Artificial Intelligence in Medicine

How AI can help doctors and patients? And why it is not that easy…

Antons Mislēvičs
Head of AI / Machine Learning, C.T.Co

Andrey Bondarenko
Senior Data Scientist, C.T.Co

Slides: https://link.makingaireal.com/rsu2024

Agenda

1. Why there is so much buzz around AI in medicine?
2. What is a current state of AI?
3. What are examples from real projects?
4. Why it is not that easy?

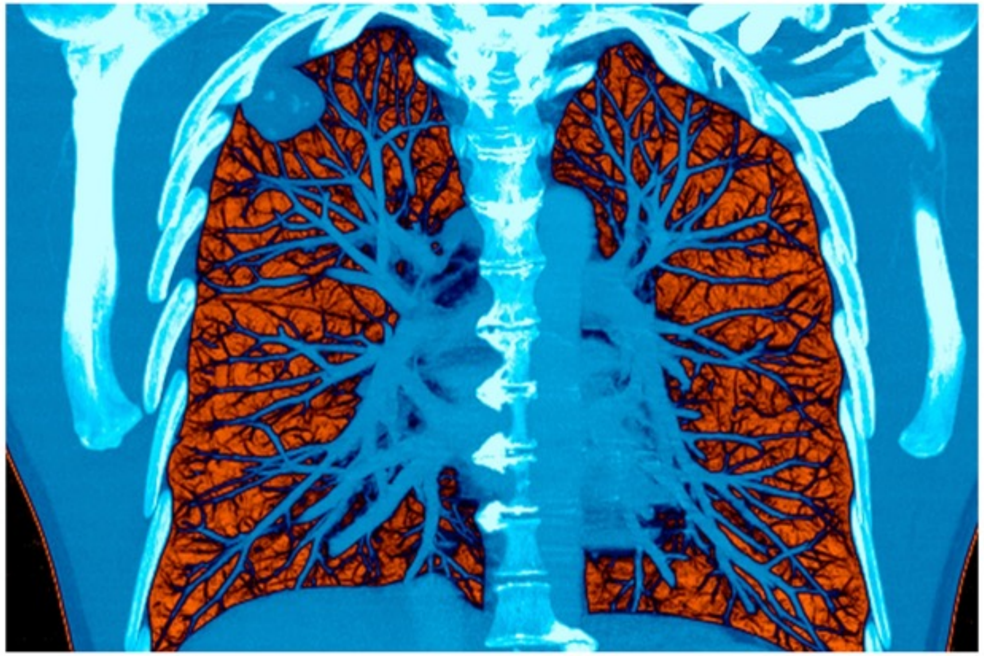
Slides: https://link.makingaireal.com/rsu2024

What do people expect from AI in medicine?

The New York Times

A.I. Took a Test to Detect Lung Cancer. It Got an A.

Artificial intelligence may help doctors make more accurate readings of CT scans used to screen for lung cancer.



A colored CT scan showing a tumor in the lung. Artificial intelligence was just as good, and sometimes better, than doctors in diagnosing lung tumors in CT scans, a new study indicates. Voisin/Science Source

By Denise Grady

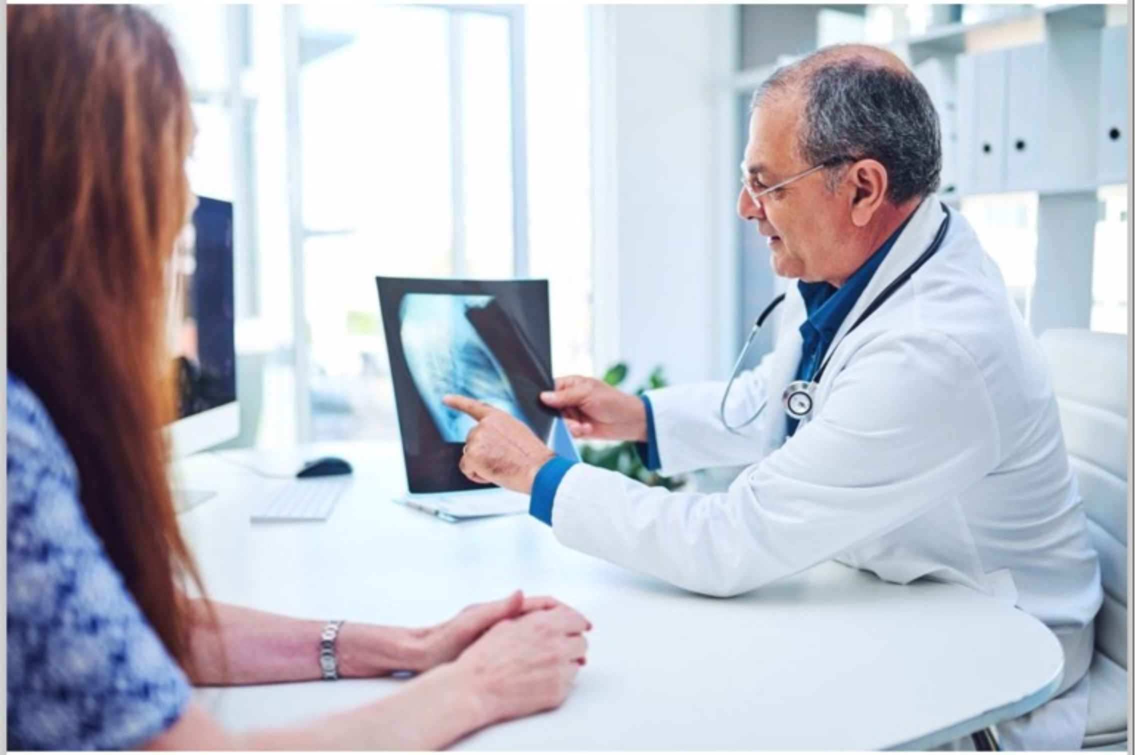
May 20, 2019

40

Could artificial intelligence be the future of cancer diagnosis?

Tim Newman • Tuesday 16 July 2019

In a recent study, researchers trained an algorithm to differentiate between malignant and benign lesions in scans of breast tissue.



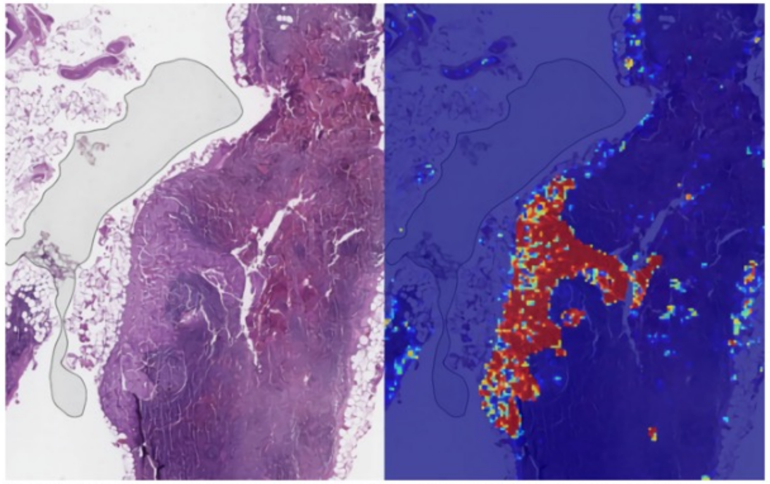
A new study asks whether artificial intelligence could streamline cancer diagnosis.

https://www.nytimes.com/2019/05/20/health/cancer-artificial-intelligence-ct-scans.html
https://www.medicalnewstoday.com/articles/325750.php

AI

Google AI claims 99% accuracy in metastatic breast cancer detection

KYLE WIGGERS @KYLE_L_WIGGERS OCTOBER 12, 2018 9:00 AM



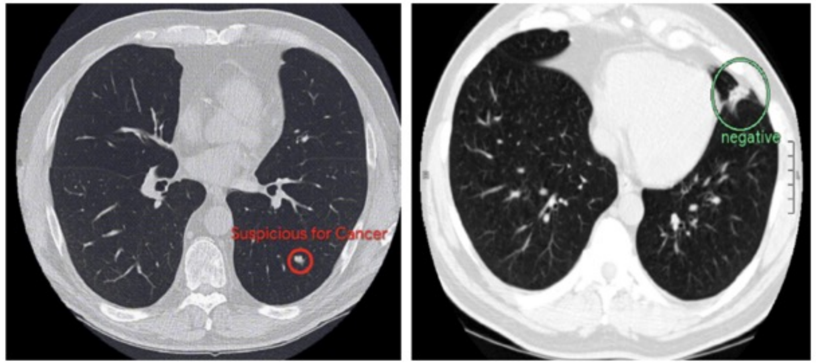
Above: Left: a slide containing lymph nodes. Right: LYNA identifying the tumor region.

Image Credit: Google

AI

Google's lung cancer detection AI outperforms 6 human radiologists

KHARI JOHNSON @KHARIJOHNSON MAY 20, 2019 8:00 AM



Google AI researchers working with Northwestern Medicine created an AI model capable of detecting lung cancer from screening tests better than human radiologists with an average of eight years experience.

https://venturebeat.com/2018/10/12/google-ai-claims-99-accuracy-in-metastatic-breast-cancer-detection/
https://venturebeat.com/2019/05/20/googles-lung-cancer-detection-ai-outperforms-6-human-radiologists/

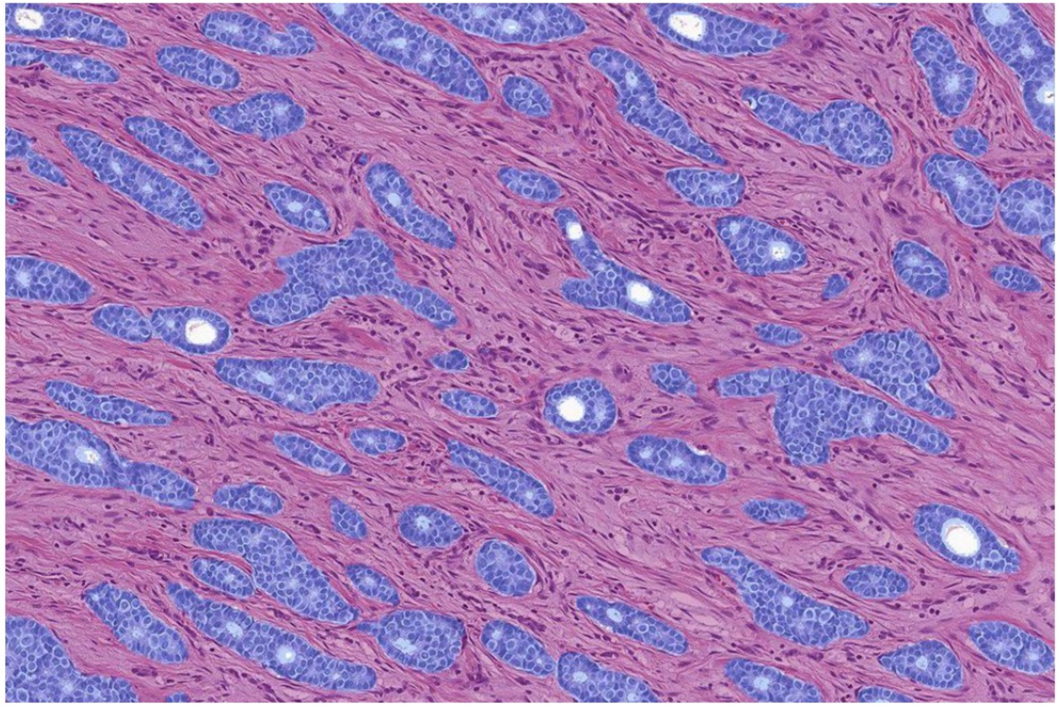
6,566 views | Apr 30, 2019, 11:46am

The Cutting-Edge Of AI Cancer Detection

Charles Towers-Clark Contributor

AI & Big Data

I write about AI, data, deep tech & self-management in the digital age



Detecting and treating cancer could benefit from a dose of AI, whether in pathological H&E slides or an entirely new branch of diagnosis altogether. LUNIT.IO

Deep Learning for Detection of Diabetic Eye Disease

Tuesday, November 29, 2016

Posted by Lily Peng MD PhD, Product Manager and Varun Gulshan PhD, Research Engineer

Diabetic retinopathy (DR) is the fastest growing cause of blindness, with nearly 415 million diabetic patients at risk worldwide. If caught early, the disease can be treated; if not, it can lead to irreversible blindness. Unfortunately, medical specialists capable of detecting the disease are not available in many parts of the world where diabetes is prevalent. We believe that Machine Learning can help doctors identify patients in need, particularly among underserved populations.

A few years ago, several of us began wondering if there was a way Google technologies could improve the DR screening process, specifically by taking advantage of recent advances in Machine Learning and Computer Vision. In "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs", published today in JAMA, we present a deep learning algorithm capable of interpreting signs of DR in retinal photographs, potentially helping doctors screen more patients in settings with limited resources.

One of the most common ways to detect diabetic eye disease is to have a specialist examine pictures of the back of the eye (Figure 1) and rate them for disease presence and severity. Severity is determined by the type of lesions present (e.g. microaneurysms, hemorrhages, hard exudates, etc), which are indicative of bleeding and fluid leakage in the eye. Interpreting these photographs requires specialized training, and in many regions of the world there aren't enough qualified graders to screen everyone who is at risk.

Figure 1. Examples of retinal fundus photographs that are taken to screen for DR. The image on the left is of a healthy retina (A), whereas the image on the right is a retina with referable diabetic retinopathy (B) due a number of hemorrhages (red spots) present.

https://www.forbes.com/sites/charlestowersclark/2019/04/30/the-cutting-edge-of-ai-cancer-detection/
https://ai.googleblog.com/2016/11/deep-learning-for-detection-of-diabetic.html

Artificial Intelligence

IBM's Watson is better at diagnosing cancer than human doctors

By **IAN STEADMAN**

Monday 11 February 2013

22,861 views | May 25, 2011, 05:02pm

IBM's Watson Now A Second-Year Med Student

Bruce Upbin Former Contributor

Tech

I manage our technology coverage.

https://www.wired.co.uk/article/ibm-watson-medical-doctor
https://www.forbes.com/sites/bruceupbin/2011/05/25/ibms-watson-now-a-second-year-med-student/

Technology & Ideas

IBM's Watson Hasn't Beaten Cancer, But A.I. Still Has Promise

The company made bold claims that haven't yet panned out. But someday artificial intelligence could crack the code of individualized diagnosis and treatment.

By Faye Flam
24 August 2018, 18:00 EEST

02 Apr 2019 | 15:00 GMT

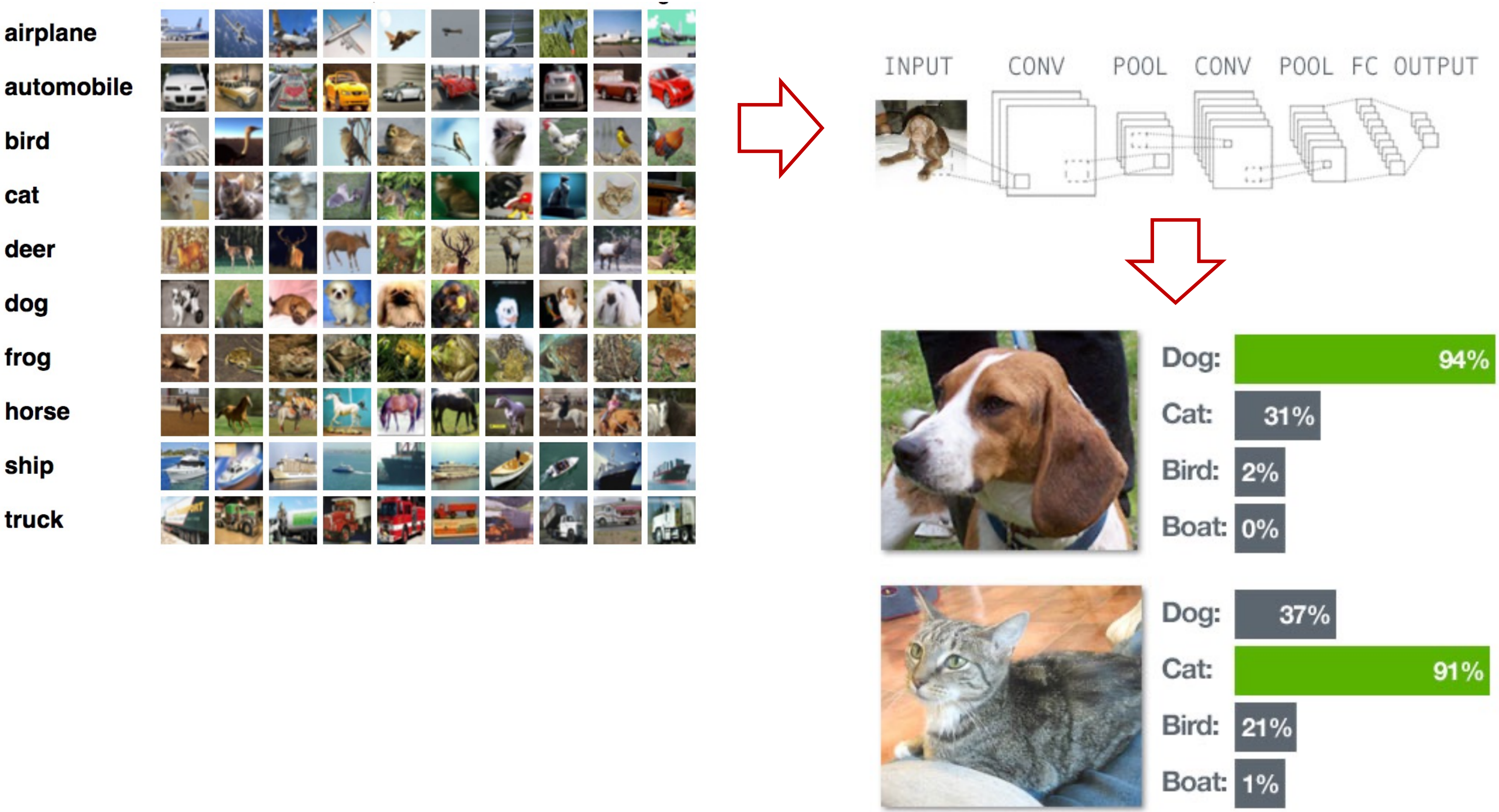
How IBM Watson Overpromised and Underdelivered on AI Health Care

