing is modifying existing partial sketches. A first instinct upon seeing how CLIPDraw works might be to just let it draw over our partially completed design, since a simple sum over $h(x)$ would preserve the model's differentiability. There are two issues with this approach. Firstly: CLIPDraw can (and often will) simply draw over the partial drawing, completely disregarding the user's design. Secondly, the opposite is also a problem: if the agent is prevented from making any adjustments to the user's input, then it becomes inflexible.

With this in mind, we start by formally defining our partial sketch as a set of Bézier curves $\bar{x} \in \mathcal{X}$, and a text prompt $\bar{t}$ as a string describing the desired end result of our drawing. In practice this partial drawing would be something like an SVG image created by a user.

Curve Fidelity

Let us denote by K_0 the number of Bézier curves in $\bar{x}$ and by K_a the number of additional curves we are going to allow the agent to draw. Finally, let x be the variable associated to the total set of $K = K_0 + K_a$ curves in the model. The idea is that the first K_0 curves produced by the method resemble those of the provided sketch, and we can enforce that by adding the following term to the cost function:

$$L_b(x, \bar{x}) \doteq \sum_{k=1}^{K_0} \sum_{m=1}^{3} \lambda_m \|\bar{x}_k^{(m)} - x_k^{(m)}\|^2, \tag{2}$$

where the index $m = 1, \ldots, 3$ represents one of three variable types: color, coordinates or width, and $\lambda_m > 0$ are regularization parameters, dependant on the type of variable. More specifically, $x_k^{(1)} \in \mathbb{R}^{D_k}$ is a vector containing the path coordinates, $x_k^{(2)} \in [0, 1]^4$ is a vector with the RGBA components of the color of the trace, and $x_k^{(3)} > 0$ represents the width of the trace.

By using this penalisation term, we enforce x to keep the original traces from the partial sketch. Furthermore, by tuning the λ_m parameters, we can control the strength of this constraint, setting large values to strictly maintain the original traces, and smaller values to allow the agent to sensibly move, adjust the width or change the color of the traces.

Drawing within a specified region

Despite the above constraints that enforce similarity on the curves, our agent might still "choose" to draw over the user's partial sketch. To overcome this, we define a region Ω of the canvas where the agent is allowed to draw, by penalizing image discrepancies outside of it. In practice we envisage that this could be provided by the user, or potentially suggested automatically through a process analogous to neural attention.

Notice we want to penalize discrepancies, but not prohibit them. Breaking the rules should always be possible during creative processes, if there is a good reason to do so. To

Figure 2: On the left, a user's partial sketch $\bar{x}$. After that, the outputs $h(\hat{x})$ obtained with three random initializations of additional Bézier curves, using the prompt $\bar{t}$ ="A red chair" and the drawing area Ω set as the top half of the canvas.

Algorithm 1 Co-Creative Drawing

Set $x_k = \bar{x}_k,\ \forall k = 1, \ldots, K_0$.
Let $x_k \sim \mathcal{U}[0,1],\ \forall k = K_0 + 1, \ldots, K$.
Establish a drawing region Ω.
while $\|h(x) - h(x_{(p)})\|_F^2 > \delta$
  $x_{(p)} \leftarrow x$
  $x \leftarrow x - \eta \nabla_x L(x; \bar{t}, \bar{x})$
return $h(x)$

accomplish this we define an additional cost function as:

$$L_i(x, \bar{x}) \doteq \alpha \|h(\bar{x}) - h(x)\|^2_{L^2(\Omega^c)}, \tag{3}$$

where $\alpha > 0$ is a regularisation parameter, and $\|\cdot\|_{L^2(\Omega^c)}$ is the L^2 norm defined in the complement of the drawing region Ω.[1] Here again, the fidelity of the image outside the designated drawing area can be enforced (or relaxed) by increasing (or decreasing) the value of α.

