TOWARDS A PRINCIPLED EVALUATION OF MACHINE-GENERATED ART

Lia Coleman\textsuperscript{1}, Panos Achlioptas\textsuperscript{2}, Mohamed Elhoseiny\textsuperscript{1,2}

\textsuperscript{1} KAUST \textsuperscript{2} Stanford University
lia.coleman@kaust.edu.sa; optas@cs.stanford.edu; mohamed.elhoseiny@kaust.edu.sa


MOTIVATION

Do artists like AI generated “paintings” and can they tell them apart from human-made ones?

How are their responses to the above questions different from those given by non-experts?

METHODS

\begin{itemize}
  \item We trained a CAN model \cite{95, 67} on the WikiArt dataset using multiclass cross entropy for our loss as in \cite{67}.
  \item Human evaluation study on 120 images: half synthetic generations; half human-art.
  \item Ask a group of 13 professional artists: Rate on a scale of 1-5 how much you like this work. Do you think this artwork was created by a human or generated by an AI?
  \item 5 responses per image from Amazon Mechanical Turk raters.
  \item In evaluation, we simulate real-world evaluation conditions: Split the data into two sets: the ‘seen’ set, and the ‘unseen’ set. Our evaluation procedure is:
    \begin{enumerate}
      \item From the ‘seen’ set, calculate the Cohen’s Kappa (a measure of similarity) between each turker’s likeability responses and the artist majority vote on likeability.
      \item Choose a Cohen’s Kappa threshold to exclude turkers from the most.
      \item (proof-of-viability) Verify that this Cohen’s Kappa, which is derived from the ‘seen’ set, performs well on the ‘unseen’ set.
    \end{enumerate}
\end{itemize}

RESULTS

\begin{itemize}
  \item Our best machine-generated image is on par with the 3rd best human-created artwork, both attaining 75\% of artists’ votes.
  \item We can use abundant Amazon MTurk labels to supplement the scarce labels of artists. Using Cohen’s Kappa as a cutoff threshold for Turkers’ labels, we get a large gain in Cohen’s Kappa agreement with artists.
  \item Some aspects of likability in art are shared and learnable. A linear SVM operating on VGG16 image features achieves 78.2\% ± 0.4\% accuracy in likability prediction, while (educated) random guessing achieves 70.8\%.
\end{itemize}

CONCLUSION

\begin{itemize}
  \item Our analysis highlights the importance of considering artists’ opinion when evaluating AI generated art.
  \item Using MTurk to help scale artist responses on likability for unseen images is a viable approach.
  \item Preliminary results that likability of artwork is learnable.
\end{itemize}

REFERENCES

\cite{1, 2, 3, 4, 5, 6, 7, 8, 9}

Are you an ARTIST, DESIGNER or an ML RESEARCHER interested in art or design?

We value your input in measuring progress on AI generated art. If you are interested in helping us with your feedback, please fill out this form.

https://tinyurl.com/v5kapde