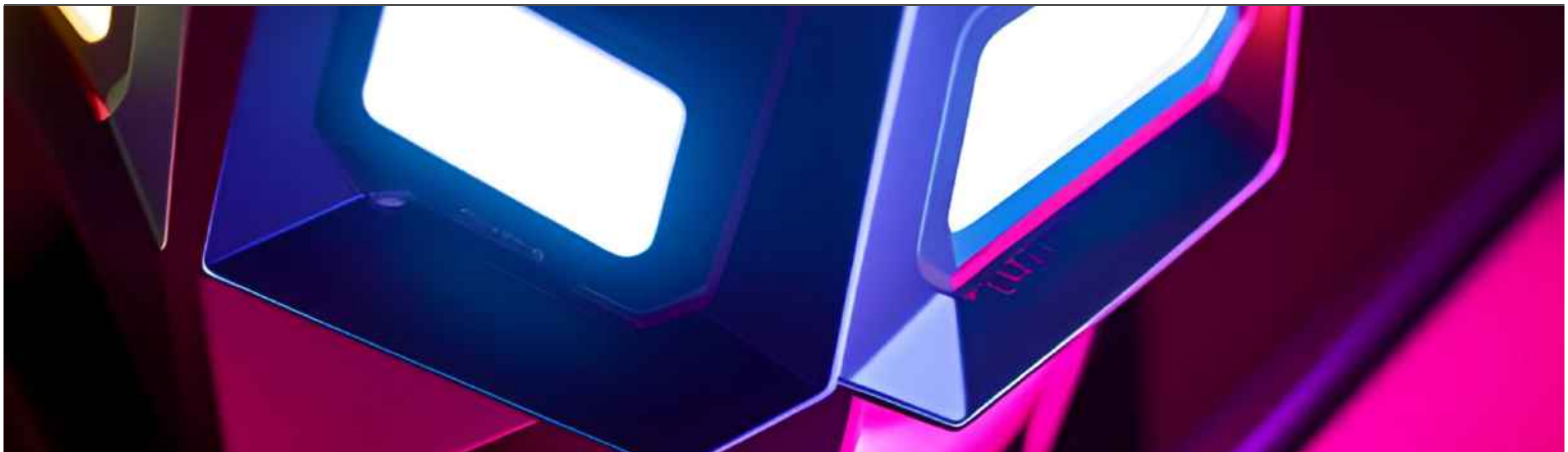
Synced



AI MACHINE LEARNING & DATA SCIENCE RESEARCH

Genius or Subpar AI Mathematician? New Study Questions ChatGPT's Mathematical Capabilities

BY SYNCED
2023-02-03

COMMENTS 3

The November release of ChatGPT garnered unprecedented public and media attention. OpenAI's conversational large language model (LLM) was widely applauded for its ability to answer complex queries, generate correct computer code and coherent long-form essays, and even solve math problems. But might that last claim have been premature?

In the new paper *Mathematical Capabilities of ChatGPT*, a research team from the University of Oxford, TU Wein, University of Cambridge, University of Vienna, and Princeton University tests ChatGPT's mathematical capabilities on publicly available and hand-crafted datasets and evaluates its suitability as an assistant to professional mathematicians. The team concludes that despite the glowing media reviews, ChatGPT's mathematical abilities "are significantly below those of an average mathematics graduate student."

Mathematical Capabilities of ChatGPT

Simon Frieder*[1,4], Luca Pinchetti[1], Ryan-Rhys Griffiths[3], Tommaso Salvatori[2], Thomas Lukasiewicz[1,2], Philipp Christian Petersen[4,5], Alexis Chevalier[6], and Julius Berner[4]

[1]Department of Computer Science, University of Oxford, Oxford, UK
[2]Institute of Logic and Computation, TU Wien, Vienna, Austria
[3]Department of Physics, University of Cambridge, Cambridge, UK
[4]Faculty of Mathematics, University of Vienna, Vienna, Austria
[5]Research Network Data Science, University of Vienna, Vienna, Austria
[6]School of Mathematics, Institute for Advanced Study, Princeton, US

February 1, 2023

The team summarizes their main contributions as follows:

1. Insight for mathematical use is provided. We show for which types of questions and which domains of mathematics ChatGPT may be useful and how it could be integrated into the workflow of a mathematician.
2. The failure modes of ChatGPT are identified, as well as the limits of its capabilities. This can aid future efforts to develop LLMs that perform better in mathematics.
3. We provide benchmarks for testing the mathematical capabilities of future LLMs so that they can be compared to ChatGPT across a range of aspects regarding advanced mathematical comprehension.