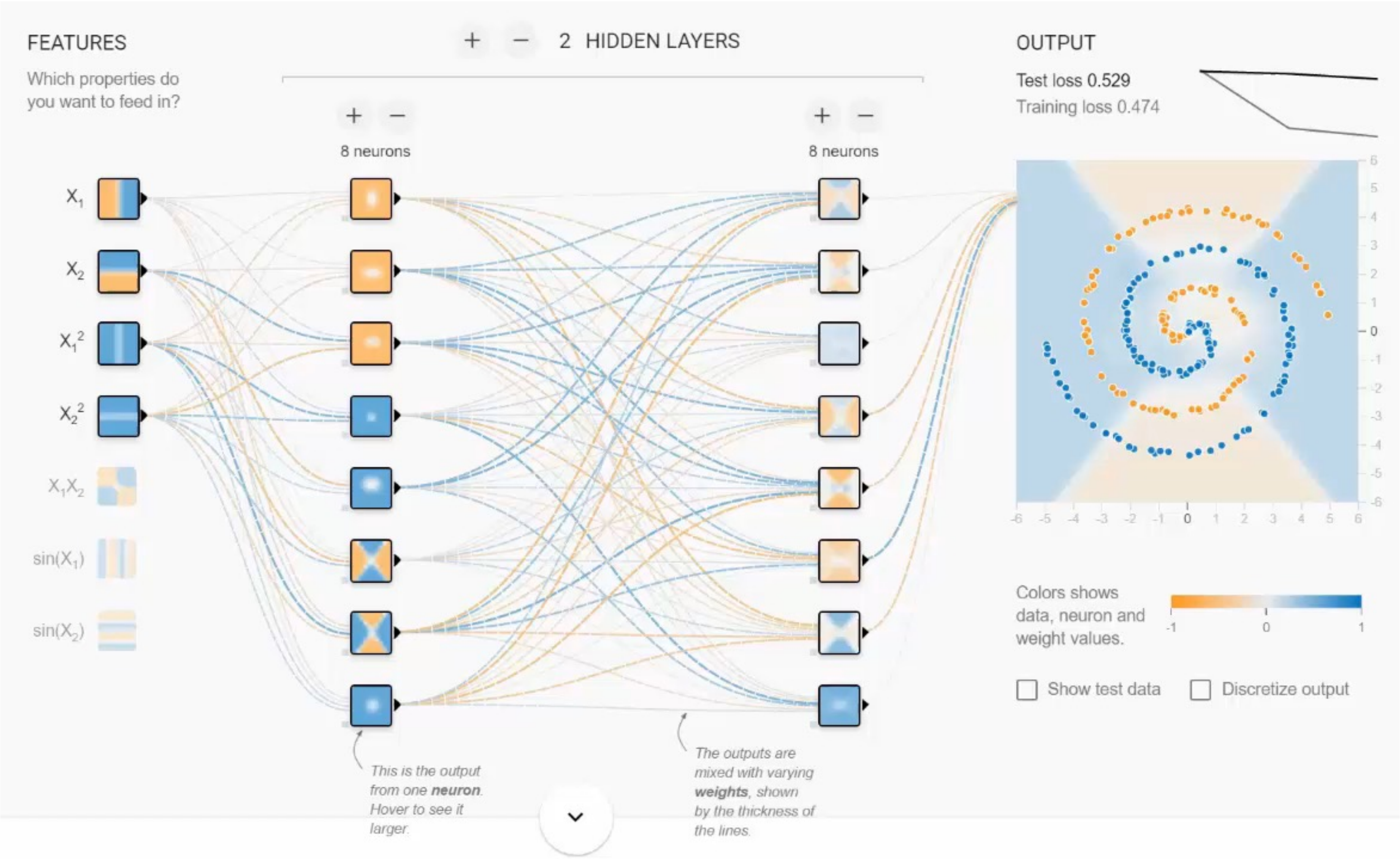
After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

By **Eliza Strickland**

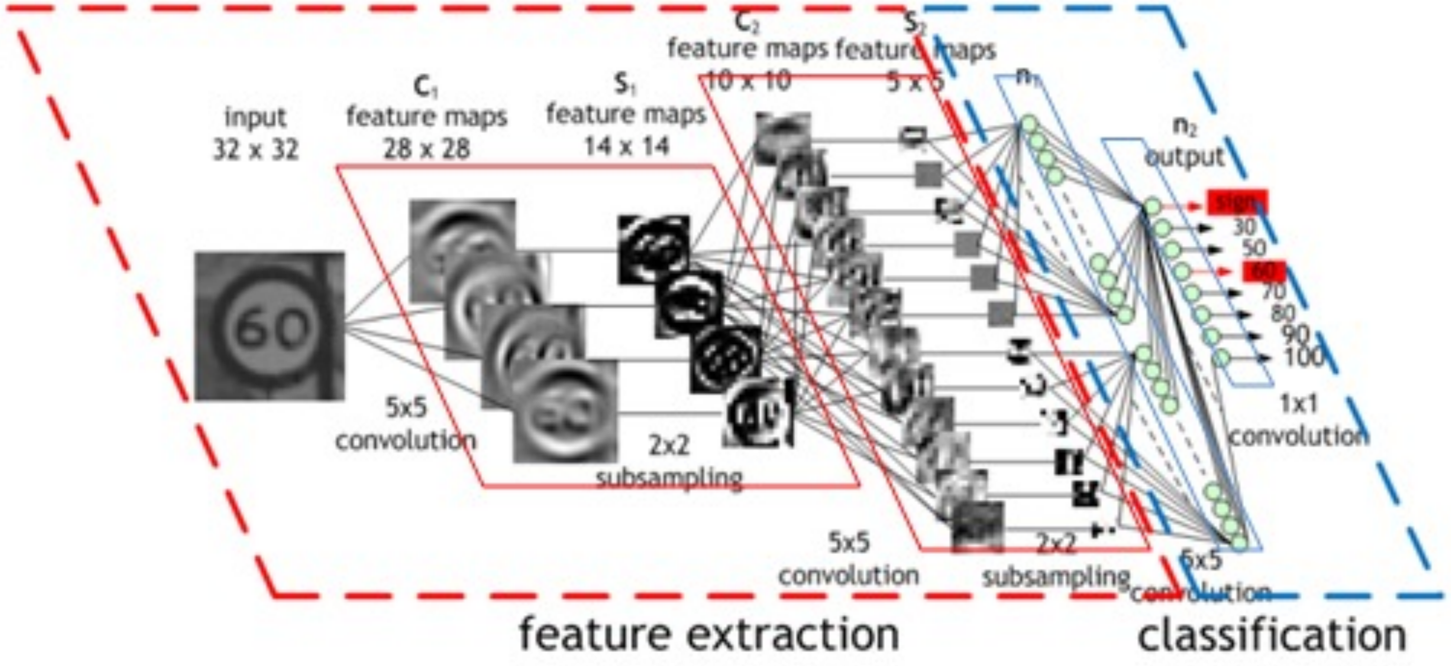
https://www.bloomberg.com/opinion/articles/2018-08-24/ibm-s-watson-failed-against-cancer-but-a-i-still-has-promise
https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

What is a current state of AI?


airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck
INPUT
CONV
POOL
CONV
POOL
FC
OUTPUT
Dog: 94%
Cat: 31%
Bird: 2%
Boat: 0%
Dog: 37%
Cat: 91%
Bird: 21%
Boat: 1%



http://playground.tensorflow.org/



https://developer.nvidia.com/discover/convolutional-neural-network

Allows to process different types of data

Audio

Image / Video

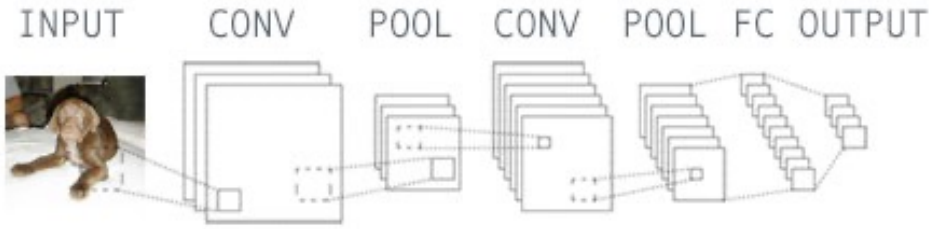
Text

The cat lives in the city.

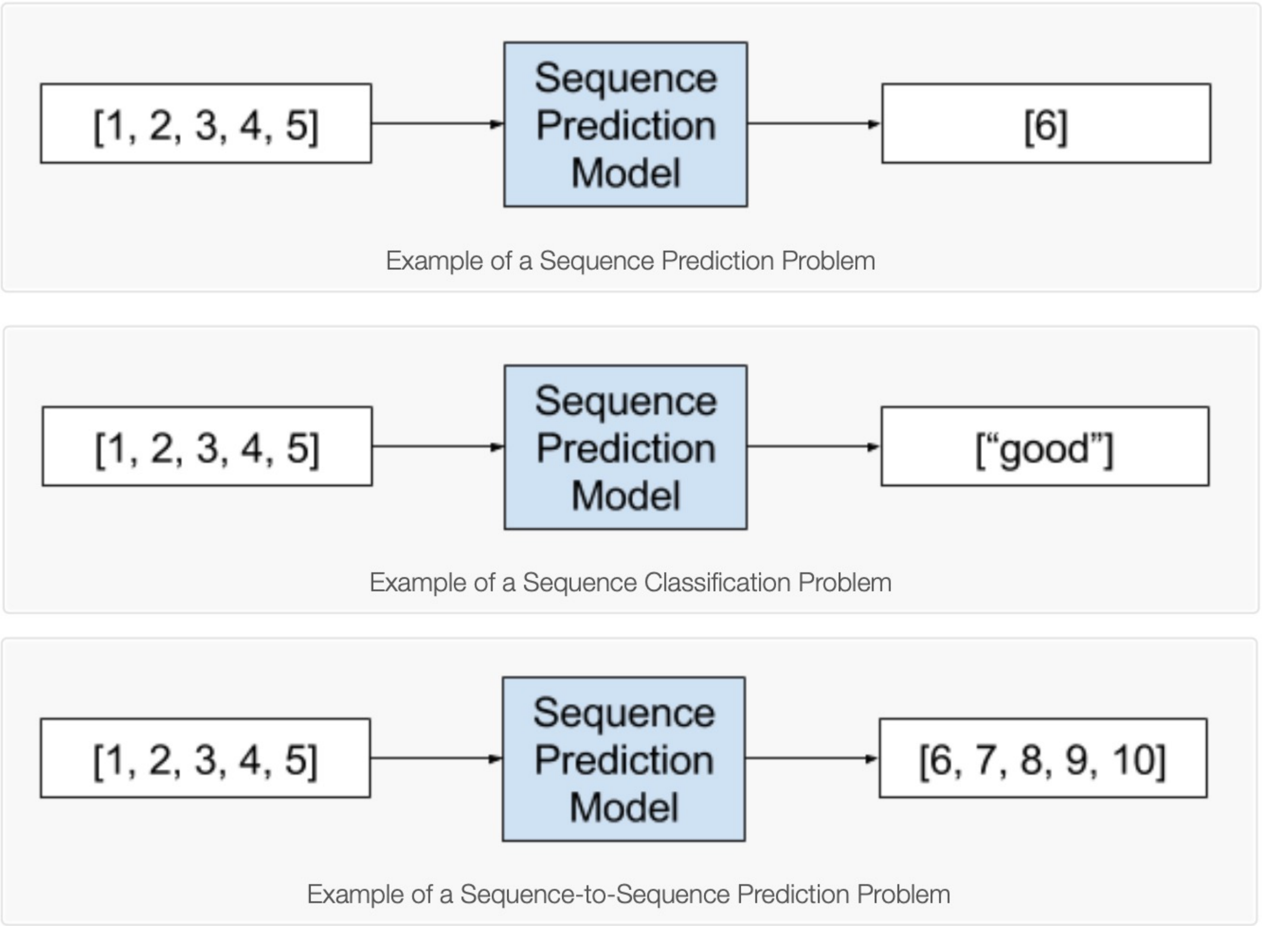
The tiger lives in the forest.

Embeddings

0.1	0.1	0.3	...	0.1	0.5	0.9

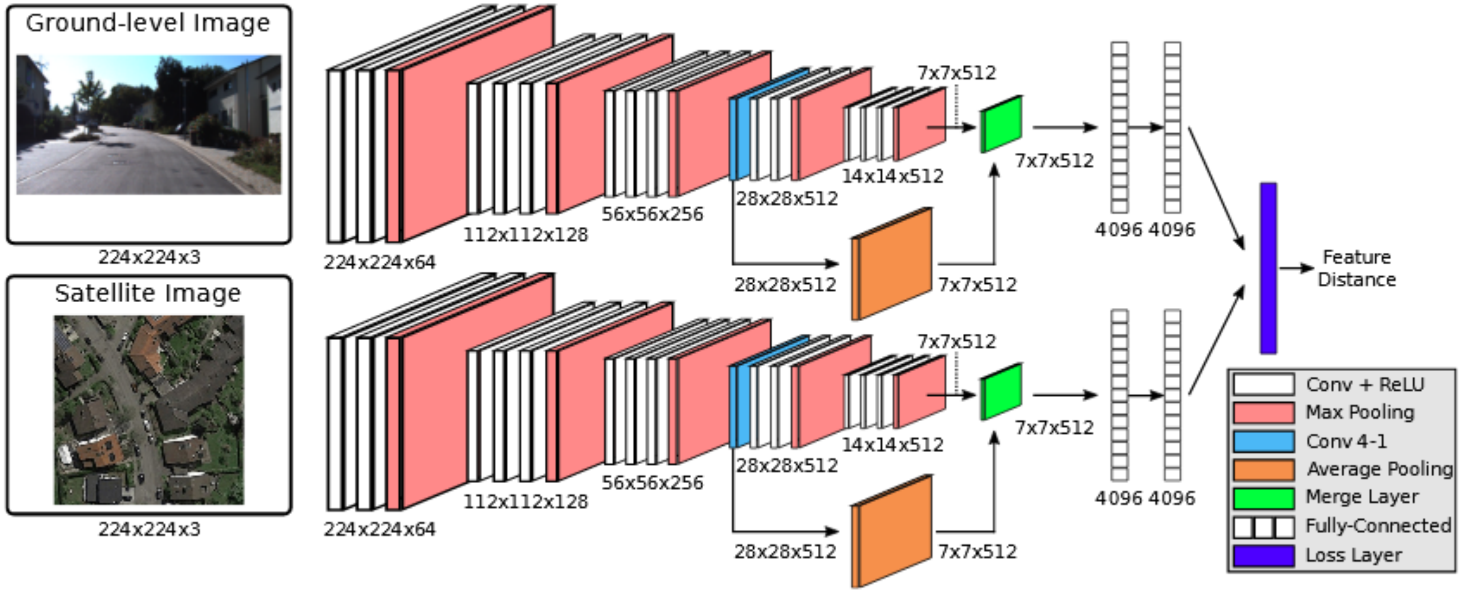


Allows to process time series / sequences



Making Predictions with Sequences: https://machinelearningmastery.com/sequence-prediction/

Allows to combine different inputs



https://news.developer.nvidia.com/satellite-images-help-track-a-vehicle/

Self-supervised learning / learning from structure of the data

Predict next word
(causal language modelling)

My				→	name
My	name			→	is
My	name	is		→	Sylvain
My	name	is	Sylvain	→	.

Improve image resolution

Predict masked word
(masked language modelling)

My [MASK] is Sylvain .

[MASK] ↓ name

Text to image / image to text

How do Transformers work?: https://huggingface.co/learn/nlp-course/chapter1/4
Image Super-Resolution: An Overview of the Current State of Research: https://towardsdatascience.com/image-super-resolution-an-overview-of-the-current-state-of-research-94294a77ed5a
Scaling up GANs for Text-to-Image Synthesis: https://mingukkang.github.io/GigaGAN/

AI writes fake news (2019)

New AI fake text generator may be too dangerous to release, say creators

The Elon Musk-backed nonprofit company OpenAI declines to release research publicly for fear of misuse

▲ The AI wrote a new passage of fiction set in China after being fed the opening line of Nineteen Eighty-Four by George Orwell (pictured). Photograph: Mondadori/Getty Images

The creators of a revolutionary AI system that can write news stories and works of fiction - dubbed "deepfakes for text" - have taken the unusual step of not releasing their research publicly, for fear of potential misuse.

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

Pérez and the others then ventured further into the valley. "By the time we reached the top of one peak, the water looked blue, with some crystals on top," said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns.

While examining these bizarre creatures the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common 'language,' something like a dialect or dialectic."

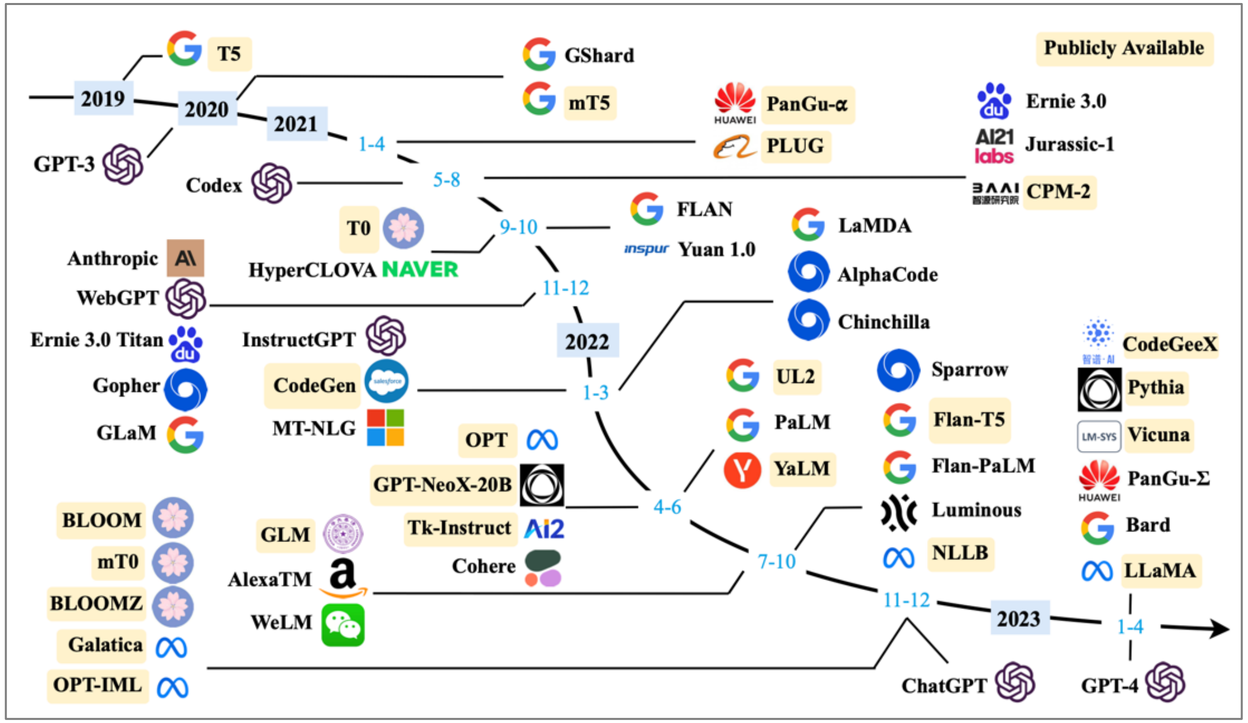
Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

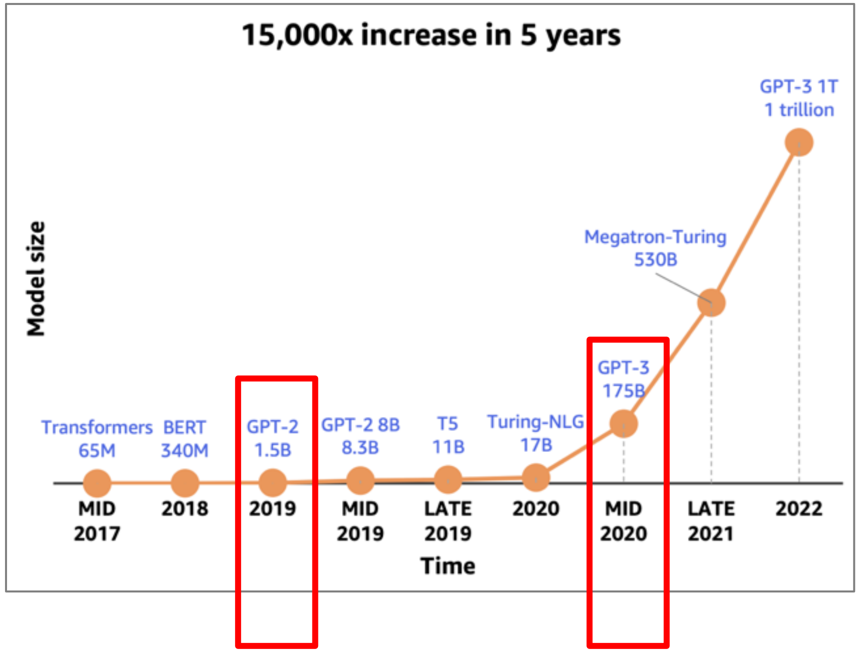
While their origins are still unclear, some believe that perhaps the creatures were created when a human and a unicorn met each other in a time before human civilization. According to Pérez, "In South America, such incidents seem to be quite common."