Algorithm

Finally, we can add the terms on (2) and (3) to the cosine distance (see Figure 1) to build our overall cost function as

$$L(x; \bar{t}, \bar{x}) \doteq -\langle g(\bar{t}), f \circ h(x) \rangle + L_b(x, \bar{x}) + L_i(x, \bar{x}).$$

The goal is now to find a solution $\hat{x}$ minimizing L. Even though differentiable, L is non-convex and hence finding a global minimum is an intractable problem. Nonetheless, we have found using a gradient descent approach such as (1) often yields good results in practice, and hence we propose to use the method summarised in Algorithm 1.

While the existence of local minima is considered a problem in most settings, it is the opposite here. A high-quality solution $\hat{x}$ within our framework can be understood as one with a low value $L(\hat{x}; \bar{t}, \bar{x})$, while a set of diverse solutions corresponds to a set of elements within different regions of $\mathcal{X}$. This means that the set of highest-quality diverse solutions is a set of local minimizers, and hence a subset of the possible convergence points of the proposed algorithm.

Results

Although creativity is a very tricky concept to define, let alone measure, there is certain consensus on the conjunction of value/utility/quality and novelty/originality/diversity being a good approximation to assess it (McCormack and Gambardella 2022). Both dimensions, however, have their own subjectivity, so we attempt to operationalise them in ways that make sense for co-creative drawing.

As a first intuitive test of our method, we drew a partial sketch, defined a very simple drawing region, ran Algorithm 1 and inspected the outputs (see Figure 2). These images were obtained by providing the agent with a sketch of a stool, and asking it to draw on the top half of the canvas to match the description "A red chair". As a first-order measure of quality: if you the reader were readily able to recognise the drawings as red chairs without reading the caption, then we can attest some subjective standard of quality. Some scribbles appear in the background, which are a consequence of the original CLIP having been trained mostly with natural images, with complexly textured backgrounds. Even ignoring the scribbles, there is also (again, naively) some degree of diversity present among the four chairs, for example in their orientations or the height of their backs.

Quality Assessment

As a simple yet robust way of assessing the quality of the outputs we checked whether CLIP itself recognises the generated drawings as matching the prompt. CLIP can be used as a classifier, with 2343 nouns from its corpus as labels.[2] Evaluating 100 samples from the tasks in Figs 2 and 3 account for a 98% recognition rate for the categories "chair" and "hat", with a confidence of $69.9\% \pm 19.6\%$. This accuracy and confidence (estimated as a softmax probability over the very large set of labels) is quite encouraging as a first assessment: our drawings are at least recognisable as the objects they are supposed to be.

Diversity Assessment

Quantifying diversity is yet another task without a standardised method, but recent papers (McCormack and Gambardella 2022) aim to measure it using the intermediate layers of Convolutional Neural Networks (CNNs). It has been shown that different layers encode geometric properties at different scales, which can capture the "style" of images (Ecker, Bethge, and Gatys 2015). Bearing this in mind, we

[1] Better, more complex penalisation functions may be feasible and will be explored in future work.

[2] Ideally, we would want to use a different NN architecture, but to the best of our knowledge, CLIP is the most complete domain-agnostic image classifier currently available.