Dataset name	Comprised of the file(s)	Tags
Grad-Text	W. Rudin, Functional Analysis (ch. 1)	M3 Q4
	W. Rudin, Functional Analysis (ch. 2)	M3 Q4
	J. Munkres, Topology (ch. 1)	M3 Q4
	J. Munkres, Topology (ch. 2)	M3 Q4
	R. Durret, Probability Theory	M3 Q4
Holes-in-Proofs	Proofs Collection A	M3 Q2 Q5
	Proofs Collection B Prealgebra	M1 Q5
	Proofs Collection B Precalculus	M1 Q5
Olympiad-Problem-Solving	Olympiad Problem Solving	M4 Q4 D2
Symbolic-Integration	Symbolic Integration	M2 Q3 D1
MATH	MATH Algebra	M1 M2 M3 Q3 Q4
	MATH Counting and Probability	M1 M2 M3 Q3 Q4
	MATH Prealgebra	M1 Q3 Q4
	MATH Precalculus	M1 Q3 Q4
Search-Engine-Aspects	Definition Retrieval	M3 Q1 Q2 D3
	Reverse Definition Retrieval	M3 Q2 D3
	Named Theorem Proof Completion	M3 Q1 Q2 D3

Table 1: A summary of all datasets, together with their associated tags. The tags Mi, Qi, and Di relate to the level of Mathematical difficulty, the Question type, and the Out-of-Distribution type from Section 3.3, respectively.

To effectively evaluate ChatGPT on advanced math problems, the researchers build a new dataset, GHOSTS, comprising a total of 728 prompts in six carefully crafted subdatasets: Grad-Text, Holes-in-Proofs, Olympiad-Problem-Solving, Symbolic-Integration, MATH, and Search-Engine-Aspects. The researchers say the GHOST datasets surpass publicly available benchmark mathematical datasets in terms of sophistication and reasoning difficulty.

The researchers use LaTeX to encode the mathematical inputs for most of their subdatasets, which are categorized into four dimensions with ascending difficulty: 1) elementary arithmetic problems, 2) symbolic problems, 3) (under)graduate-level exercises from well-known textbooks and questions from math.stackexchange.com, and 4) exercises in the style of Mathematical Olympiad problems.

The team applied ChatGPT on the GHOST datasets and considered output length, stability of the answer under prompt engineering, and how close they judged ChatGPT to be to the correct answer.

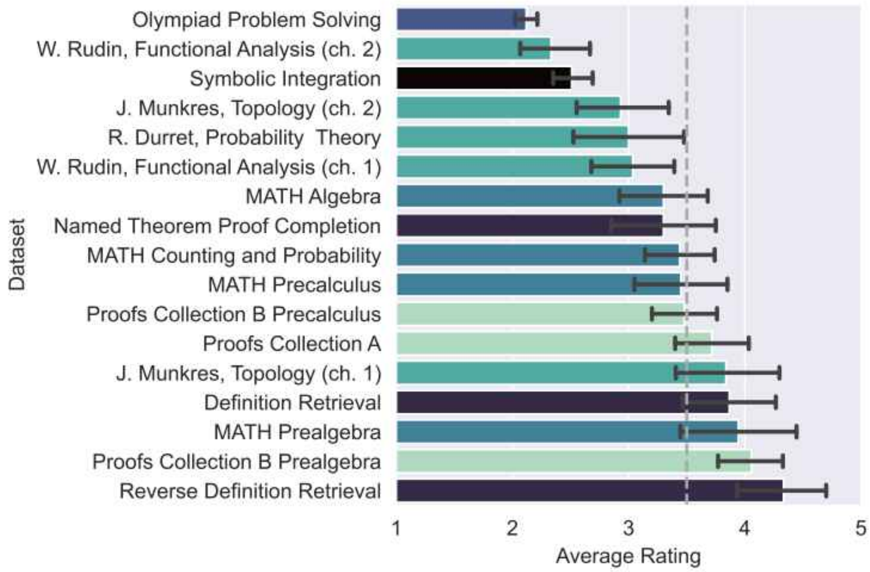


Figure 2: Average rating for each file in each subdataset (same colors indicate the files come from the same subdataset). Since the maximal ranking is 5, and the minimal ranking where the question was at least understood is 2, a passing grade (50% of points) would be 3.5, as indicated by the dotted line. The error bars represent 95% confidence intervals.