However, Pérez also pointed out that it is likely that the only way of knowing for sure if unicorns are indeed the descendants of a lost alien race is through DNA. "But they seem to be able to communicate in English quite well, which I believe is a sign of evolution, or at least a change in social organization," said the scientist.

Better Language Models and Their Implications: https://openai.com/blog/better-language-models/#sample1
GPT-2: 1.5B Release: https://openai.com/blog/gpt-2-1-5b-release/
New AI fake text generator may be too dangerous to release, say creators: https://www.theguardian.com/technology/2019/feb/14/elon-musk-backed-ai-writes-convincing-news-fiction
OpenAI built a text generator so good, it's considered too dangerous to release: https://techcrunch.com/2019/02/17/openai-text-generator-dangerous/
The AI Text Generator That's Too Dangerous to Make Public: https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/

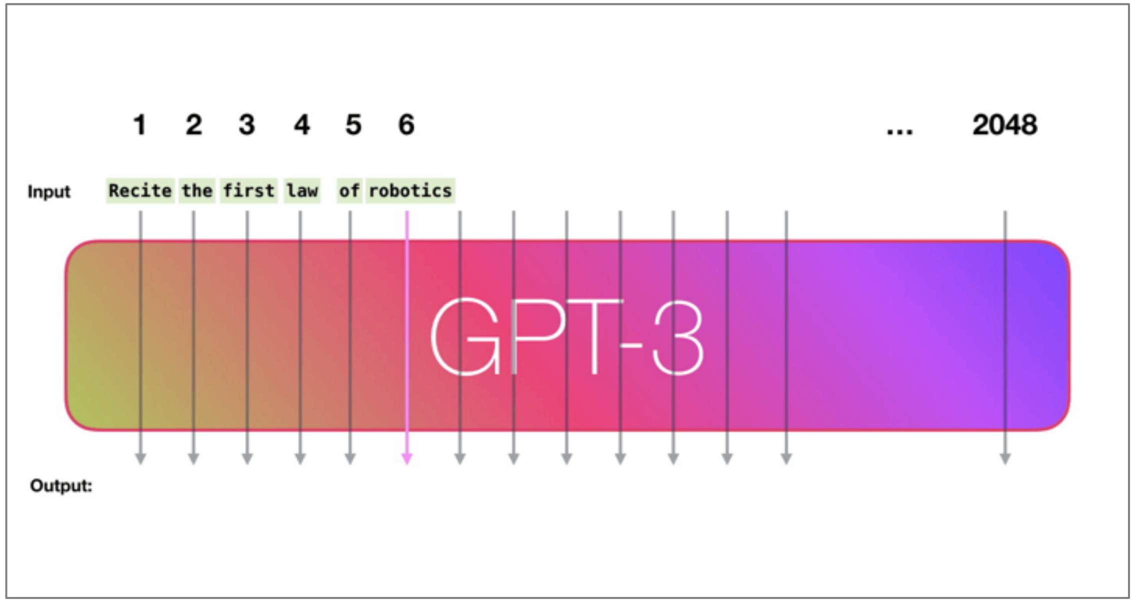
Large Language Models: ChatGPT is not the only one...

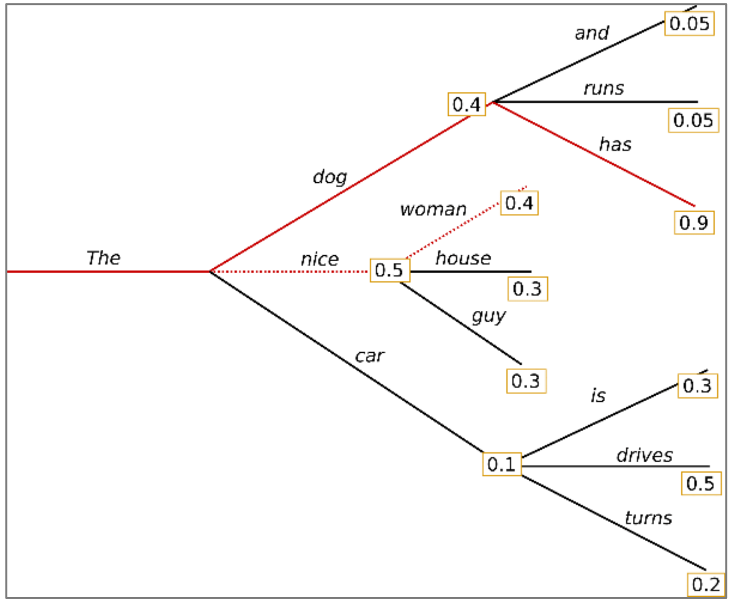




A Survey of Large Language Models (2023): https://arxiv.org/pdf/2303.18223.pdf
A Survey Of Large Language Models (Kaggle): https://www.kaggle.com/general/400499
Deploying Large NLP Models: Infrastructure Cost Optimization: https://neptune.ai/blog/nlp-models-infrastructure-cost-optimization
Train and deploy large language models on Amazon SageMaker: https://d1.awsstatic.com/events/Summits/reinvent2022/AIM405_Train-and-deploy-large-language-models-on-Amazon-SageMaker.pdf

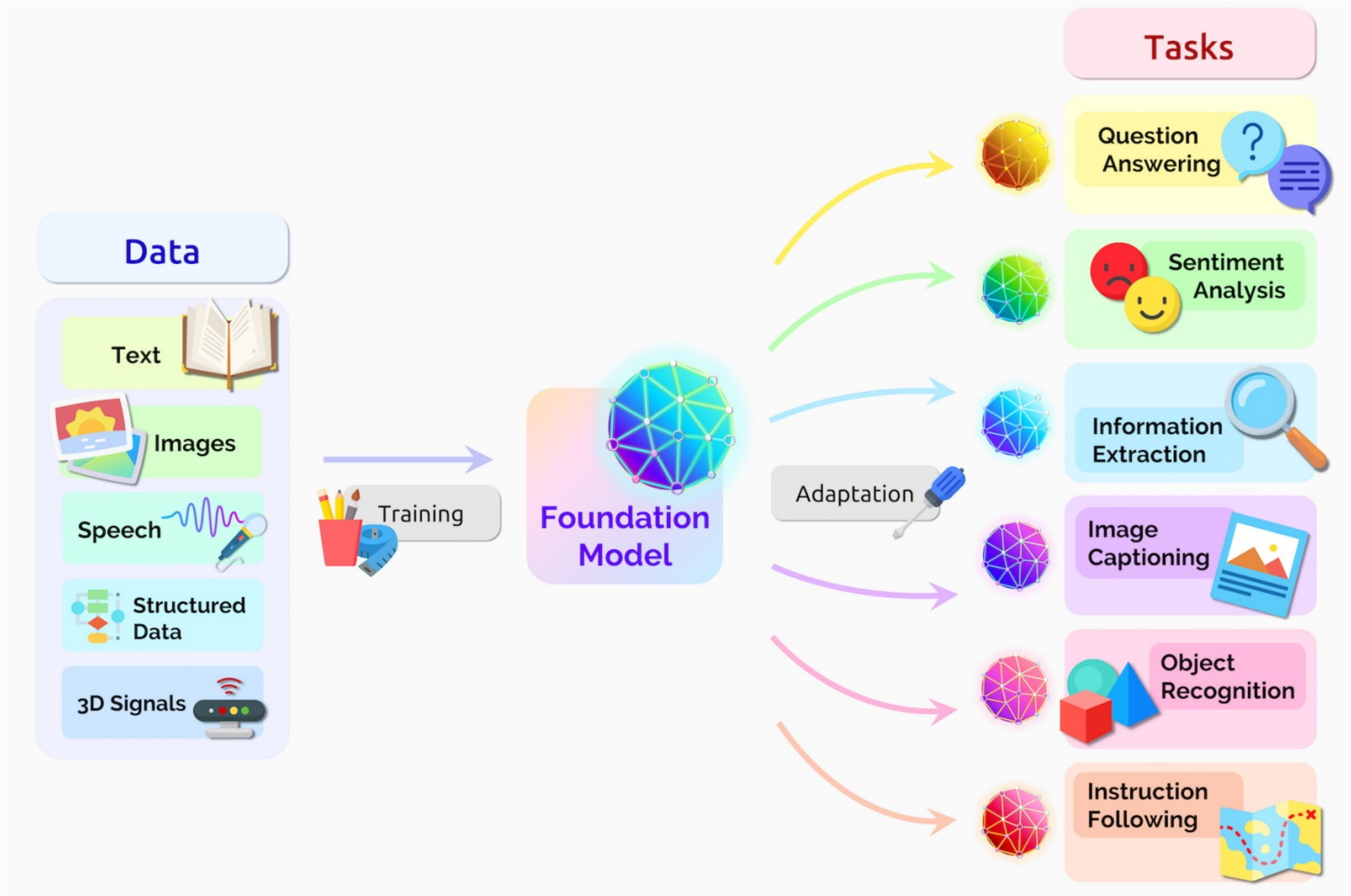
How Large Language Models generate text?



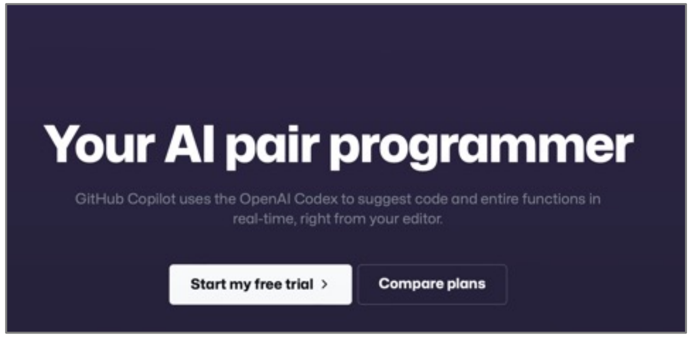


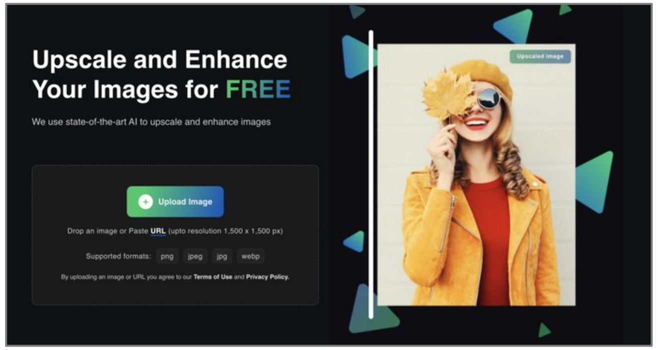
How to generate text: using different decoding methods for language generation with Transformers: https://huggingface.co/blog/how-to-generate
How GPT3 Works - Visualizations and Animations: http://jalammar.github.io/how-gpt3-works-visualizations-animations/

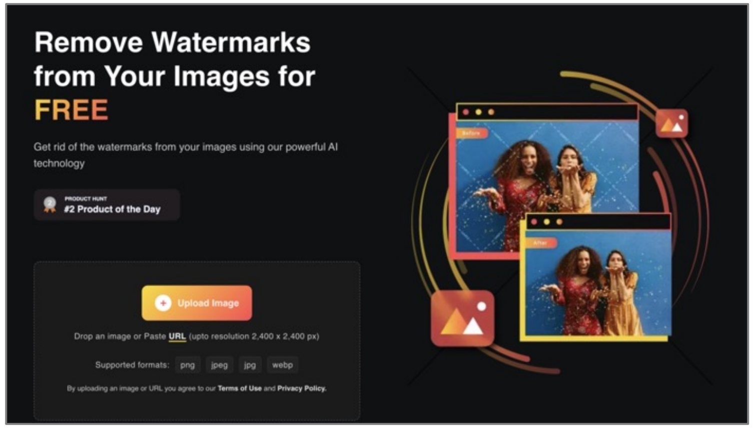
Foundation Models and multimodal transformers

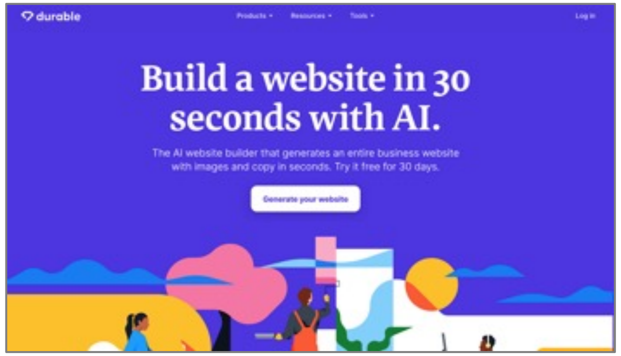


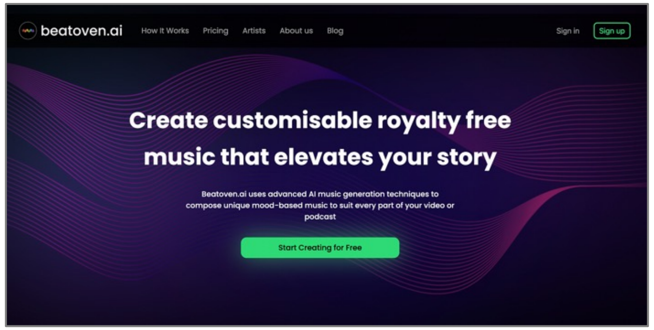
What Is a Transformer Model? https://blogs.nvidia.com/blog/2022/03/25/what-is-a-transformer-model/

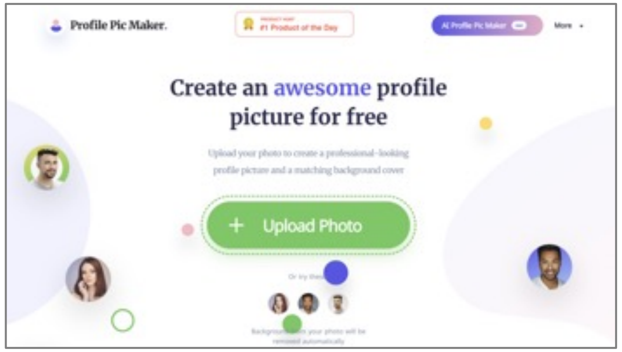


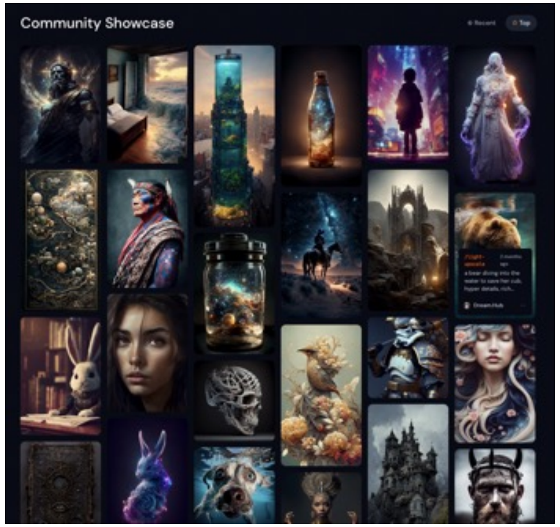


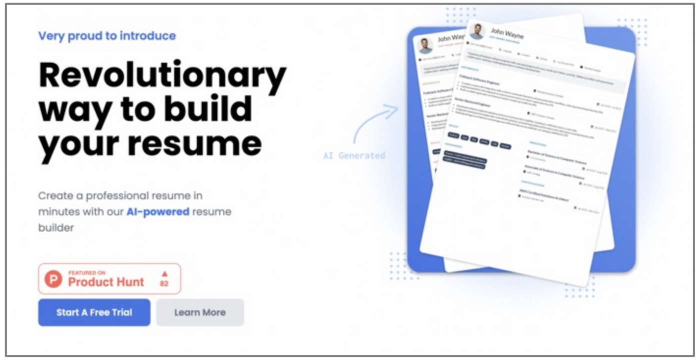












GitHub Copilot: https://github.com/features/copilot
Durable: https://durable.co
Tome: https://tome.app
Upscale.media: https://www.upscale.media
beatoven.ai: https://www.beatoven.ai
Profile Pic Maker: https://pfpmaker.com
Resume Studio: https://resumestudio.careers
WatermarkRemover.io: https://www.watermarkremover.io
AutoDraw: https://www.autodraw.com
Midjourney: https://www.midjourney.com

Is AI already as smart as we are?

Artificial Narrow Intelligence (ANI), aka Machine Learning/Deep Learning → Artificial Capable Intelligence (ACI) → Artificial General Intelligence (AGI) → Artificial Super Intelligence (ASI)

Mustafa Suleyman: My new Turing test would see if AI can make $1 million

The Modern Turing Test would measure what an AI can do in the world, not just how it appears. And what is more telling than making money?

By Mustafa Suleyman

July 14, 2023

The truth is, I think we're in a moment of genuine confusion (or, perhaps more charitably, debate) about what's really happening. Even as the Turing test falls, it doesn't leave us much clearer on where we are with AI, on what it can actually achieve. It doesn't tell us what impact these systems will have on society or help us understand how that will play out.

We need something better. Something adapted to this new phase of AI. So in my forthcoming book *The Coming Wave*, I propose the Modern Turing Test—one equal to the coming AIs. What an AI can say or generate is one thing. But what it can achieve in the world, what kinds of concrete actions it can take—that is quite another. In my test, we don't want to know whether the machine is intelligent as such; we want to know if it is capable of making a meaningful impact in the world. We want to know what it can *do*.

Mustafa Suleyman: My new Turing test would see if AI can make $1 million: https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/

TECH

Meta's AI chief doesn't think AI super intelligence is coming anytime soon, and is skeptical on quantum computing

PUBLISHED SUN, DEC 3 2023•11:00 AM EST | UPDATED SUN, DEC 3 2023•AT 11:20 EST

Jonathan Vanian
@IN/JONATHAN-VANIAN-B704432/

SHARE

KEY POINTS

- Facebook parent Meta held a media event this week in San Francisco highlighting the 10-year anniversary of its Fundamental AI Research team.
- Society is more likely to get "cat-level" or "dog-level" AI years before human-level AI, Meta chief scientist Yann LeCun said.
- Unlike Google, Microsoft and other tech giants, Meta is not making a big bet on quantum computing.

Interview

AI pioneer Fei-Fei Li: 'I'm more concerned about the risks that are here and now'

Zoë Corbyn

AI is promising to transform the world in ways that don't necessarily seem for the better: killing jobs, supercharging disinformation and surveillance, and causing harm through biased algorithms. Do you take any responsibility for how AI is being used?

First, to be clear, AI is "promising" nothing. It is people who are promising - or not promising. AI is a piece of software. It is made by people, deployed by people and governed by people.

Second, of course I don't take responsibility for how all of AI is being used. Should Maxwell take responsibility for how electricity is used because he developed a set of equations to describe it? But I am a person who has a voice and I feel I have a responsibility to raise important issues - which is why I created Stanford HAI. We cannot pretend AI is just a bunch of math equations and that's it. I view AI as a tool. And like other tools our relationship with it is messy. Tools are invented by and large to deliver good but there are unintended consequences and we have to understand and mitigate their risks well.