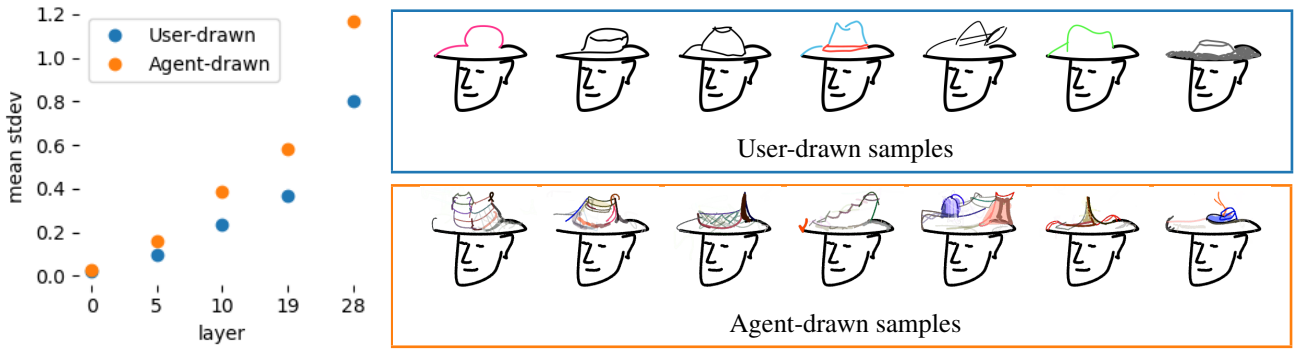

Figure 3: On the left, the mean standard variation over each layer's neuron activations for the 10 tested samples. On the right, some samples of the hat-design task outputs as completed by the users and the agent.

propose to use the variability of the activation of intermediate CNN layers as a measure of diversity.

We provided 10 human subjects with the same partial sketch of a person wearing a hat, and asked them to complete the design as they wish (some samples can be seen in Figure 3). We then put the images through the CNN proposed in (Simonyan and Zisserman 2014) and got the outputs of the five intermediate layers used in Style Transfer. We computed the standard deviation over the 10 samples for every neuron, and averaged over every layer, getting five points of comparison. We then did the same with 10 randomly generated samples from our model. Comparing the two sets (see Figure 3) shows that our generated samples have a higher variance. Although we cannot assure how well these measurements align with our intuitive notion of diversity, the results do suggest at least comparable, if not higher than inter-human design diversity in our results. Of course, this small-scale study has limitations: we neither asked our human subjects to be diverse nor did we recruit skilled milliners to design for us.

Conclusions

We have introduced a model intended for a designer to interact with a sketch-generation agent. Preliminary quantitative results account for the model being capable of producing diverse and quality drawings. Qualitatively, the process and its outputs show potential as a useful fit for co-creative drawing.

The proposed idea is flexible enough to explore the use of other image generative models as the core of the co-creative agent. Future work shall also deal with the formalization and expansion of the introduced experimental setting.

Author Contributions

F. Ibarrola was in charge of defining the cost functions, writing the code, and drafting the paper, and participated in experiment design. O. Bown proposed the initial ideas and did the final reviews. K. Grace was in charge of guidance and editing, and contributed to the model and experiments designs. All the authors participated in the ideas discussions.

Acknowledgments

We would like to acknowledge the Australian Research Council for funding this research (ID #DP200101059).

References

Bown, O.; Grace, K.; Bray, L.; and Ventura, D. 2020. A speculative exploration of the role of dialogue in human-computerco-creation. In *ICCC*, 25–32.

Davis, N.; Hsiao, C.-P.; Yashraj Singh, K.; Li, L.; and Magerko, B. 2016. Empirically studying participatory sense-making in abstract drawing with a co-creative cognitive agent. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, 196–207.

Ecker, A.; Bethge, A.; and Gatys, L. 2015. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Frans, K.; Soros, L.; and Witkowski, O. 2021. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *arXiv preprint arXiv:2106.14843*.

Kantosalo, A.; Ravikumar, P. T.; Grace, K.; and Takala, T. 2020. Modalities, Styles and Strategies: an interaction framework for human-computer co-creativity. In *ICCC*, 57–64.

Karimi, P.; Grace, K.; Davis, N.; and Maher, M. L. 2018. Creative sketching apprentice: Supporting conceptual shifts in sketch ideation. In *International Conference on-Design Computing and Cognition*, 721–738. Springer.

Li, T.-M.; Lukáč, M.; Gharbi, M.; and Ragan-Kelley, J. 2020. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics* 39(6):1–15.

Liu, V., and Chilton, L. B. 2021. Design guidelines for prompt engineering text-to-image generative models. *arXiv preprint arXiv:2109.06977*.

McCormack, J., and Gambardella, C. C. 2022. Quality-diversity for aesthetic evolution. *arXiv preprint arXiv:2202.01961*.

Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2020. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*.