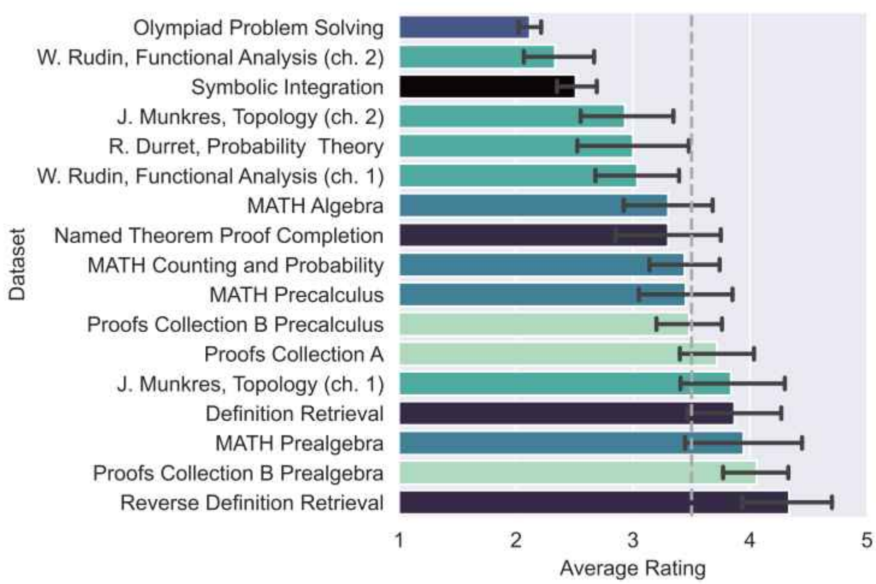


Figure 2: Average rating for each file in each subdataset (same colors indicate the files come from the same subdataset). Since the maximal ranking is 5, and the minimal ranking where the question was at least understood is 2, a passing grade (50% of points) would be 3.5, as indicated by the dotted line. The error bars represent 95% confidence intervals.

ChatGPT failed on most of the problems, faring especially poorly on questions requiring deep insights and original solutions such as those found in the Mathematical Olympiads. The paper concludes that while ChatGPT can effectively search for mathematical objects when given information about them, it struggles with advanced mathematics and delivering consistent, high-quality proofs or calculations.

The team hopes their work will inspire other professional mathematicians to contribute to building a more thorough benchmark for assessing and improving LLMs' mathematical abilities.

Just one day before this paper was published, OpenAI announced it had upgraded ChatGPT with improved mathematical capabilities. It's unclear how this latest version would perform in the experiments presented here.

The GHOSTS dataset will be released on the project's GitHub. The paper *Mathematical Capabilities of ChatGPT* is on arXiv.

Author: Hecate He | **Editor**: Michael Sarazen

We know you don't want to miss any news or research breakthroughs. **Subscribe to our popular newsletter *Synced Global AI Weekly* to get weekly AI updates.**

TAGS • AI • ARTIFICIAL INTELLIGENCE • CHAT BOT • CHATGPT • DEEP NEURAL NETWORKS • MACHINE LEARNING • ML • RESEARCH • TECHNOLOGY

ABOUT SYNCED

Machine Intelligence | Technology & Industry | Information & Analysis

---

3 COMMENTS ON "GENIUS OR SUBPAR AI MATHEMATICIAN? NEW STUDY QUESTIONS CHATGPT'S MATHEMATICAL CAPABILITIES"

Basket Random
2023-02-07

As a math fan, this topic caught my attention.

Loading...

↵Reply

Pingback: Microsoft's MathPrompter Dramatically Improves LLM Performance on Mathematical Reasoning Tasks | Synced

Darius Quid
2023-03-29

It's pretty interesting, and that's another proof that math can be actually complicated. Even Artificial Intelligence struggles with it, so I genuinely don't understand people who enjoy solving math problems, especially mathematical olympiad ones, it's way too much for me. I used to hate this class at school, and I even decided to get online math classes from brighterly math for my daughter because I know for a fact that I won't be able to explain anything to her besides the simplest basics. To be honest, I'm pretty sure that she's even better than me at math right now.