To what extent do you worry about the existential risk of AI systems - that they could gain unanticipated powers and destroy humanity - as some high-profile tech leaders and researchers have sounded the alarm about, and which was a large focus of last week's UK AI Safety Summit?

I respect the existential concern. I'm not saying it is silly and we should never worry about it. But, in terms of urgency, I'm more concerned about ameliorating the risks that are here and now.

Bill Gates does not expect GPT-5 to be much better than GPT-4

The buzz around generative AI is huge, in part because the technology has evolved so rapidly in such a short time. Can it keep up the pace?

In an interview with the German business newspaper Handelsblatt, Microsoft founder Bill Gates says there are many reasons to believe that GPT technology has reached a plateau.

Meta's AI chief doesn't think AI super intelligence is coming anytime soon, and is skeptical on quantum computing: https://www.cnbc.com/2023/12/03/meta-ai-chief-yann-lecun-skeptical-about-agi-quantum-computing.html
AI pioneer Fei-Fei Li: 'I'm more concerned about the risks that are here and now': https://www.theguardian.com/technology/2023/nov/05/ai-pioneer-fei-fei-li-im-more-concerned-about-the-risks-that-are-here-and-now
Bill Gates does not expect GPT-5 to be much better than GPT-4: https://the-decoder.com/bill-gates-does-not-expect-gpt-5-to-be-much-better-than-gpt-4/

CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation

Zhihong Chen[1*] Maya Varma[1*] Jean-Benoit Delbrouck[1*] Magdalini Paschali[1]
Louis Blankemeier[1] Dave Van Veen[1] Jeya Maria Jose Valanarasu[1] Alaa Youssef[1]
Joseph Paul Cohen[1†] Eduardo Pontes Reis[1] Emily B. Tsai[1] Andrew Johnston[1]
Cameron Olsen[1] Tanishq Mathew Abraham[2] Sergios Gatidis[1]
Akshay S. Chaudhari[1] Curtis Langlotz[1]

[1]Stanford University [2]Stability AI
{zhihongc,mvarma2,jbdel,paschali,akshaysc,langlotz}@stanford.edu

5.2 Evaluation Results

Tables 3, 4 and Figure 3 provide results on the two tasks associated with evaluation axis 2, medical text generation and summarization. We provide a detailed breakdown of CheXagent performance below:

- On findings section generation, CheXagent outperforms all medical FMs across all metrics on both the private dataset and MIMIC-CXR. In particular, CheXagent achieves an average improvment of 6.5 points on RadGraph scores and 0.5 points on CheXbert scores; this is notable since these metrics directly evaluate factual correctness. Figure 3 shows results from automated evaluations using GPT-4. For 152 randomly selected samples from our private dataset, we provided GPT-4 with a reference report, the findings section generated by CheXagent, and the findings section generated by each of the medical FM baselines; GPT-4 was then prompted to select the report with the highest accuracy. As shown in Figure 3, our GPT-4 evaluations demonstrate that CheXagent generates high-quality reports when compared to other medical FMs.
- For findings summarization, as shown in Table 4, CheXagent outperforms LLMs of comparable size on the Rouge-L metric and achieves comparable performance to LLMs with more than twice the number of parameters [80, 81].

6 Human Evaluation

To complement the quantitative results presented in Section 5.2, we conduct a reader study in which five radiologists compare text generated by CheXagent against text written by a physician. This study includes the two textual understanding tasks described in Section 5.1: findings generation and

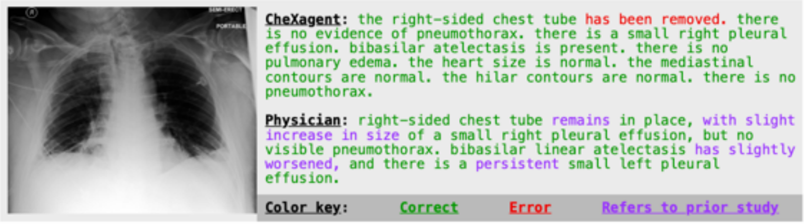


Figure 5: Comparison of CheXagent (top) against physician (bottom) on report generation. Radiologist reports often refer to past studies (**purple**), a context not available to CheXagent. The presence of these references in the training set perhaps contributes to CheXagent's error (**red**) in an otherwise accurate report (**green**), motivating future work for vision-language models.

CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation : https://arxiv.org/pdf/2401.12208.pdf

What are good uses of Large Language Models (LLMs)?

1. Content Generation

- ✓ Generating first draft of content:
  - There is no single right answer
  - It is easier to choose from options, than start from blank page
  - Similar content was generated in the past
  - It is acceptable if the result is not factually correct
- ✓ Agentic flows:
  - Prompt Engineering + Workflow + Roles
  - Examples: CrewAI, AutoGen, Agentic Teams

Examples:

- Crawling web pages to extract required information / generate summaries
- Generating SQL queries based on user prompt and DB schema description in documentation
- Generating UI form flow based on user free text description

2. Search / QnA

- ✓ Advanced semantic search:
  - RAG - Retrieval Augmented Generation
- ✓ AI Experts / Targeted GPTs:
  - Fine-tuned models or advanced RAG configurations

Examples:

- Use existing Knowledge Base / documentation to answer user queries
- Answer questions about specific document

3. Language Processing

- ✓ LLM is parsing unstructured input:
  - The list of entities / values is extracted
  - The result is used by rule-based or ML / DL approach
- ✓ LLM is extracting complex features:
  - The result is used by ML / DL model
  - The model can be fine-tuned on specific task, adding classification head

Examples:

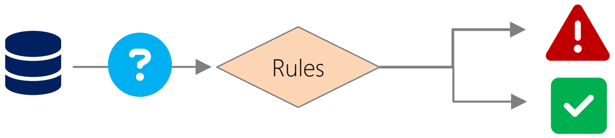
- User intent classification in a chatbot
- Automated document / e-mail processing
- Information extraction from documents – agentic flow with dynamic prompt engineering

Potential Issues

- ! Hallucinations
- ! Biases from data seen during training
- ! Slow response times
- ! Large (potentially uncontrolled) cost, when using 3rd party models
- ! Strict dependency on 3rd party service provider
- ! Data is sent to 3rd party service provider (if not using self-hosted small language models)

AI is not only Large Language Models (LLMs)

1. Rule-based – algorithmic approach



Good when:

- ✓ Expert knowledge is available
- ✓ Rules are known and can be implemented

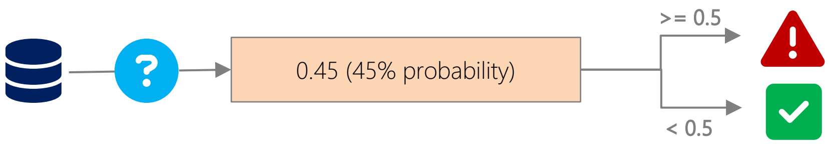
Pros:

- \+ Deterministic – the outcome is always known
- \+ The output is machine readable
- \+ Can be implemented by software developers
- \+ Do not require historical data

Cons:

- – Domain expert needs to define rules
- – Rules need maintenance / updates
- – Maintenance gets hard as the number of rules grows

2. ML / Deep Learning – statistical learning approach



Good when:

- ✓ Historical data on which the model can be trained is available

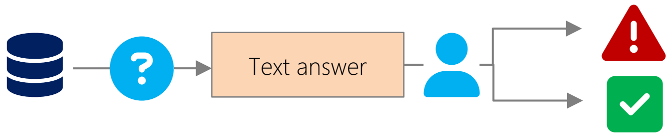
Pros:

- \+ Deterministic
- \+ The output is machine readable
- \+ The model detects "rules" automatically from historical data
- \+ Easy to update in the future – retrain the model on new data (with Deep Learning and AutoML)
- \+ The model returns probabilities what allows to fine-tune strategies of how to handle predictions, when integrating into business process

Cons:

- – Typically requires large labeled datasets to train (for Deep Learning)
- – Resulting model is typically a black box, and requires thorough statistical evaluation

3. Large Language Model – Generative AI approach



Good when:

- ✓ No previous data on which the model can be trained / validated is available

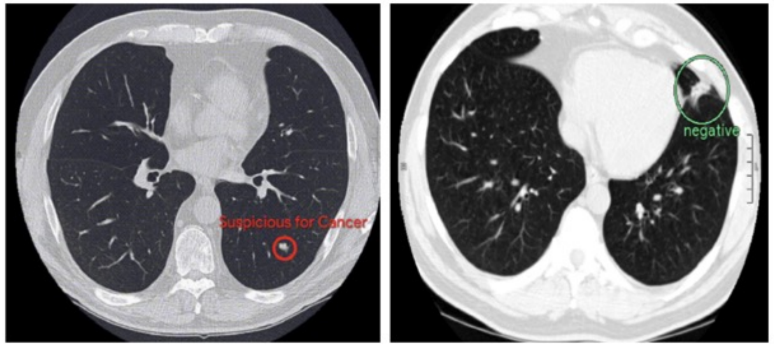
Pros:

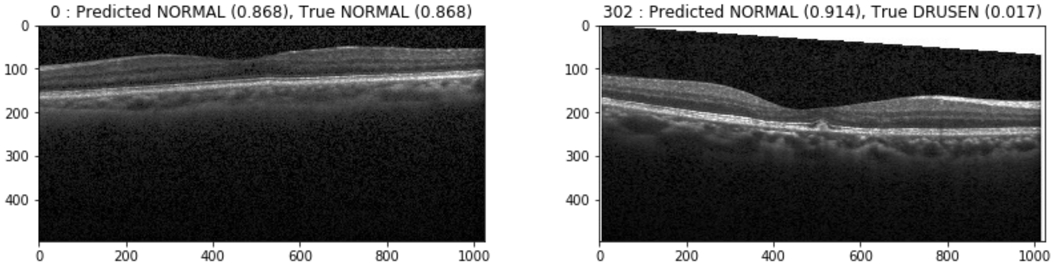
- \+ The model relies on "knowledge" that LLM had learnt from massive datasets during pre-training / fine-tuning
- \+ Easy to invoke - ask questions in natural language
- \+ Easy to integrate - invoke as Service Web API

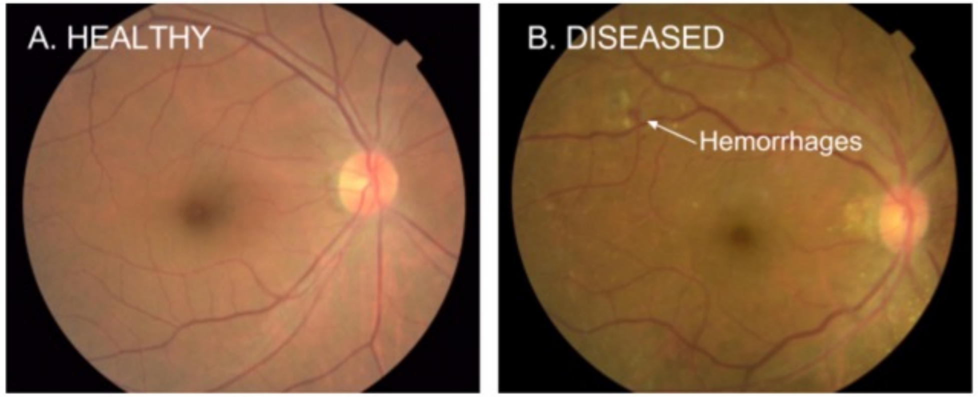
Cons:

- – Validating the output is mandatory
- – The output is hard to evaluate
- – Strict dependency (coupling) with service provider
- – The output is non-deterministic (similar query can produce very different output

Applications for medical imagery

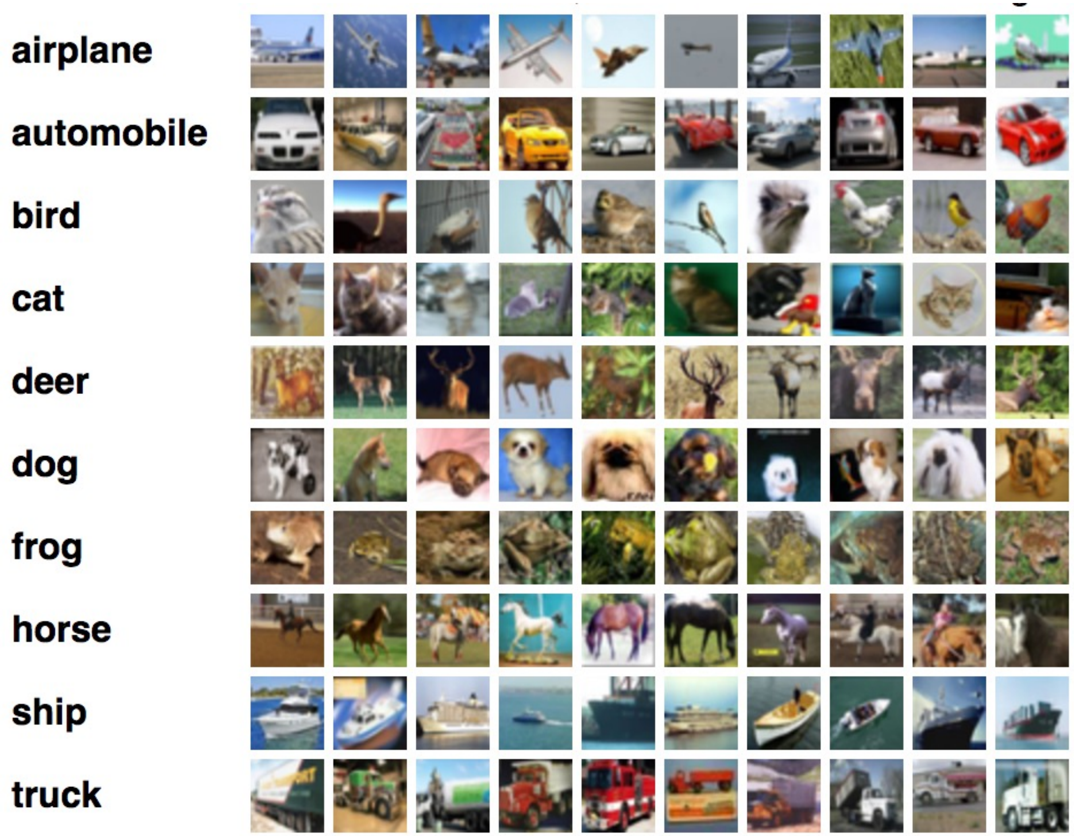

Suspicious for Cancer
negative

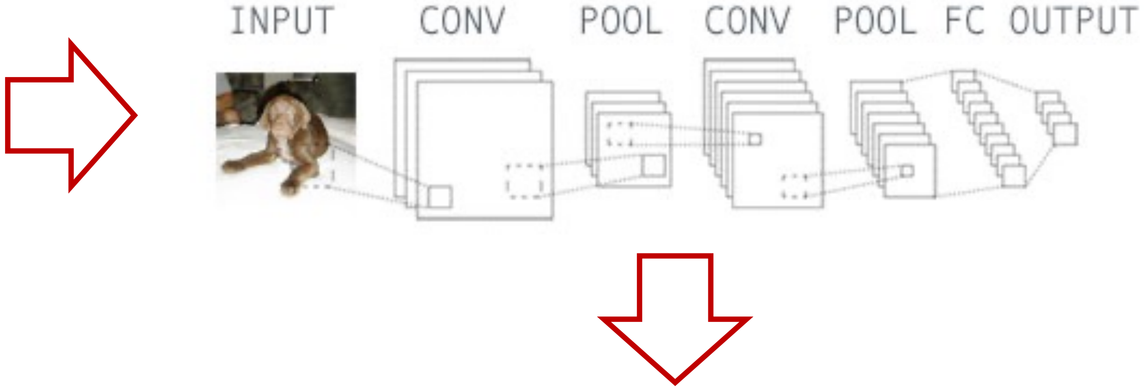

0 : Predicted NORMAL (0.868), True NORMAL (0.868)
302 : Predicted NORMAL (0.914), True DRUSEN (0.017)
0
100
200
300
400
0
200
400
600
800
1000

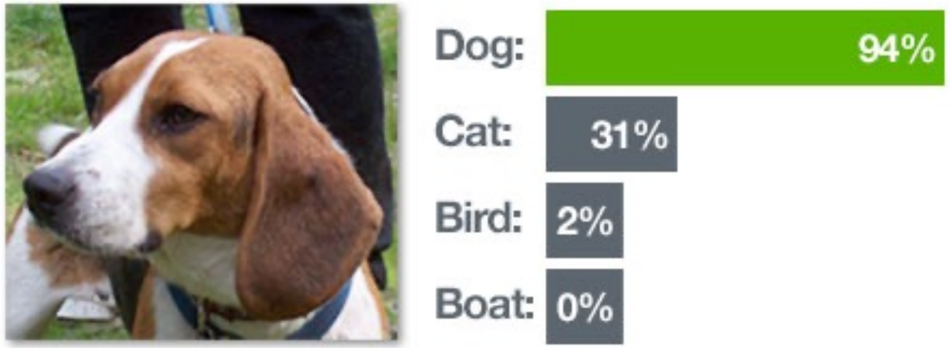

A. HEALTHY
B. DISEASED
Hemorrhages

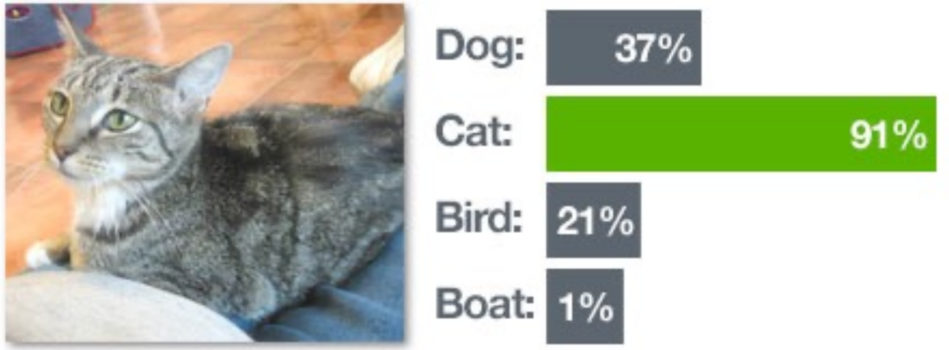
or

1. Image Classification





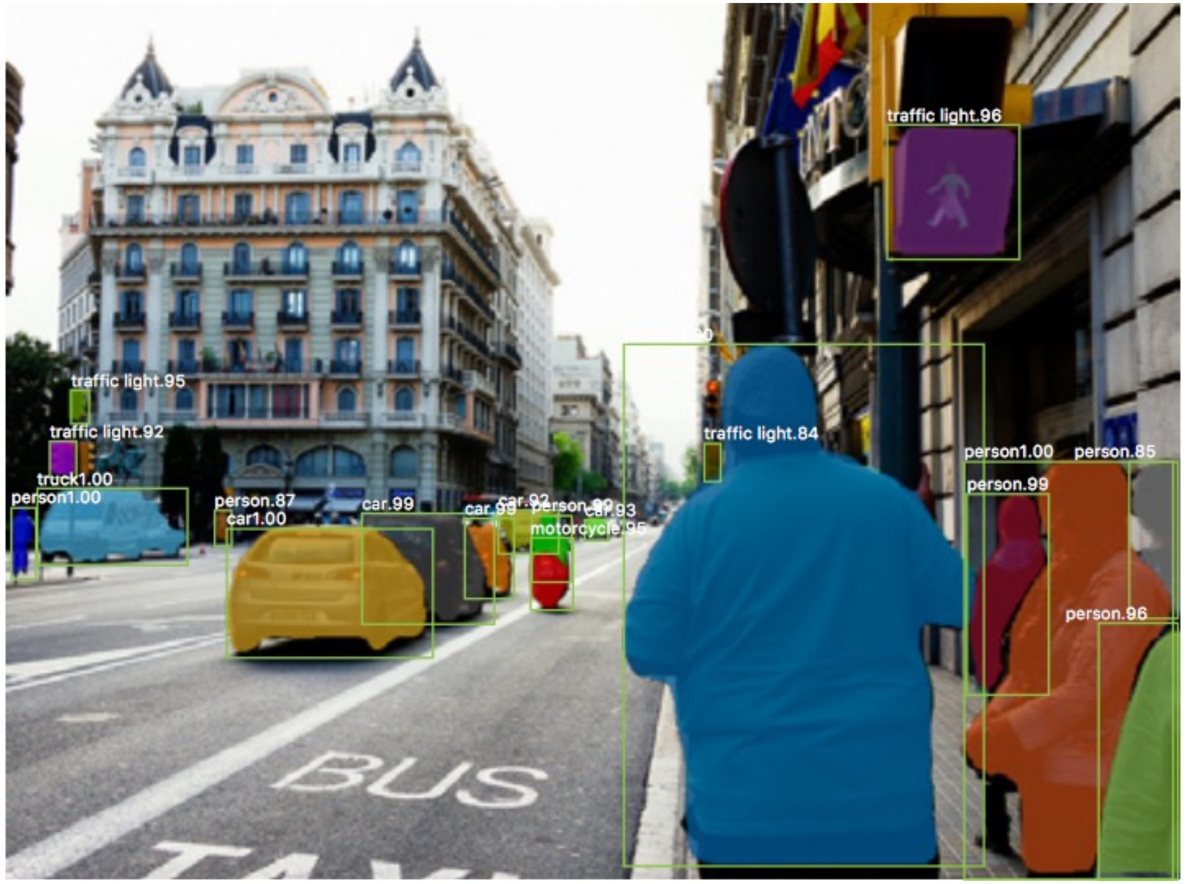




Scenarios:

1. Assign image (part of image) to a class:
   - good/bad
   - diagnosis 1/2/3
2. Give probabilities/confidence scores to predictions

2. Image Segmentation

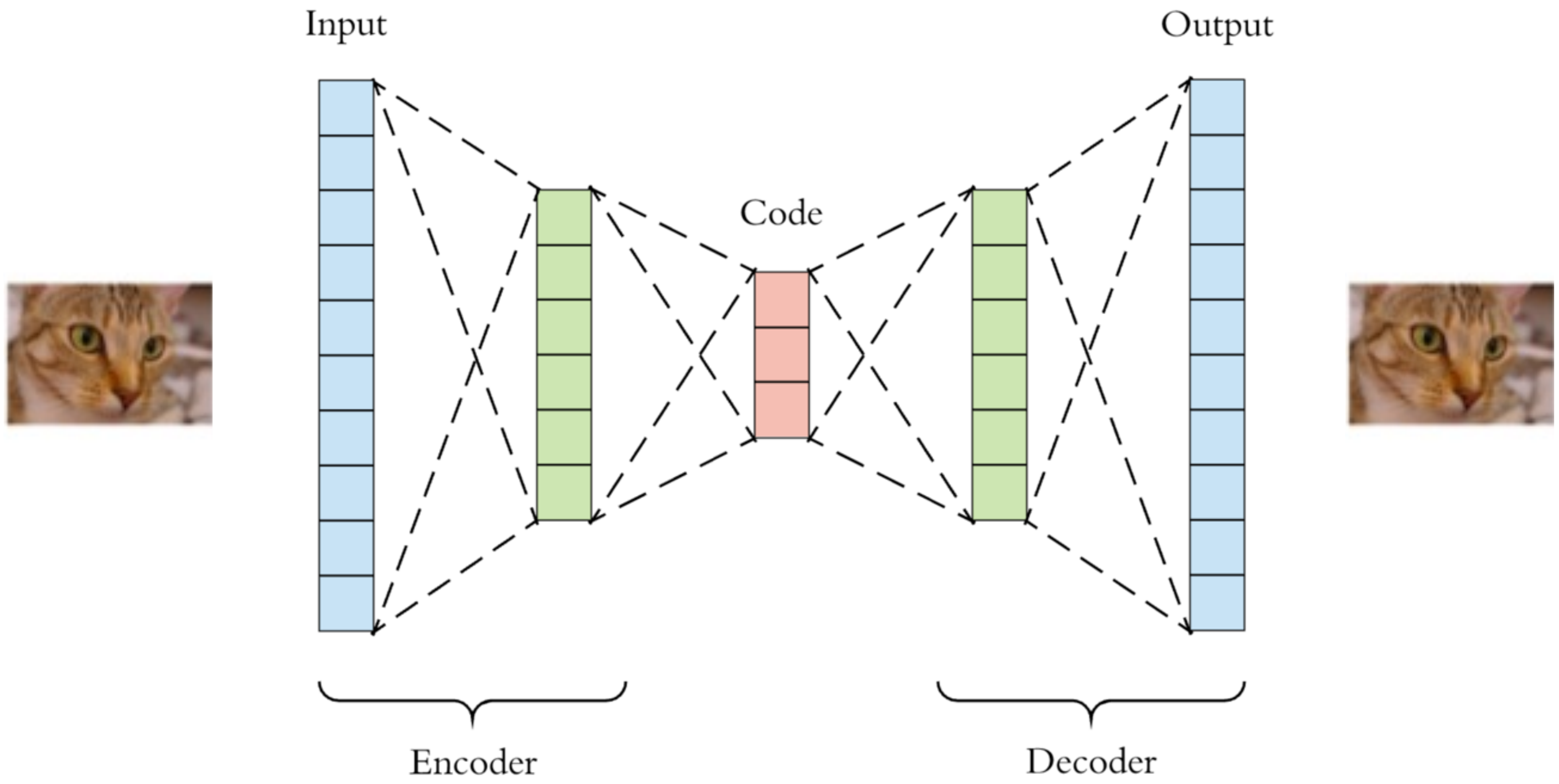


Scenarios:

1. Segment image into regions, so it is easier to analyze
2. Highlight segments that require attention (for example, with specific disease)

Mask R-CNN: https://arxiv.org/abs/1703.06870
A Brief History of CNNs in Image Segmentation: From R-CNN to Mask R-CNN:
https://blog.athelas.com/a-brief-history-of-cnns-in-image-segmentation-from-r-cnn-to-mask-r-cnn-34ea83205de4

3. Autoencoders

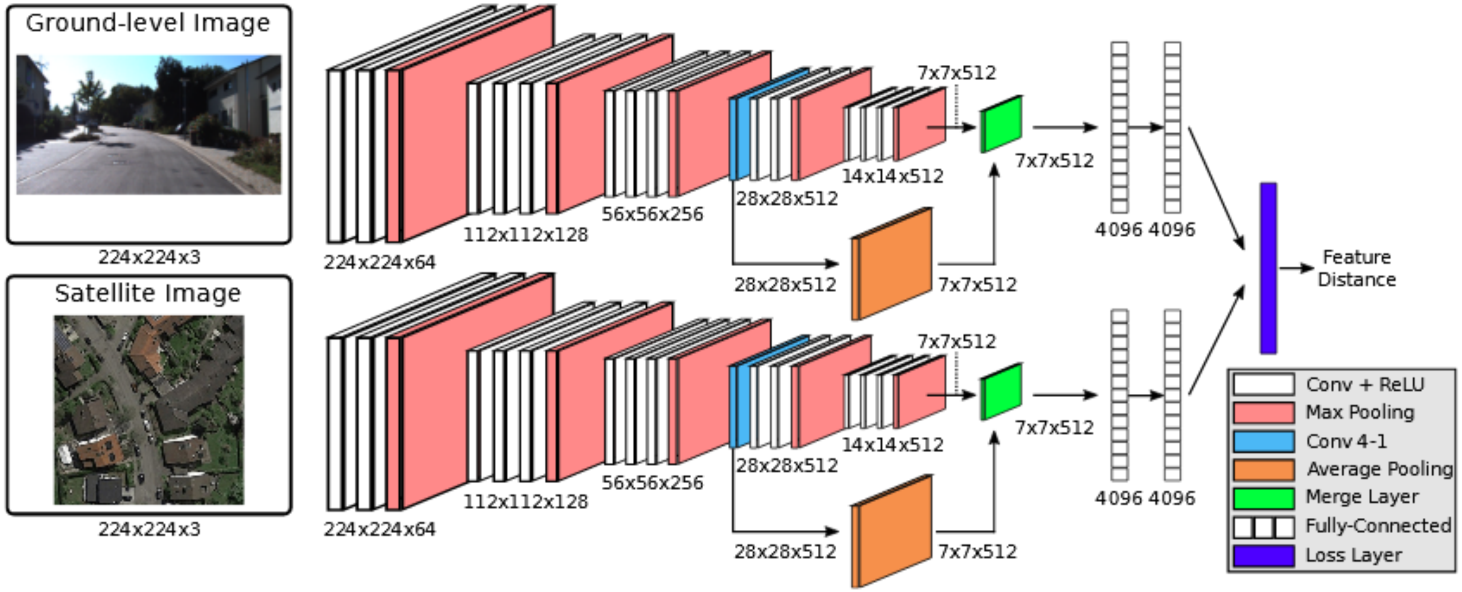


Scenarios:

1. Train on good cases => detect bad cases
2. Train on specific bad cases => detect if the case is from this category
3. Train on common cases => detect anomalies/rare cases

Comprehensive Introduction to Autoencoders: https://towardsdatascience.com/generating-images-with-autoencoders-77fd3a8dd368

4. Ensembles/multi-input models

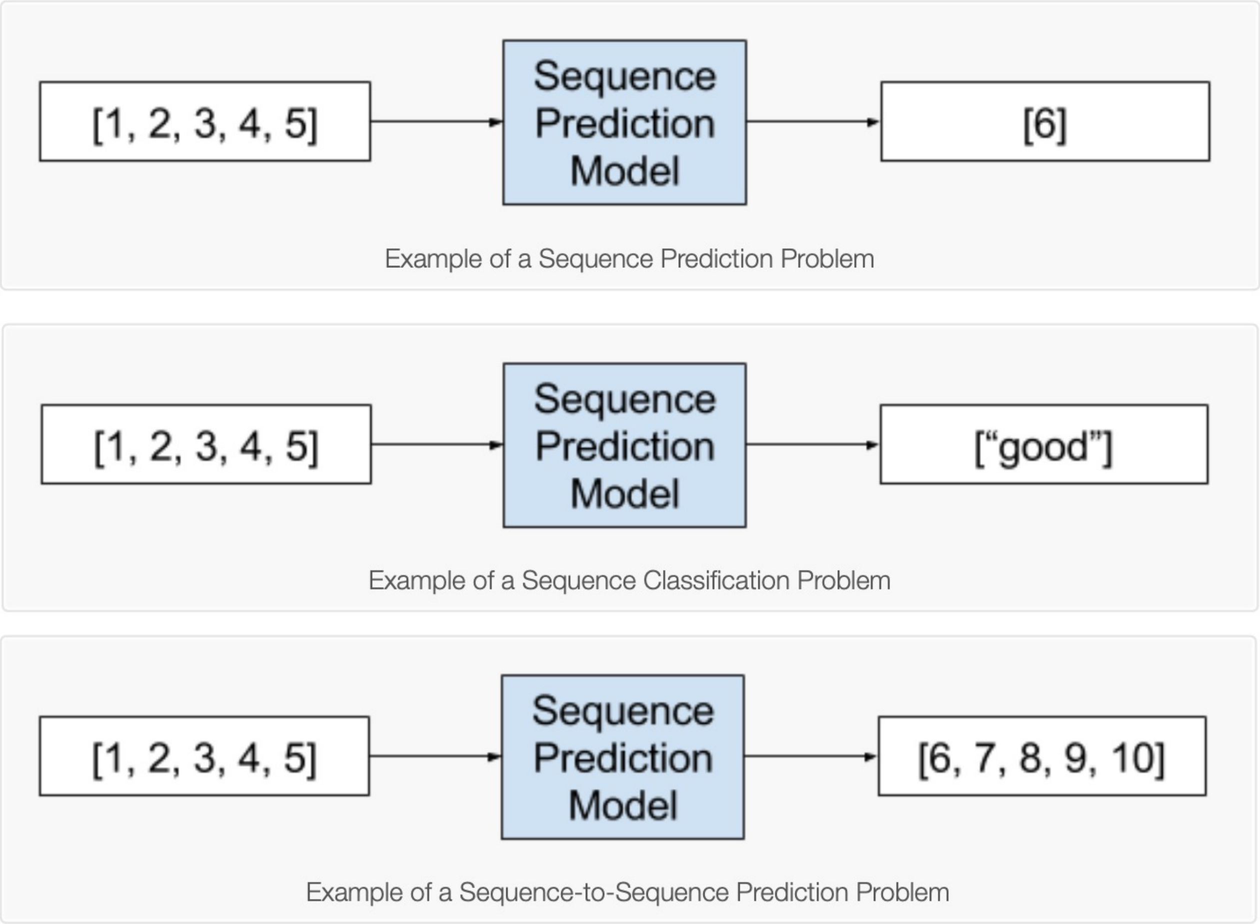


Scenarios:

1. Combine medical imagery with other data sources (medical records, results of tests, etc.)
2. Get insights from large data volumes, where correlations are not clear

https://news.developer.nvidia.com/satellite-images-help-track-a-vehicle/

5. Models on sequential data



Scenarios:

1. Classification of sequential data (for example, multiple medical images, or test results)
2. Predict treatment outcome
3. Predict future development of disease – how medical images or test results in the future will look

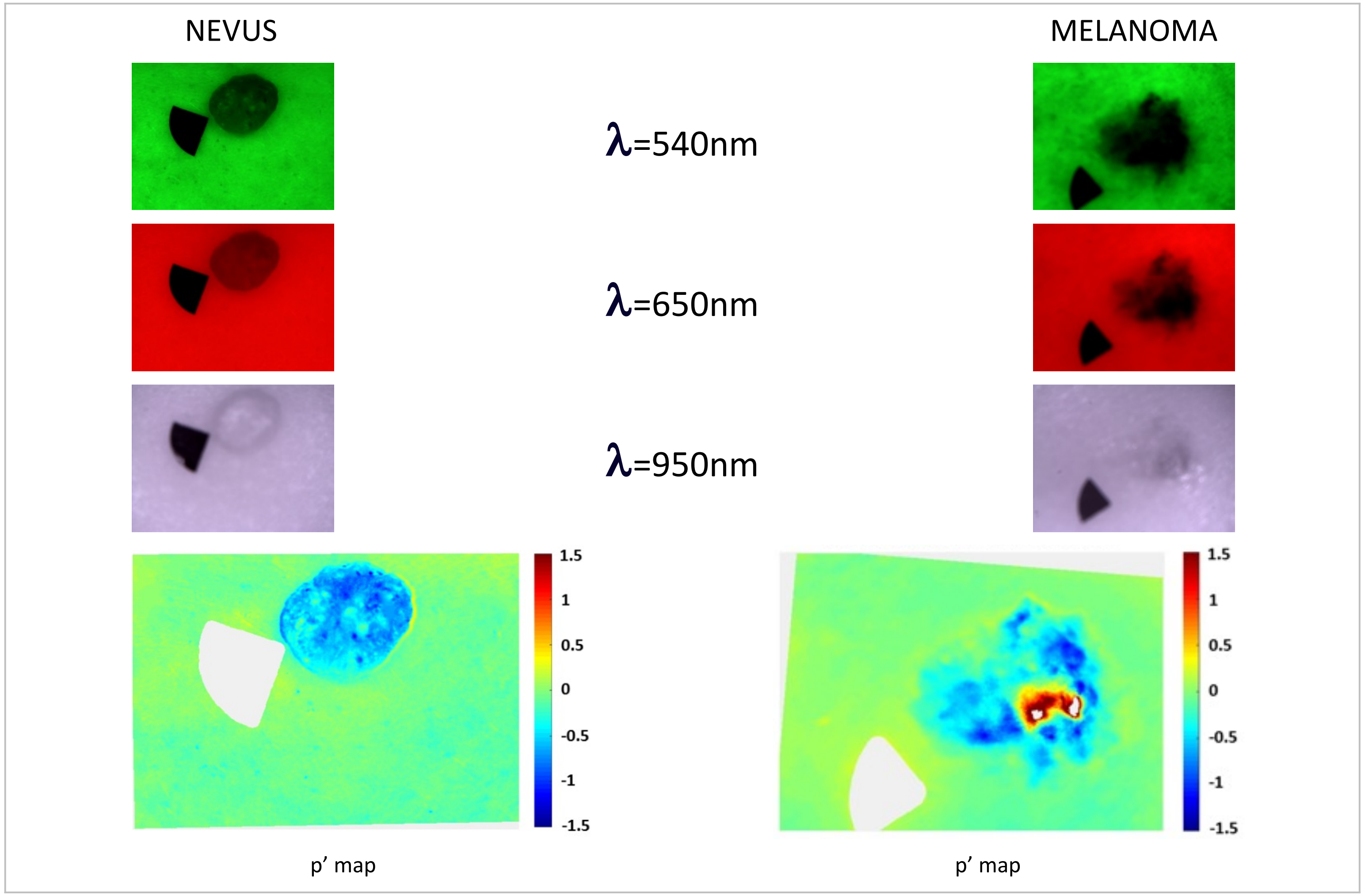
Making Predictions with Sequences: https://machinelearningmastery.com/sequence-prediction/

Examples from real projects

Melanoma Detection

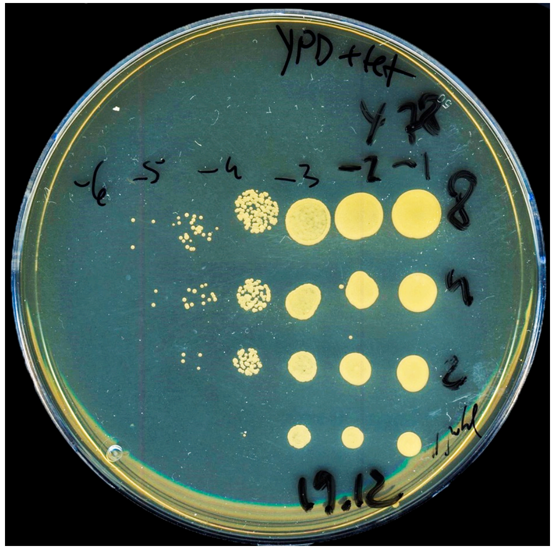
In collaboration with Riga Technical University and University of Latvia

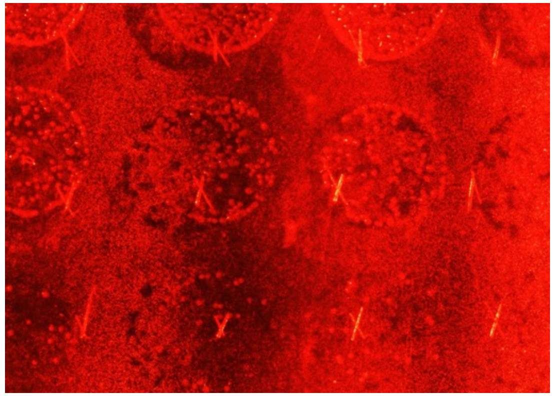
The system uses 550nm, 650nm and 950nm LEDs for diffuse reflectance imaging