Loading...

↳Reply

LEAVE A REPLY

Comment *

Name

Email

Website

☐ Notify me of follow-up comments by email.

☐ Notify me of new posts by email.

Post Comment

PREVIOUS POST **NEXT POST**

Select Language

Search ...

Search

Sort by Relevance

Follow us on Twitter

Tweets from @Synced_Global

Synced
@Synced_Global · Apr 11

Adobe & UCL's Pix2Video: Text-Guided Video Editing via Image Diffusion Without Preprocessing or Finetuning

syncedreview.com
Adobe & UCL's Pix2Video: Text-Guided Video Editing via Image Diffusion Without Preprocessing or...

Synced
@Synced_Global · Apr 10

UC Berkeley's Instruct-NeRF2NeRF Edits 3D Scenes With Text Instructions

syncedreview.com
UC Berkeley's Instruct-NeRF2NeRF Edits 3D Scenes With Text Instructions | Synced

1

Synced
@Synced_Global · Apr 6

Google Reveals Its Latest TPU v4-Based Supercomputer, Which Betters Nvidia's A100s in Speed and Efficiency

syncedreview.com
Google Reveals Its Latest TPU v4-Based Supercomputer, Which Betters Nvidia's A100s in Speed and...

1

Synced
@Synced_Global · Apr 5

AI Needs a Therapist: Columbia U & IBM's SafeguardGPT Leverages Psychotherapy & RL to Build Healthy AI Systems

Tags

AI Artificial Intelligence Artificial Intelligence_ China computer vision Deep Learning DeepMind Deep Neural Networks Facebook Google Machine Learning ML Neural Network Neural Networks NLP NVIDIA Reinforcement Learning Research robotics technology

Archives

Select Month

FOLLOW US AND DON'T MISS ANOTHER ARTICLE

Email Address

Subscribe

Synced

GLOBAL CENTER (CANADA)

56 Temperance St, #700
Toronto, ON M5H 3V5

MASSACHUSETTS OFFICE (US)

One Broadway, 14th Floor, Cambridge, MA 02142

CALIFORNIA OFFICE (US)

75 E Santa Clara St, 6th Floor, San Jose, CA 95113

CONTACT

Contact Us @ global.general@jiqizhixin.com

Visit Us @ Synced China

Contribute to Synced Review

Synced

About Us

Founded in March 2014, *Synced*, is China's first media entity that focuses on reporting machine intelligence and related technologies. We are currently the industry's most influential content producer, principled to produce professional, authoritative, and edifying content relating to the fields of Artificial Intelligence, robotics, neuroscience, and other forefront technologies. We are dedicated to providing a platform for those who wish to learn more about cutting-edge technology and discover how to deal with the upcoming waves of technological revolution; A platform that proactively initiates discussions regarding human-technology relationship that seek to inspire imagination and creativity, meanwhile encourages readers to reflect upon the interactive future of humanity and technology.

On China's largest social network and content sharing platform, WeChat, *Synced* has more than 190,000 subscribers with more than 50,000 readership per day. It also has over 20,000 daily total page visit on other platforms such as Baidu, TouTiao.com, Netease and Sohu. *Synced* fills the gap in China's science and technology media scene. Based on subscriber and third-party feedbacks, our professionalism, practicality and thoughtfulness are widely recognized and commended by company executives, experts, scholars, well-known investors, industry practitioners, among other technology enthusiasts. We are the key content provider for *Tencent*, *Baidu*, and *TouTiao.com*, winning the title of *Huxiu*'s *Top Ten Content Provider of 2015* and *TouTiao's Best vertical-Media of 2015*. We were proud to be serving as *APEC Global VR Summit*'s chief technology media, the one and only Chinese Media partner of *AI frontier conference* in the silicon valley, and invited to report in the *Brain Forum 2016* in Switzerland as the only Chinese media.