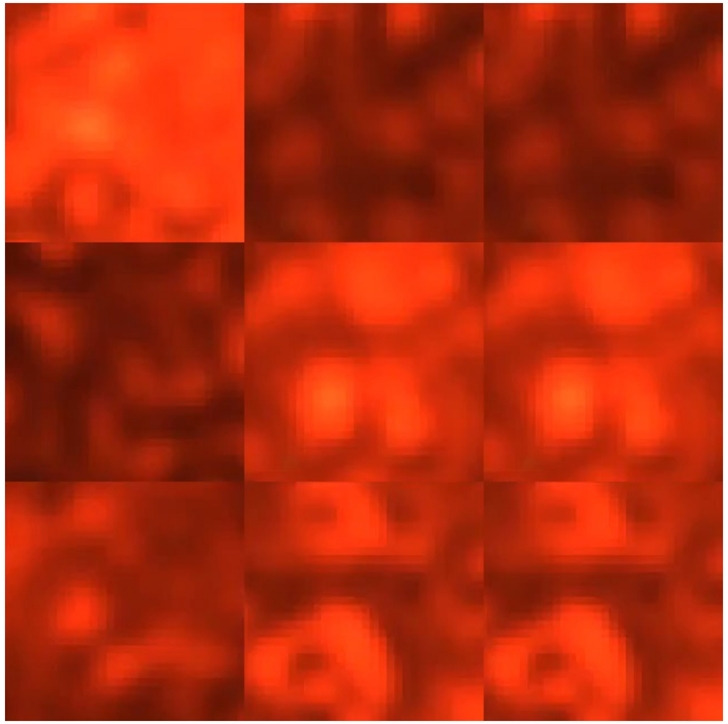


Microorganisms Growth Analysis

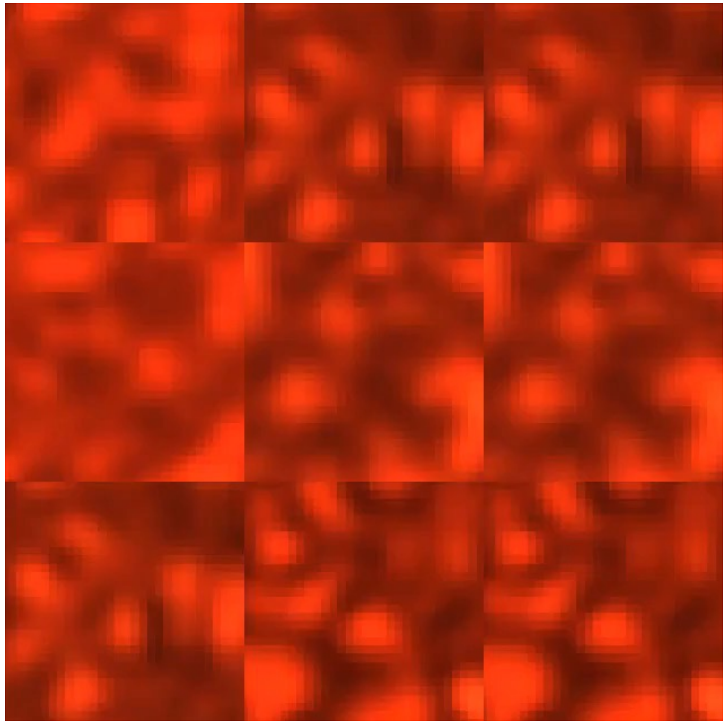
In collaboration with Riga Technical University and University of Latvia







No bacteria



With bacteria

Endometrioma Detection in Loparoscopic Surgery

In collaboration with Riga Technical University

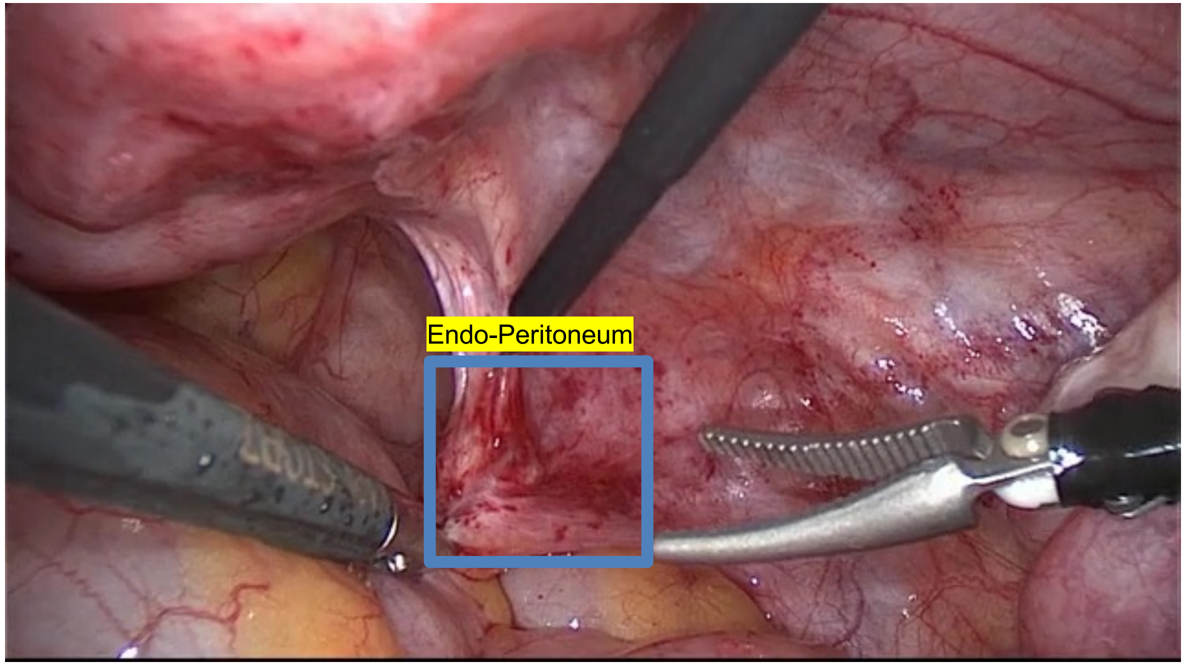


Figure A. Example of the object detection task annotation of the peritoneum endometrioma

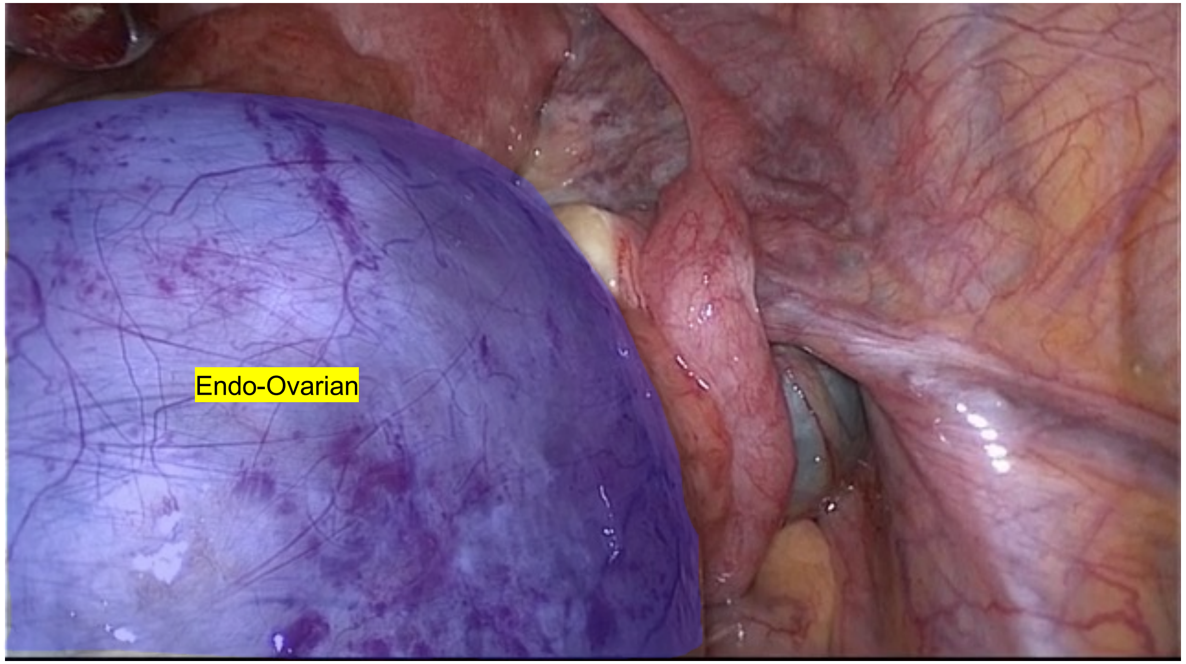


Figure B. Example of the Segmentation task annotation of the ovarian endometrioma

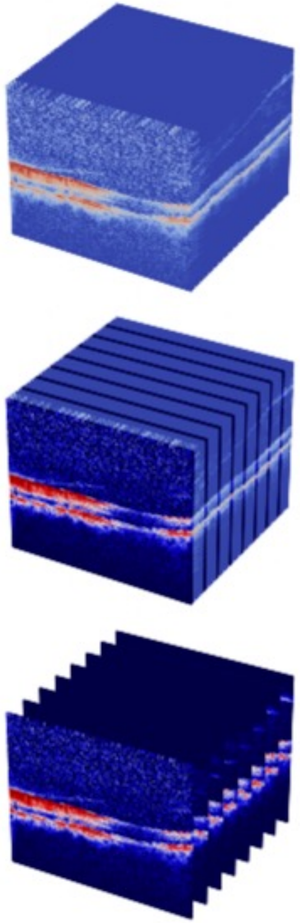
The goal is to provide real-time/near real-time aid/suggestions to surgeons during the surgery.

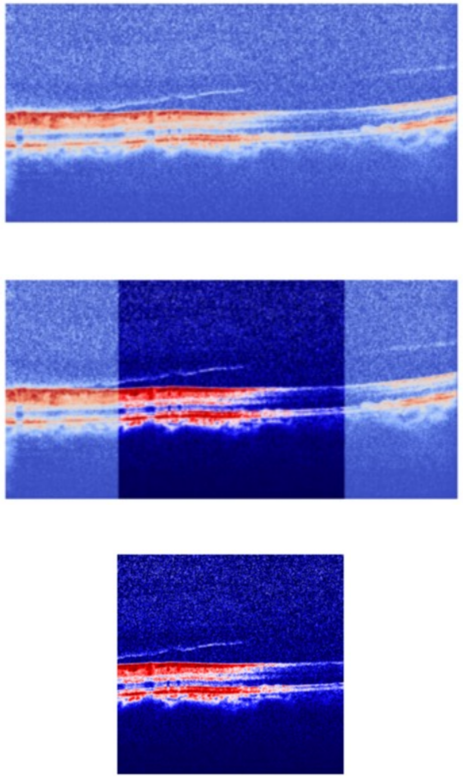
Dataset included labels for Object Detection (Fig. A) and Segmentation (Fig. B) tasks. Two sets of experiments were conducted to train the best model on available data, to perform detection (9 classes) and segmentation (3 classes) of objects on images.

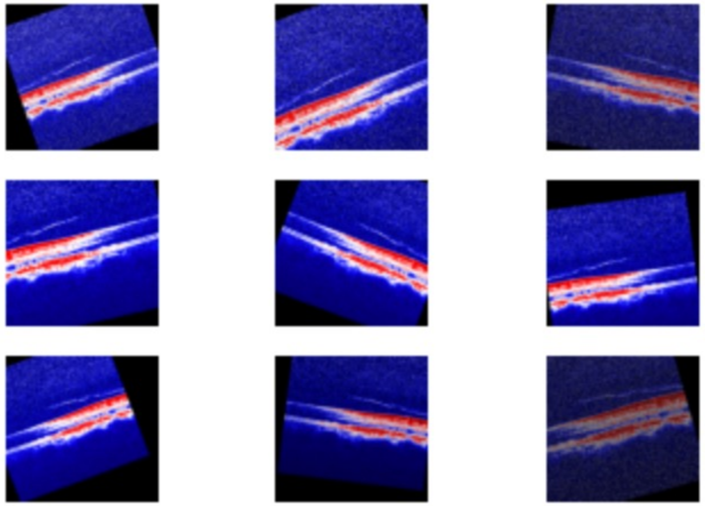
Note: Due to privacy restrictions visualizations above (Fig A, B) come not from actual dataset used during the research, but from similar public dataset.

Detecting disease progression based on retina OCT scan
(Optical Coherence Tomography Analysis)

In collaboration with University Hospital of Zürich


1. Pick 20 b-scans from OCT volume


2. Crop middle part to handle different scan angles


3. Augment image to increase dataset and avoid overfitting (training dataset only)

Detecting conversion to exudative neovascular age-related macular degeneration using machine learning: https://iovs.arvojournals.org/article.aspx?articleid=2774002

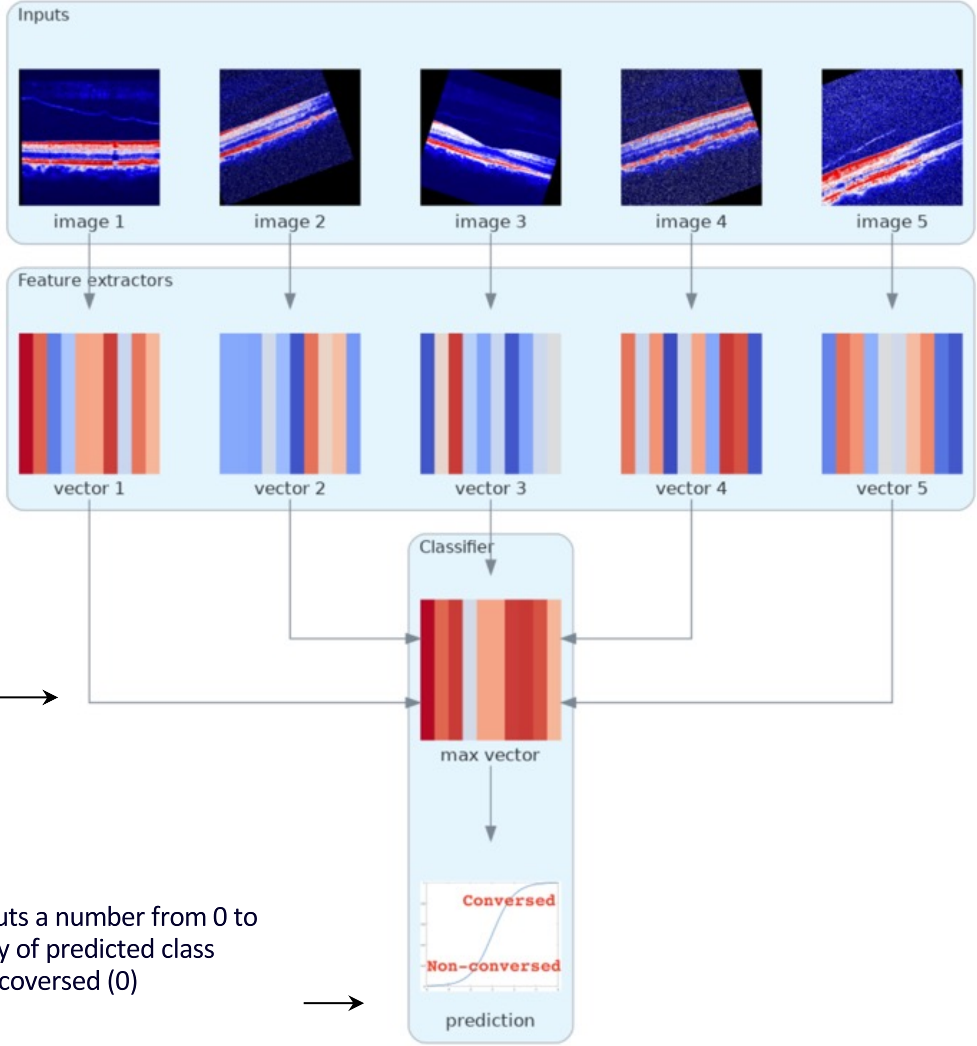
Detecting disease progression based on retina OCT scan

In collaboration with University Hospital of Zürich

Deep Convolutional Neural Network is applied to each b-scan individually to extract vectors of 512 features describing the image

Global Max Pooling is applied on these vectors to extract the most prominent features

Classification network outputs a number from 0 to 1, that represents probabilty of predicted class being conversed (1) or non-coversed (0)



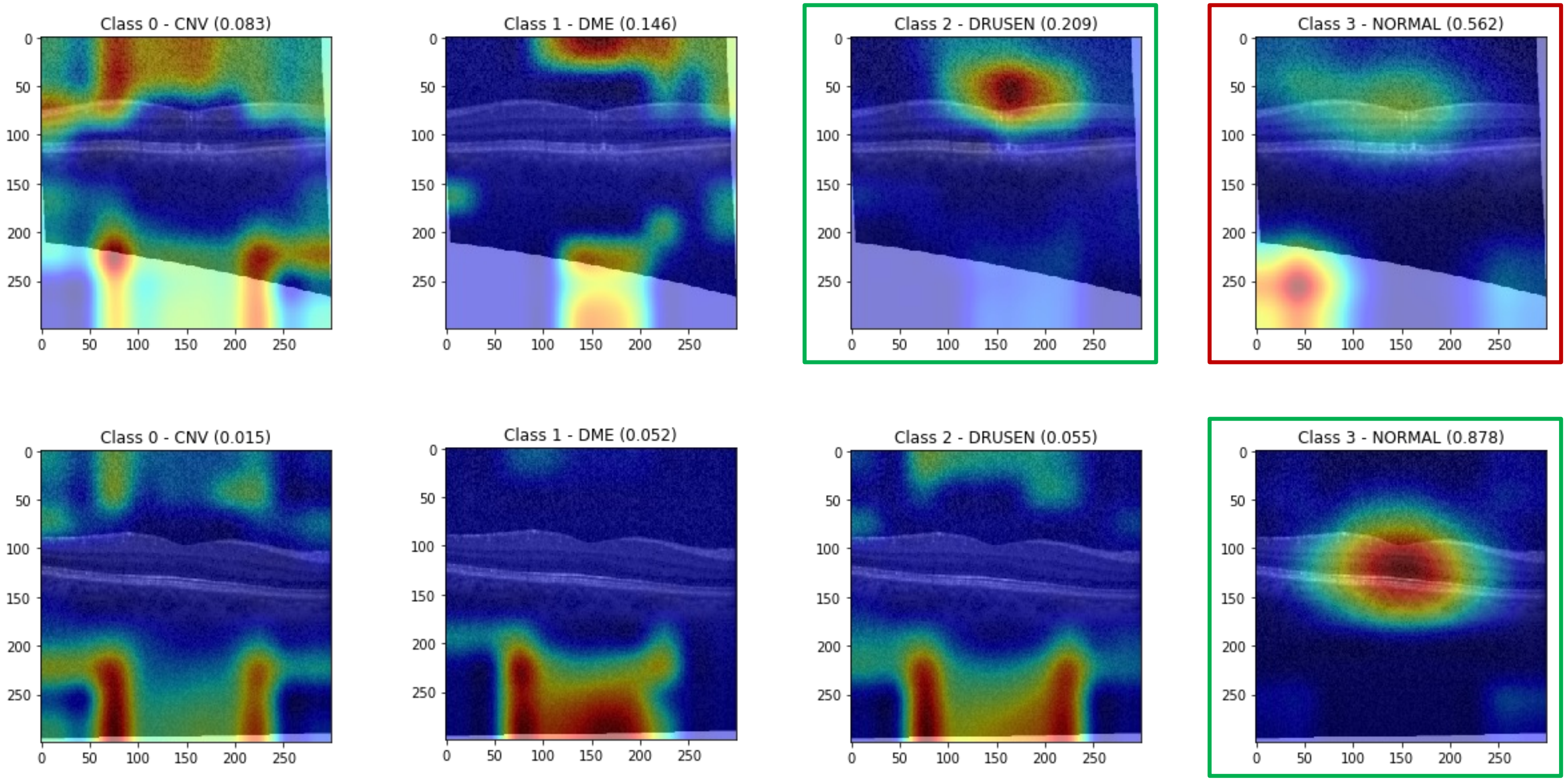
Detecting conversion to exudative neovascular age-related macular degeneration using machine learning: https://iovs.arvojournals.org/article.aspx?articleid=2774002

Detecting disease progression based on retina OCT scan

In collaboration with University Hospital of Zürich

Wrong prediction

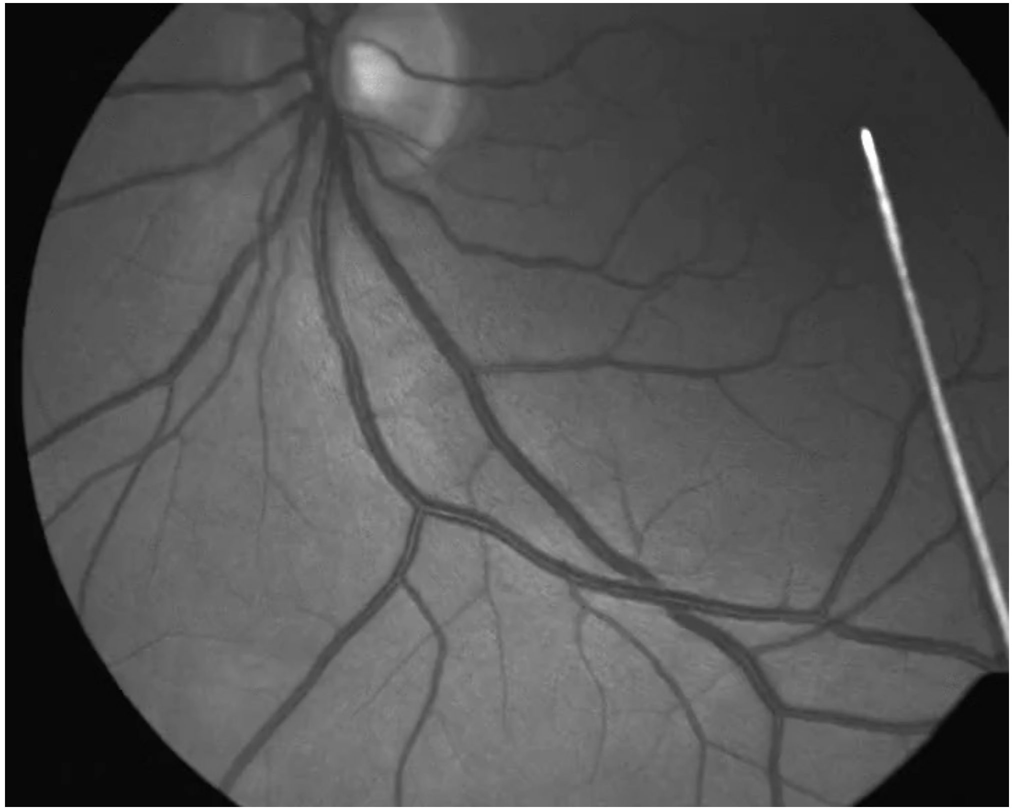
Correct prediction



Using saliency maps to explain model predictions

Detect Spontaneous Venous Pulsations based on RVA videos

In collaboration with University Hospital of Zürich



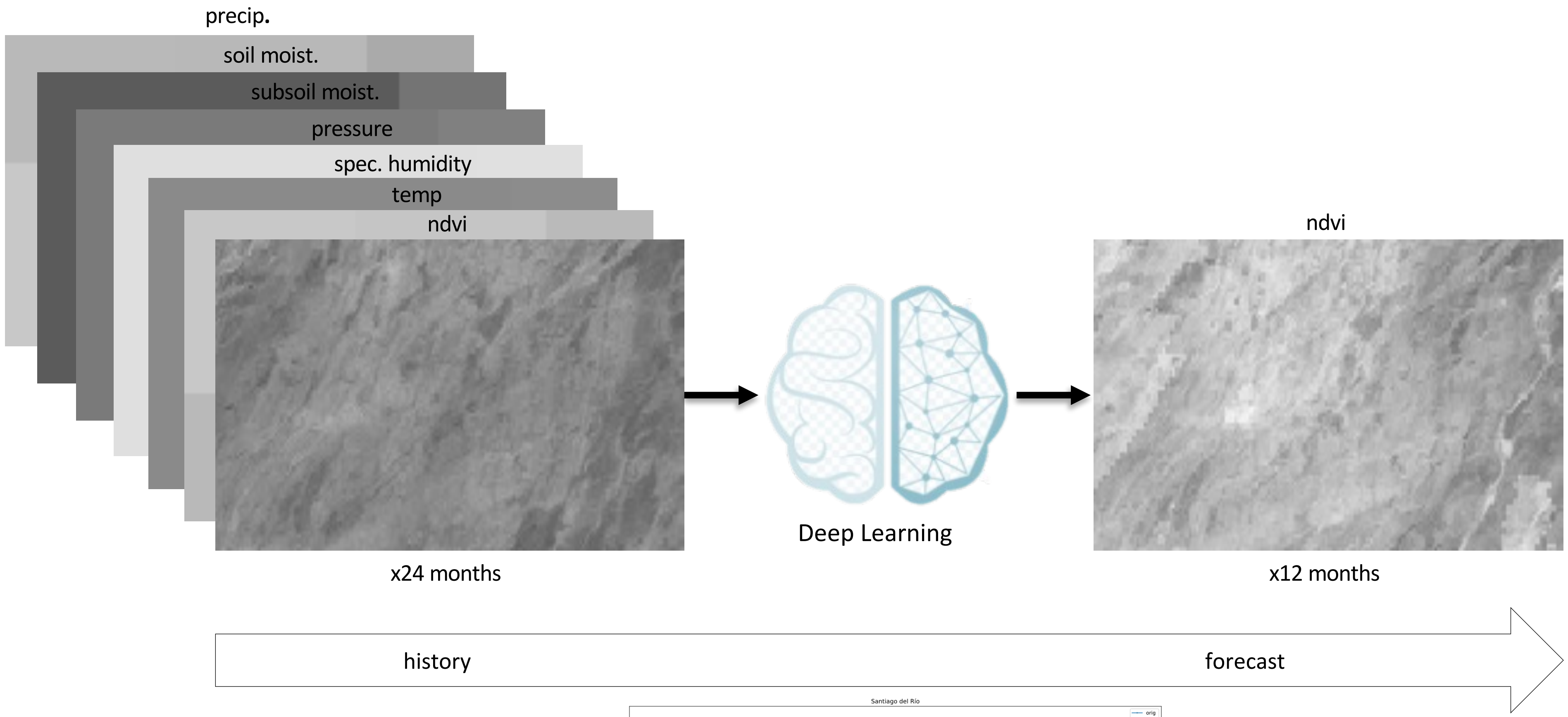
ORIGINAL

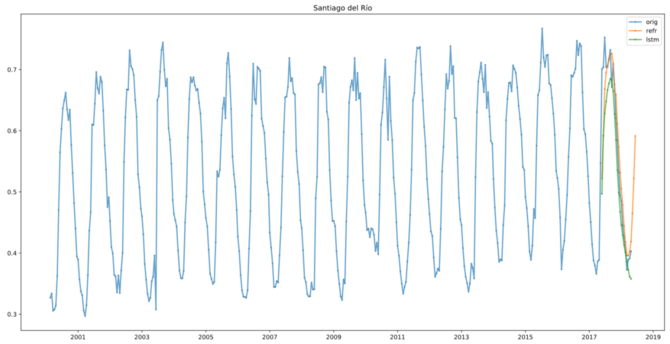
PROCESSED BY AI

First Zurich Machine Intelligence in Clinical Neuroscience Symposium: https://micnlab.com/symposium2021/
Dr. Timothy Hamann - At the Pulse of Time: Machine Vision in Retinal Videos: https://www.youtube.com/watch?v=MqqRNjXplYU

Multi-channel forecasting for parametric insurance

NDVI forecasting (Normalized Difference Vegetation Index)





Why implementing AI in medicine is not that easy?

1. AI relies on data…

…and data is not always of a good quality

Original Article

Lack of Reduction in Racial Disparities in Cancer-Specific Mortality Over a 20-Year Period

Ayal A. Aizer, MD, MHS[1]; Tyler J. Wilhite, BA[2]; Ming-Hui Chen, PhD[3]; Powell L. Graham, AB[4]; Toni K. Choueiri, MD[5]; Karen E. Hoffman, MD, MPH, MHS[6]; Neil E. Martin, MD, MPH[4]; Quoc-Dien Trinh, MD[7]; Jim C. Hu, MD, MPH[8]; and Paul L. Nguyen, MD[4]

BACKGROUND: To the authors' knowledge, it remains unknown whether race-based differences in cancer outcomes have changed with time. In the current study, the authors assessed whether racial disparities in cancer-specific mortality have improved over the last 20 years. **METHODS:** The Surveillance, Epidemiology, and End Results program was used to identify 2,713,474 patients diagnosed between 1988 and 2007 with either lung, breast, prostate, or colorectal cancer (the leading 3 causes of cancer-related mortality among each sex). After exclusions, 1,001,978 patients remained eligible for analysis. The impact of race on cancer-specific mortality was assessed using the regression model of Fine and Gray; an interaction model evaluated trends over time. **RESULTS:** African Americans presented with a more advanced stage of disease ($P < .001$) and underwent definitive therapy less often ($P < .001$) than whites. After adjustment for demographics and year of diagnosis, African Americans were found to have higher estimates of cancer-specific mortality than whites for all cancers combined (hazards ratio, 1.28; 95% confidence interval, 1.26-1.30 [$P < .001$]) and within each individual cancer (each $P < .05$). These differences did not change significantly between 1988 through 1997 and 1998 through 2007, except among patients with breast cancer, in whom survival disparities increased. These findings remained significant after adjustment for stage of disease at presentation and receipt of definitive therapy (hazards ratio for breast cancer mortality in African Americans vs whites: 1.37 from 1988-1997 and 1.53 from 1998-2007; P for interaction, $< .001$). **CONCLUSIONS:** The survival gap for African Americans has not closed over time. Race-based differences in outcome persist independent of stage of disease and treatment, suggesting that additional strategies beyond screening and improving access to care, such as further research into tumor biologies disproportionately affecting African Americans, are needed to improve survival for African American patients with cancer. ***Cancer* 2014;120:1532–9.** *© 2014 American Cancer Society.*

KEYWORDS: African American, Asian American, black, cancer, Hispanic, metastatic, mortality, radiation, surgery, white.

It is disappointing that over the last 20 years the survival gap between African American patients and white patients does not appear to be closing. To the extent that this persistent gap reflects differences in biology, there needs to be increased research specifically among African Americans with cancer, and further emphasis on ensuring that African Americans are represented in clinical trials, because African Americans may be nearly 30% less likely to participate in a trial than white patients.[27] Possible reasons for the limited participation of African Americans in clinical trials include a mistrust of the medical system given prior abuses, limited communication regarding trial availability to African American patients, and economic factors.[28] Clinical trial enrollment, or lack thereof, may be

https://onlinelibrary.wiley.com/doi/epdf/10.1002/cncr.28617
https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/

Linguistic Features Identify Alzheimer's Disease in Narrative Speech

Kathleen C. Fraser[a], Jed A. Meltzer[b] and Frank Rudzicz[a,c,*]
[a]*Department of Computer Science, University of Toronto, Toronto, Canada*
[b]*Rotman Research Institute, Toronto, Canada*
[c]*Toronto Rehabilitation Institute-UHN, Toronto, Canada*

Handling Associate Editor: Peter Garrard

Accepted 20 August 2015

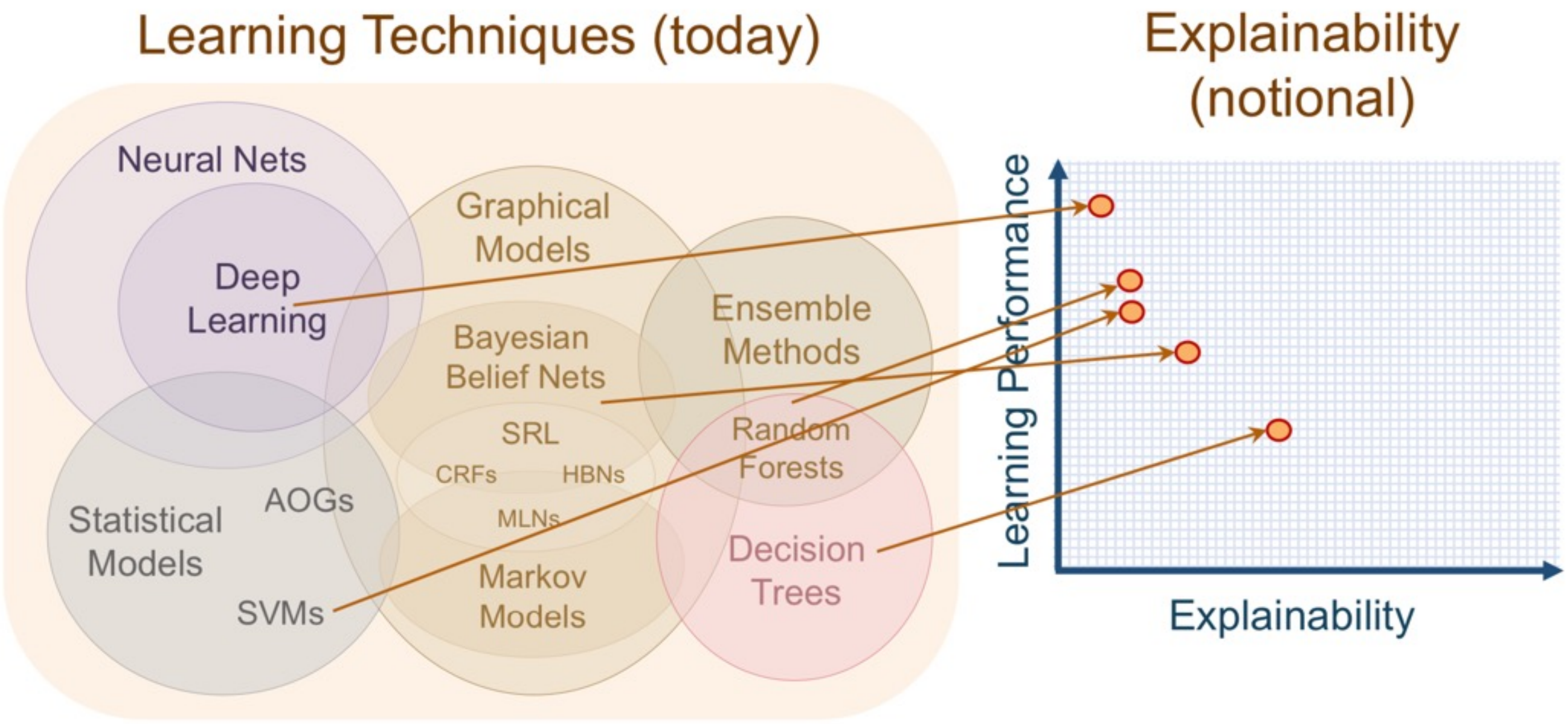
Abstract.
Background: Although memory impairment is the main symptom of Alzheimer's disease (AD), language impairment can be an important marker. Relatively few studies of language in AD quantify the impairments in connected speech using computational techniques.
Objective: We aim to demonstrate state-of-the-art accuracy in automatically identifying Alzheimer's disease from short narrative samples elicited with a picture description task, and to uncover the salient linguistic factors with a statistical factor analysis.
Methods: Data are derived from the DementiaBank corpus, from which 167 patients diagnosed with "possible" or "probable" AD provide 240 narrative samples, and 97 controls provide an additional 233. We compute a number of linguistic variables from the transcripts, and acoustic variables from the associated audio files, and use these variables to train a machine learning classifier to distinguish between participants with AD and healthy controls. To examine the degree of heterogeneity of linguistic impairments in AD, we follow an exploratory factor analysis on these measures of speech and language with an oblique promax rotation, and provide interpretation for the resulting factors.
Results: We obtain state-of-the-art classification accuracies of over 81% in distinguishing individuals with AD from those without based on short samples of their language on a picture description task. Four clear factors emerge: semantic impairment, acoustic abnormality, syntactic impairment, and information impairment.
Conclusion: Modern machine learning and linguistic analysis will be increasingly useful in assessment and clustering of suspected AD.

Imagine there was a simple test to see whether you were developing Alzheimer's disease. You would look at a picture and describe it, software would assess the way you spoke, and based on your answer, tell you whether or not you had early-stage Alzheimer's. It would be quick, easy, and over 90% accurate—except for you, it doesn't work.

That might be because you're from Africa. Or because you're from India, or China, or Michigan. Imagine most of the world is getting healthier because of some new technology, but you're getting left behind.

http://www.cs.toronto.edu/~kfraser/Fraser15-JAD.pdf
https://qz.com/1367177/if-ai-is-going-to-be-the-worlds-doctor-it-needs-better-textbooks/

2. Modern AI is a black box…



Explainable Artificial Intelligence (XAI), David Gunning, DARPA/I20, Nov 2017: https://www.darpa.mil/attachments/XAIProgramUpdate.pdf

Input → BLACK BOX → Output

Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana, Paul Koch, Yin Lou, Marc Sturm , Johannes Gehrke, Noemie Elhadad

Download BibTex

KDD'15, August 10-13, 2015, Sydney, NSW, Australia | August 2015
Published by ACM
View Publication

In machine learning often a tradeoff must be made between accuracy and intelligibility. More accurate models such as boosted trees, random forests, and neural nets usually are not intelligible, but more intelligible models such as logistic regression, naive-Bayes, and single decision trees often have significantly worse accuracy. This tradeoff sometimes limits the accuracy of models that can be applied in mission-critical applications such as healthcare where being able to understand, validate, edit, and trust a learned model is important. We present two case studies where high-performance generalized additive models with pairwise interactions (GA2Ms) are applied to real healthcare problems yielding intelligible models with state-of-the-art accuracy. In the pneumonia risk prediction case study, the intelligible model uncovers surprising patterns in the data that previously had prevented complex learned models from being fielded in this domain, but because it is intelligible and modular allows these patterns to be recognized and removed. In the 30day hospital readmission case study, we show that the same methods scale to large datasets containing hundreds of thousands of patients and thousands of attributes while remaining intelligible and providing accuracy comparable to the best (unintelligible) machine learning methods.

ing [1]. Although models based on rules were not as accurate as the neural net models, they were *intelligible*, i.e., interpretable by humans. On one of the pneumonia datasets, the rule-based system learned the rule "HasAsthama(x) ⇒ LowerRisk(x)", i.e., that patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia than the general population. Needless to say, this rule is counterintuitive. But it reflected a true pattern in the training data: patients with a history of asthma who presented with pneumonia usually were admitted not only to the hospital but directly to the ICU (Intensive Care Unit). The good news is that the aggressive care received by asthmatic pneumonia patients was so effective that it lowered their risk of dying from pneumonia compared to the general population. The bad news is that because the prognosis for these patients is better than average, models trained on the data incorrectly learn that asthma lowers risk, when in fact

https://www.microsoft.com/en-us/research/wp-content/uploads/2017/06/KDD2015FinalDraftIntelligibleModels4HealthCare_igt143e-caruanaA.pdf
https://www.microsoft.com/en-us/research/publication/intelligible-models-healthcare-predicting-pneumonia-risk-hospital-30-day-readmission/

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

Craig M. Bennett[1], Abigail A. Baird[2], Michael B. Miller[1], and George L. Wolford[3]

[1] Psychology Department, University of California Santa Barbara, Santa Barbara, CA; [2] Department of Psychology, Vassar College, Poughkeepsie, NY;
[3] Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for multiple comparisons should be completed with these datasets, but is often ignored by investigators. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

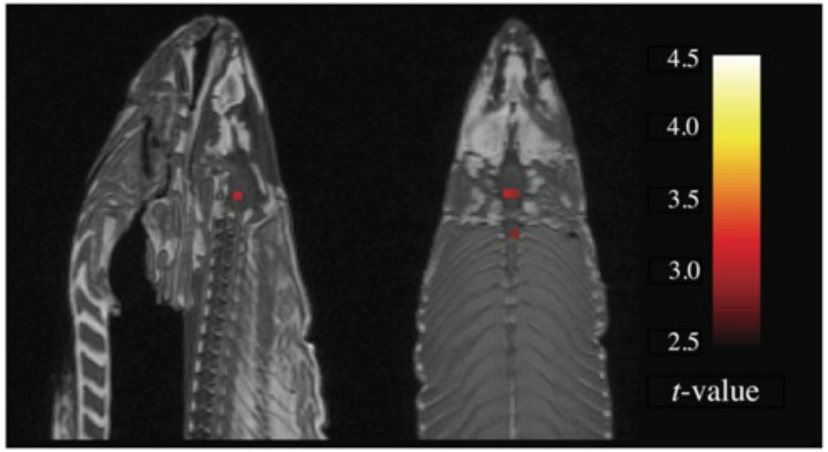
METHODS

Subject. One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

GLM RESULTS



A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, p(uncorrected) < 0.001, 3 voxel extent threshold.

Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm^3 with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed.

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction:
by Craig M. Bennett , Abigail A. Baird , Michael B. Miller , George L. Wolford http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.161.8384

ARTICLES

https://doi.org/10.1038/s42256-021-00338-7

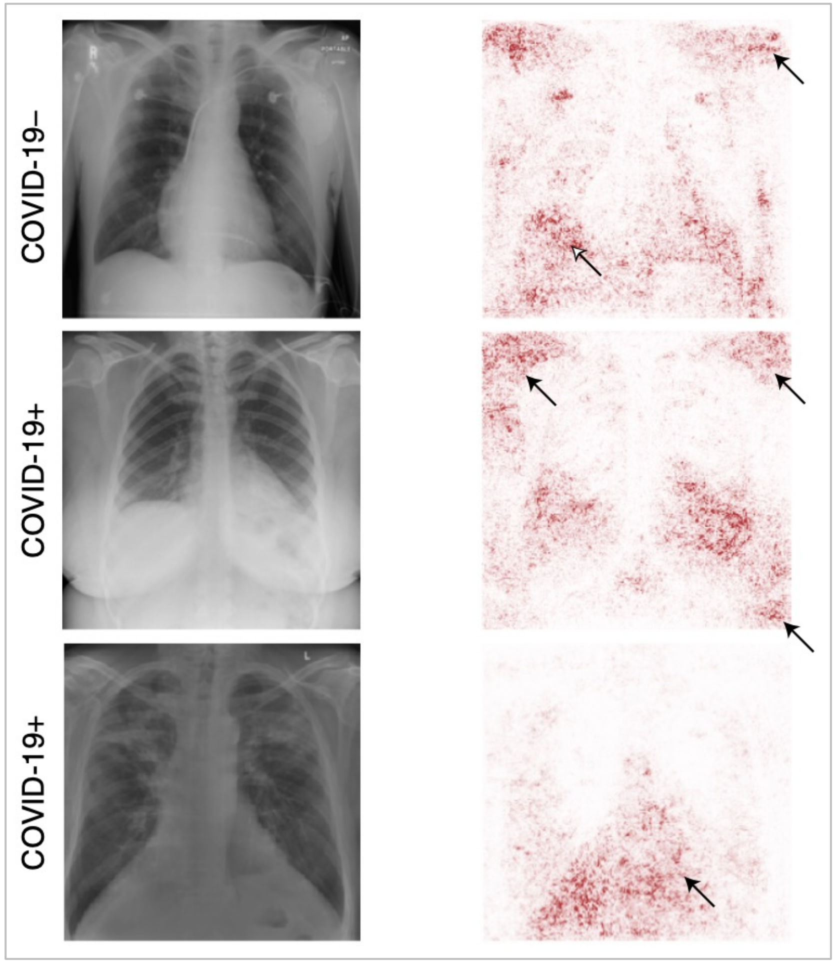
nature machine intelligence

Check for updates

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave[1,2,3], Joseph D. Janizek[1,2,3] and Su-In Lee[1] ✉

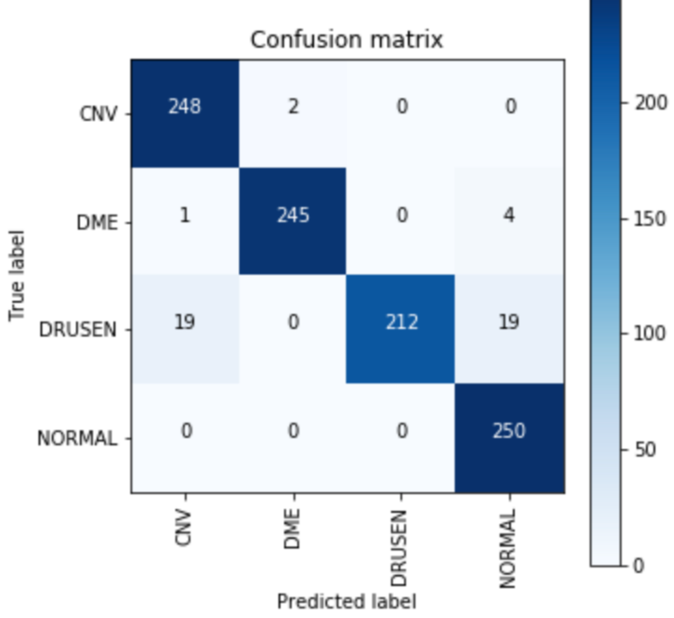
Artificial intelligence (AI) researchers and radiologists have recently reported AI systems that accurately detect COVID-19 in chest radiographs. However, the robustness of these systems remains unclear. Using state-of-the-art techniques in explainable AI, we demonstrate that recent deep learning systems to detect COVID-19 from chest radiographs rely on confounding factors rather than medical pathology, creating an alarming situation in which the systems appear accurate, but fail when tested in new hospitals. We observe that the approach to obtain training data for these AI systems introduces a nearly ideal scenario for AI to learn these spurious 'shortcuts'. Because this approach to data collection has also been used to obtain training data for the detection of COVID-19 in computed tomography scans and for medical imaging tasks related to other diseases, our study reveals a far-reaching problem in medical-imaging AI. In addition, we show that evaluation of a model on external data is insufficient to ensure AI systems rely on medically relevant pathology, because the undesired 'shortcuts' learned by AI systems may not impair performance in new hospitals. These findings demonstrate that explainable AI should be seen as a prerequisite to clinical deployment of machine-learning healthcare models.



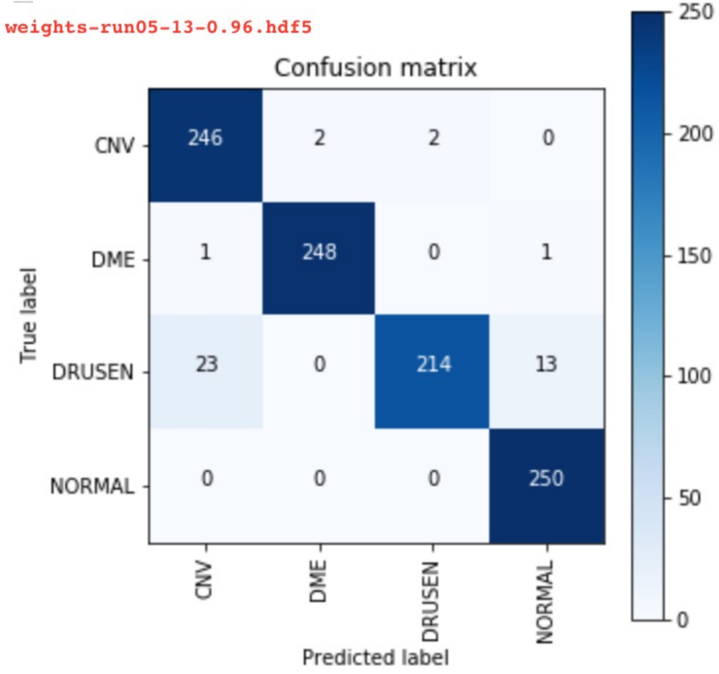
AI for radiographic COVID-19 detection selects shortcuts over signal (2021): https://www.nature.com/articles/s42256-021-00338-7

3. Accuracy % are impressive, but when AI is wrong it looks stupid to experts…

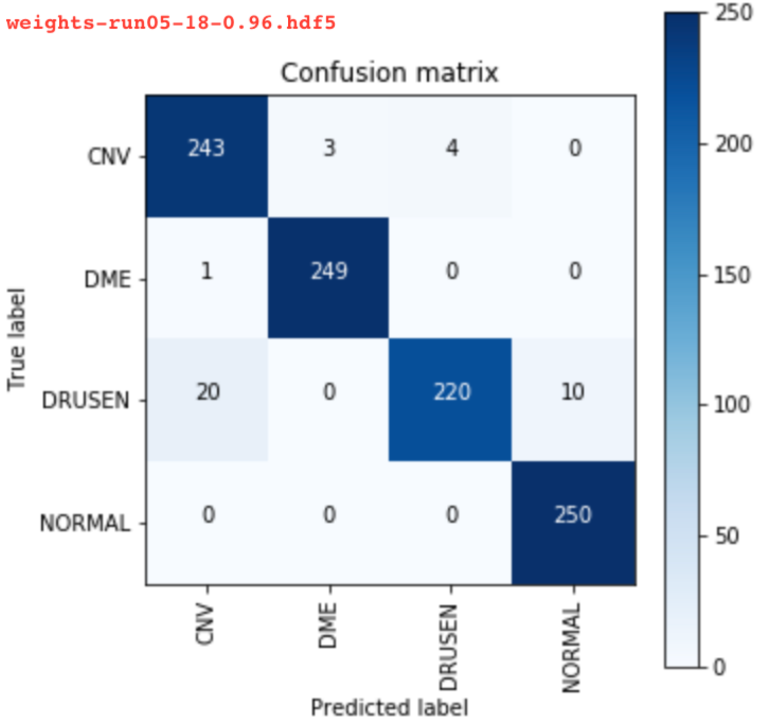
weights-run05-10-0.96.hdf5



weights-run05-13-0.96.hdf5



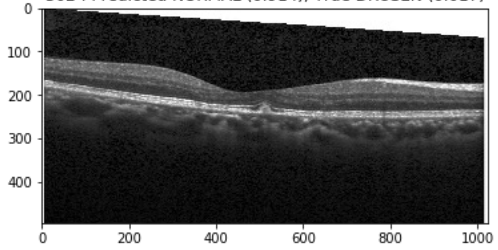
weights-run05-18-0.96.hdf5



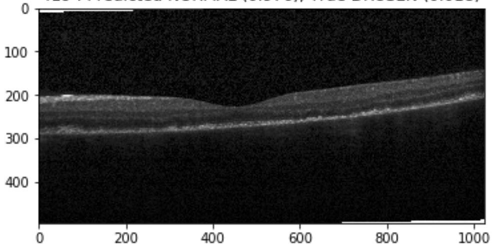
Examples where something went terribly wrong

True class is **NOT** Normal, but Normal was predicted with probability > 70%

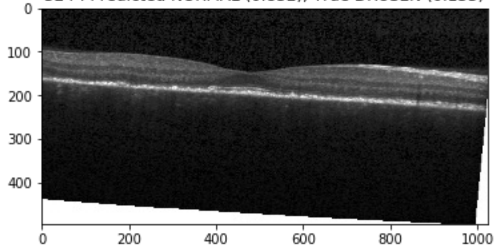
302 : Predicted NORMAL (0.914), True DRUSEN (0.017)



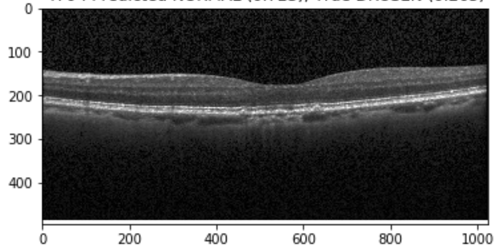
419 : Predicted NORMAL (0.976), True DRUSEN (0.018)



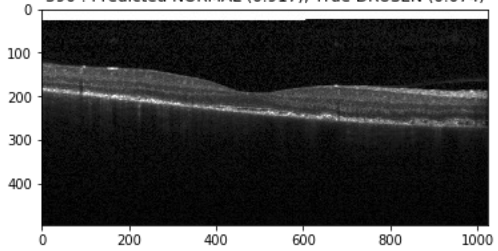
324 : Predicted NORMAL (0.832), True DRUSEN (0.133)



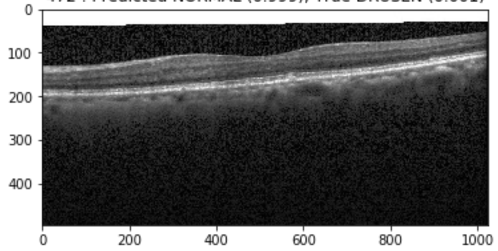
470 : Predicted NORMAL (0.723), True DRUSEN (0.265)



390 : Predicted NORMAL (0.917), True DRUSEN (0.074)



472 : Predicted NORMAL (0.999), True DRUSEN (0.001)



Show and Tell: A Neural Image Caption Generator, O. Vinyals, A. Toshev, S. Bengio, D. Erhan, 2015: http://arxiv.org/abs/1411.4555v2

> A man riding a motorcycle on a beach.

"a young boy is holding a baseball bat."

"a cat is sitting on a couch with a remote control."

"a woman holding a teddy bear in front of a mirror."

"a horse is standing in the middle of a road."

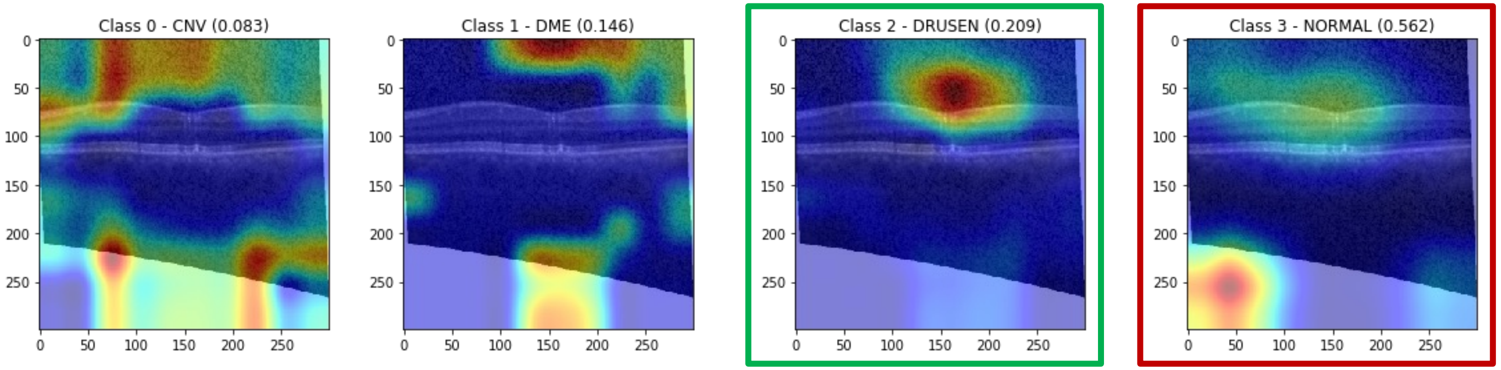
Accelerating innovation and powering new experiences with AI: https://code.facebook.com/posts/310100219388873/accelerating-innovation-and-powering-new-experiences-with-ai/
CS231n: Convolutional Neural Networks for Visual Recognition, Lecture 10: Recurrent Neural Networks: http://cs231n.stanford.edu/slides/winter1516_lecture10.pdf

Visualizing predictions

True class - **DRUSEN**

Predicted probabilities:

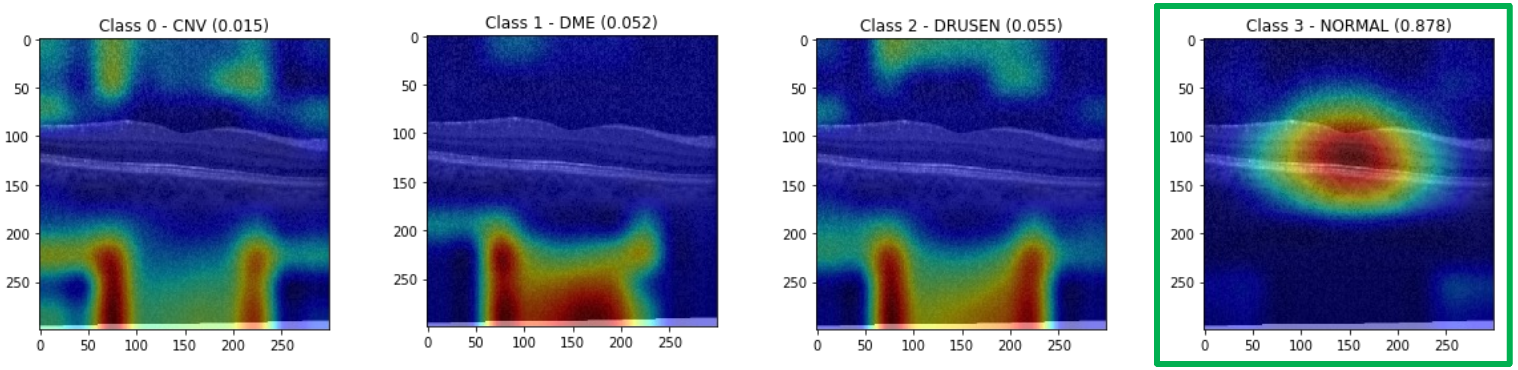
- CNV: 0.083
- DME: 0.146
- DRUSEN: 0.209
- NORMAL: 0.562



True class - **NORMAL**

Predicted probabilities:

- CNV: 0.015
- DME: 0.052
- DRUSEN: 0.055
- NORMAL: 0.878



Key takeaways

Key takeaways

1. AI will not replace doctors in near future

2. Real applications of AI in medicine:
   - Medical Administration – claims, billing, fraud detection, audit
   - Decision Support – aggregating data, finding similar cases
   - Generating New Insights – discover unknown correlations in medical data
   - Self-Service Diagnostics
   - Image Analysis:
     - Pre-processing images (segmentation, classification, find anomalies)
     - AI as a "second opinion"
     - AI controlling that formal procedures are followed accurately
     - Predictive maintenance for medical devices

3. Remember about "not alive" Atlantic Salmon and COVID patient "pose from RTG"

Thanks!

Slides: https://link.makingaireal.com/rsu2024