Synced has conducted numerous in-depth interviews with industry leaders, for who provided valuable insight on industry dynamics and expertise. Past interviewees include the god father of Reinforcement Learning professor ***Richard Sutton*** from University of Alberta, chief scientist at Baidu Research Silicon Valley and adjunct professor at Stanford University ***Andrew Ng***, Professor ***Eric P. Xing*** from *Carnegie Mellon University*, Microsoft Research's chief Artificial Intelligence scientist ***Deng Li*** , Professor ***Qiang Yang*** from Hong Kong University of Science and Technology, the founders of ***MXNet***, Duke University's vice provost for research ***Lawrence Carin***, Professor ***Jack Gallant*** from UC Berkeley, Nobel Laureate ***Randy Schekman***, and so on. We have also reported on Baidu, Google, IBM, Microsoft, among other major technology companies, alongside many outstanding Artificial Intelligence companies and robotics startups around the world.

In addition to content production, *Synced* provides services to interested individuals in the artificial intelligence industry, including financial advisory, globalization strategy and talent recruitment. At the same time, *Synced* has covered young Chinese researchers in more than 40 colleges and universities in North America and Europe.

Follow us on Twitter

Tweets from @Synced_Global

Follow on Twitter

Synced @Synced_Global · Apr 25

Look Again, YOLO: Baidu's RT-DETR Detection Transformer Achieves SOTA Results on Real-Time Object Detection

syncedreview.com
Look Again, YOLO: Baidu's RT-DETR Detection Transfo...
End-to-end transformer-based object detectors (DETRs) play a crucial role in applications such as ...

Synced @Synced_Global · Apr 21

Huawei's DiffFit Unlocks the Transferability of Large Diffusion Models to New Domains

syncedreview.com
Huawei's DiffFit Unlocks the Transferability of Large D...
The astonishing performance of diffusion models on tasks such as image synthesis, video generation and ...

1

Synced @Synced_Global · Apr 19

DeepMind & MPG Establish a Research Program for Meta-Learned Models of Cognition

syncedreview.com
DeepMind & MPG Establish a Research Program for ...
Meta-learning is a machine learning approach for developing algorithms and models that learn how to ...

1

Synced @Synced_Global · Apr 19

Microsoft & Bath U's SpectFormer Significantly Improves Vision Transformers via Frequency and Attention

syncedreview.com

syncedreview.com

Tags

AI Artificial Intelligence Artificial Intelligence_ China computer vision Deep Learning DeepMind Deep Neural Networks Facebook Google Machine Learning ML Neural Network Neural Networks NLP NVIDIA Reinforcement Learning Research robotics technology

Archives

Select Month

FOLLOW US AND DON'T MISS ANOTHER ARTICLE

Email Address

Subscribe

Synced

GLOBAL CENTER (CANADA)

MASSACHUSETTS OFFICE (US)

56 Temperance St, #700
Toronto, ON M5H 3V5

One Broadway, 14th Floor, Cambridge, MA 02142

CALIFORNIA OFFICE (US)

75 E Santa Clara St, 6th Floor, San Jose, CA 95113

CONTACT

Contact Us @ global.general@jiqizhixin.com

Visit Us @ Synced China

Contribute to Synced Review

MIT Technology Review SUBSCRIBE

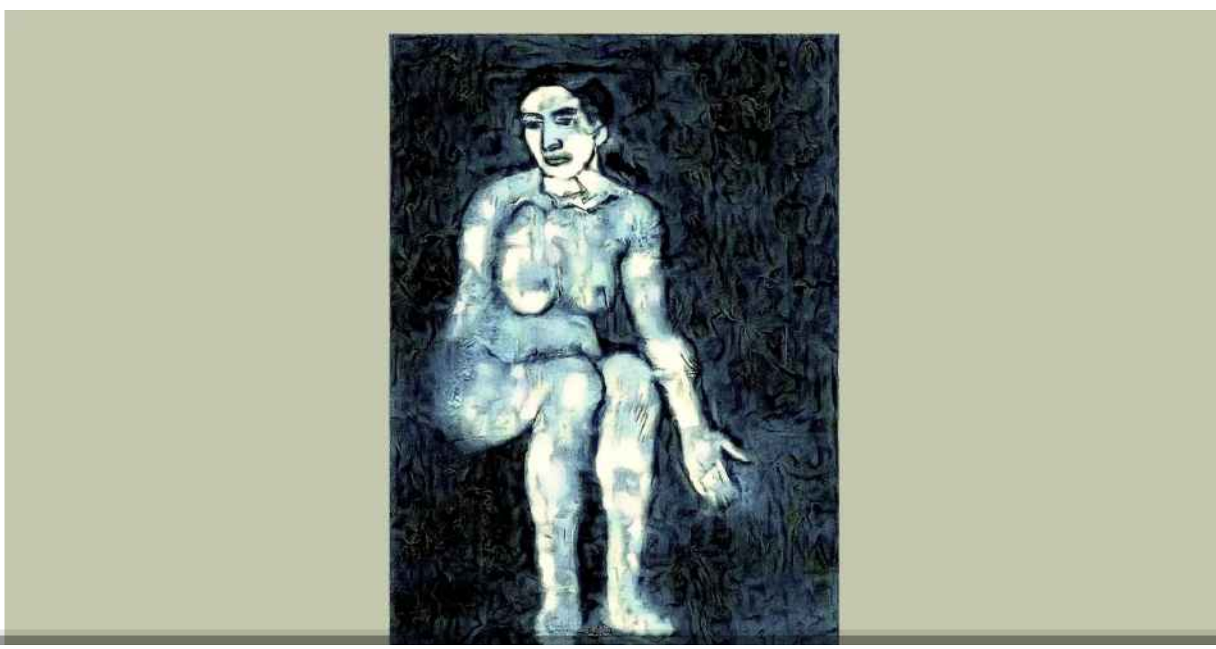
HUMANS AND TECHNOLOGY

This Picasso painting had never been seen before. Until a neural network painted it.

With help from a neural network, researchers reconstructed an image the artist created and painted over during his Blue Period.

By Emerging Technology from the arXiv

September 20, 2019



***The Old Guitarist* is probably the most famous painting from Picasso's Blue Period. It dates from 1903-**1904, when the young artist was living in poverty in Paris. Picasso used the color blue to represent the emotional pain and desolation he was experiencing at the time.

But *The Old Guitarist* is interesting for another reason. Art historians have long noted the presence of a ghostly woman's face faintly visible beneath the paint. In 1998, conservators at the Art Institute of Chicago, where the painting hangs, photographed it using x-rays and infrared light to see what lies beneath the surface.

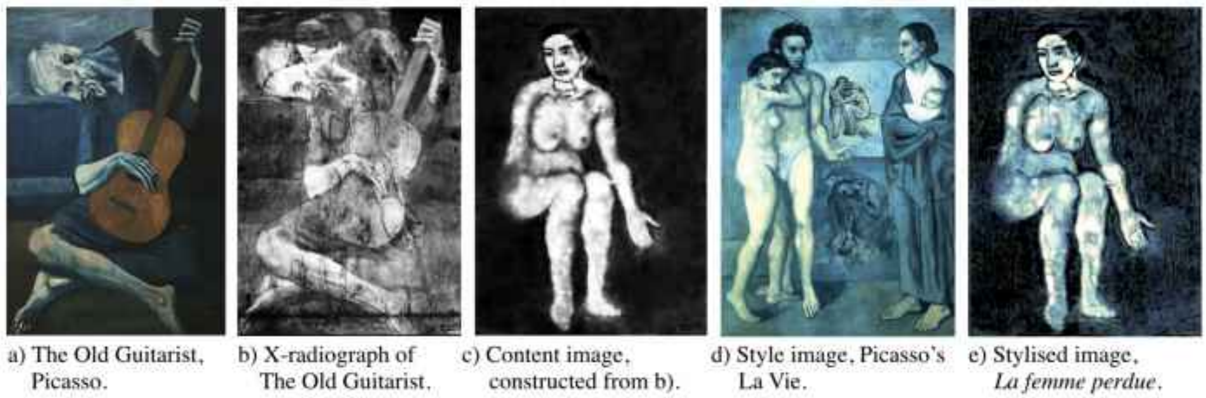
These images show an entirely different painting. It depicts a seated woman holding out her left arm. The researchers then matched this painting to a composition Picasso had sketched out in a letter to a colleague at the time.

The findings fascinated the art world. Artists often paint over earlier works, particularly during periods of penury when canvas is in short supply.

The new image provided an important insight into the progression of Picasso's work, his subjects, and his thinking during his Blue Period. Since he is perhaps the most important artist of 20th century, that's hugely significant.

But from an aesthetic point of view, what the researchers managed to retrieve is disappointing. Infrared and x-ray images show only the faintest outlines, and while they can be used to infer the amount of paint the artist used, they do not show color or style. So a way to reconstruct the lost

The researchers have used the same technique to retrieve lost paintings by other artists and say it has the potential to transform the way art historians work.



Courtesy of the researchers

First some background. Neural style transfer was developed in 2015 by Leon Gatys and colleagues at the University of Tubingen in Germany. It comes about from a fascinating insight into the way neural networks learn to recognize images of different kinds.

Neural networks consist of layers that analyze an image at different scales. The first layer might recognize broad features like edges, the next layer sees how these edges form simple shapes like circles, the next layer recognizes patterns of shapes, such as two circles close together, and yet another layer might label these pairs of circles as eyes.

This kind of network would be able to recognize eyes in paintings in a wide variety of styles, from Leonardo da Vinci to Van Gogh to Picasso. In each case, the eyes form a similar pattern that the machine can pick out.

Gatys and co went further by training such a network to recognize artistic style; for example, to distinguish a Van Gogh from a Picasso.

Their key discovery was that the ability to distinguish style was entirely separate from the ability to see faces or other objects. In fact, Gatys and co were able to separate this ability and use it in reverse. They fed a picture into the neural network, which then superimposed the style onto the image.

This work has been hugely influential. Various groups have used it to produce artworks, comics, or even movies in the style of any chosen artist.

This process works just as well with Picasso, making it possible to produce an image in the style of Picasso's cubist paintings, or his Rose Period, or indeed his Blue Period.

This is where Bourached and Cann come in. They have taken a manually edited version of the x-ray images of the ghostly woman beneath *The Old Guitarist* and passed it through a neural style transfer network. This network was trained to convert images into the style of another artwork from Picasso's Blue Period.

The result is a full-color version of the painting in exactly the style Picasso was exploring when he painted it. "We present a novel method of reconstructing lost artwork, by applying neural style transfer to x-radiographs of artwork with secondary interior artwork beneath a primary exterior, so as to reconstruct lost artwork," they say.

Of course, there is no way of knowing that Picasso painted the image this way. But Bourached and Cann say their goal is to broaden the insight into an artist's intentions, mistakes, and musings by reconstructing artwork that has been hidden. "Our method of combining original but hidden artwork, subjective human input, and neural style transfer helps to broaden an insight into an artist's creative process," they say.

That is surely just the start, and a technique that art historians might feel has the potential to go much further.

Ref: arxiv.org/abs/1909.05677 : Raiders of the Lost Art

by Emerging Technology from the arXiv

DEEP DIVE

HUMANS AND TECHNOLOGY

This couple just got married in the Taco Bell metaverse

Digital technology: The backbone of a net-zero emissions future

Organizations need to focus on accelerated digitalization to help decarbonization and emissions reduction, and to drive innovation.

By MIT Technology Review Insights

Why the definition of design might need a change

When we unpack its current meaning, we may find that we want—and need—to retool the word yet again.

By Nicholas de Monchaux

Weight-loss injections have taken over the internet. But what does this mean for people IRL?

Wegovy, Mounjaro, and Ozempic are viral TikTok sensations. But the societal impact can't be measured in views.

By Amelia Tait

STAY CONNECTED

Illustration by Rose Wong

Get the latest updates from MIT Technology Review

Discover special offers, top stories, upcoming events, and more.

Enter your email

The latest iteration of a legacy

Founded at the Massachusetts Institute of Technology in 1899, MIT Technology Review is a world-renowned, independent media company whose insight, analysis, reviews, interviews and live events explain the newest technologies and their commercial, social and political impact.

READ ABOUT OUR HISTORY

Advertise with MIT Technology Review

Elevate your brand to the forefront of conversation around emerging technologies that are radically transforming business. From event sponsorships to custom content to visually arresting video storytelling, advertising with MIT Technology Review creates opportunities for your brand to resonate with an unmatched audience of technology and business elite.

ADVERTISE WITH US

About us

Careers

Custom content

Advertise with us

International Editions

Republishing

Editorial guidelines

Privacy policy

Terms of Service

Write for us

Contact us

Who we are

Founded at the Massachusetts Institute of Technology in 1899, MIT Technology Review is a world-renowned, independent media company whose insight, analysis, reviews, interviews and live events explain the newest technologies and their commercial, social and political impacts.

MIT Technology Review derives authority from its relationship to the world's foremost technology institution and from its editors' deep technical knowledge, capacity to see technologies in their broadest context, and unequaled access to leading innovators and researchers.

Our in-depth reporting reveals what’s going on now to prepare you for what’s coming next.

Our work

MIT Technology Review SUBSCRIBE

Subscription

Our journalism is made possible by our subscribers.

Whether you prefer the connectedness of the web, or the tangibility of a printed magazine, we have an option for you.

Subscribe now

Podcasts

We produce leading podcasts on a variety of technology and business topics.

Get them wherever you find your favorite podcasts.

In Machines We Trust

The Extortion Economy

Business Lab

MIT Technology Review SUBSCRIBE

Events

We introduce emerging technologies to audiences worldwide through real-world and virtual events featuring technology's top minds.

View our events

POPULAR STORIES

Thought-leading journalism

How Facebook got addicted to spreading misinformation

Silicon Valley's latest bet on living forever

car kill the baby or the grandma?

can teach us about AI's ultimate mystery

MIT Technology Review

SUBSCRIBE

Meet the MIT Technology Review team >>

See open positions >>

Custom content
View our showcases.

Advertise with us
Review our media kit.

Contact us
Submit the contact-us form.

The latest iteration of a legacy

Founded at the Massachusetts Institute of Technology in 1899, MIT Technology Review is a world-renowned, independent media company whose insight, analysis, reviews, interviews and live events explain the newest technologies and their commercial, social and political impact.

READ ABOUT OUR HISTORY

Advertise with MIT Technology Review

Elevate your brand to the forefront of conversation around emerging technologies that are radically transforming business. From event sponsorships to custom content to visually arresting video storytelling, advertising with MIT Technology Review creates opportunities for your brand to resonate with an unmatched audience of technology and business elite.

ADVERTISE WITH US

MIT Technology Review SUBSCRIBE

Careers

Custom content

Advertise with us

International Editions

Republishing

MIT News

Help & FAQ

My subscription

Editorial guidelines

Privacy policy

Terms of Service

Write for us

Contact us

Brian Kelley: Two years of Gaussian processes

RDKit
475 subscribers
Subscribe

0 Share Download Clip Save

73 views 6 months ago
Show more

All | From RDKit | Computer Science | Statistics | Related | Listenable

Lukas Friedrich: Three Years of Project MELLODDY
RDKit
59 views • 6 months ago

Thursday afternoon lightning talks
RDKit
60 views • 6 months ago

Molecular docking methods and analysis of docking
JSS Academy of Higher Education & Research
25K views • Streamed 2 years ago

Pivot Tables Excel: Detailed Beginners Pivot Table Tutorial
Simon Sez IT
183K views • 1 year ago

Roger Sayle: Understanding tautomerism (in 2½D)
RDKit

Search

PowerPoint Slide Show - 02_Brian_GaussianProcesses.pptx - PowerPoint

Tanimoto Kernel

RELAY THERAPEUTICS

Using Ryan-Rhys Griffith's kernel (compute Tanimoto over matrix input)

$||\mathbf{x}||^2$ Xs = tf.reduce_sum(tf.square(X), axis=-1) # Squared L2-norm of X

$||\mathbf{x}'||^2$ X2s = tf.reduce_sum(tf.square(X2), axis=-1) # Squared L2-norm of X2

$< \mathbf{X}, \mathbf{X}' >$ outer_product = tf.tensordot(X, X2, [[-1], [-1]]) # outer product of the matrices X and X2

$$||\mathbf{x}||^2 + ||\mathbf{x}'||^2 - < \mathbf{x}, \mathbf{x}' >$$

denominator = -outer_product + broadcasting_elementwise(tf.add, Xs, X2s)

kernel = self.variance * outer_product / denominator

Variance is what is being trained
(in addition to noise)

This just makes tensorflow fast

https://towardsdatascience.com/gaussian-process-regression-on-molecules-in-gpflow-ee6fedab2130

12:06 / 32:09

Brian Kelley: Two years of Gaussian processes

RDKit
475 subscribers

Subscribe

0 Share Download Clip Save