Acta Scientific COMPUTER SCIENCES

Volume 4 Issue 5 May 2022

Review Article

And so this is Christmas and what has AI done in 2021?

Christian Mancas*

DATASIS ProSoft srl, Bucharest, Romania

***Corresponding Author:** Christian Mancas, DATASIS ProSoft srl, Bucharest Romania.

Received: December 25, 2021

Published: April 18, 2022

Abstract

Starting from some recent remarks of Elon Musk on AI, this minireview editorial discusses a few important narrow AI achievements from 2021.

Keywords: Artificial Intelligence; Neural Networks; CLIP; Codex; Computer Vision Models; DALL-E; DINO; GSLM; HuBERT; Machine Learning Models; Pathways; PAWS; Perceiver; SEER; Speech Recognition and Generation; T0; Transformer-based Models; Translating Natural Language to Code

Introduction

Artificial Intelligence (AI) is born in 1956, had some first 20 years of success, then a first 10 year "winter", followed by a short revival, then a much longer "winter", and, finally, since some a quarter of a century it is flourishing exponentially, at least according to Elon Musk [1].

Elon Musk's recent remarks on AI [1] are both pertinent and scaring:

- "I think that the danger of AI is much greater than the danger of nuclear warheads, by a lot".
- "Mark my words: AI is far more dangerous than nukes!"
- "I tried to convince people to slow down AI, to regulate AI. This was futile. I tried for years".
- "The biggest issue I see with the so-called AI experts is that they think they know more than they do, and they think they're smarter than they actually are. This tends to plague smart people. They define themselves by their intelligence and they don't like the idea that a machine could be way smarter than them so that they discount the idea, which is fundamentally flawed, this is the wishful thinking situation".
- "I am very close to the cutting edge in AI and it scares the hell out of me. It's capable of vastly more than almost anyone knows, and the rate of improvement is exponential".
- "It feels like we are the biological bootloader for AI, effectively".
- "We are all of us already Cyborgs. So you have a machine extension of yourself in the form of your phone and your computer and all your applications. You are already Superhumans. By far you have more power and more capability than the President of the United States had 30 years ago".
- "The fact is that we've got regulators in the aircraft industry, car industry, drugs, food, at anything that's sort of public risk and I think this has to fall in the category of public risk".
- "I'm not really worried about the short-term stuff, things that are like narrow AI is not a species-level risk: it would result in dislocations, in lost jobs, better weaponry, and that kind of thing. But it is not a fundamental species-level risk. Whereas digital superintelligence is. So, it's all about laying the groundwork to make sure that if humanity collectively decides that creating digital superintelligence is the right move, then we should do so very, very carefully".

While digital superintelligence research is both almost unavailable to us and still in its infancy, narrow AI research and products are not only accessible, but, indeed, impressively abundant. This paper briefly discusses only 11 of the most important (in our view) ones that emerged this year.

Neural networks

DALL-E

2021 saw another Open AI marvel come to life: the neural network DALL-E [2] that creates images from text. Although being a smaller version of GPT-3, it has an exceptional zero-shot performance, generating extremely high-quality images even for abstract, unreal, or absurd objects.

DALL-E is a language-transformer AI model trained on some 250 million pairs of texts and images (collected over the Internet, holding some 12 billion parameters of autoregressive transformers (taken from GPT-3), whose input is a sequence of tokens and that is trained to maximize the likelihood of sequential token generation. Training has two stages:

- The first one is image compression, so as to reduce the transformer's size without significant quality degradation, yielding a sequence of 1024 tokens (32×32 grid of image tokens).
- The second one is concatenation of 256 BPE-encoded text tokens with image tokens and training an autoregressive transformer.

DALL-E surpasses all the earlier generative models especially by its ability to widen its knowledge on previously unseen texts:

- It produces anthropomorphized versions of animals and objects, by combining completely unrelated things and by applying transformations to existing images.
- It can modify objects' attributes, the number of times they appear on an image, and visualize different perspectives in three-dimensional space.
- It can even draw objects' internal structure (which requires deep knowledge of objects and that previously required specific training).

DALL-E can help artists and designers to generate design ideas: no longer need to draw sketches, you can simply write your thoughts and choose among the options it produces, also using CLIP to detect the best ones.

CLIP

Open AI also developed this year the neural network CLIP (Contrastive Language-Image Pre-training) [3]. CLIP provides a powerful bridge between natural language processing and computer vision. CLIP does not aim to recognize the objects on images, but only to provide their most appropriate description. Its 77.1% accuracy result on an adversarial dataset explicitly designed to confuse AI models is impressive.

CLIP was trained on 400 million Internet image – text caption pairs. The model is made of two sub-models, a text encoder and an image one. Both convert texts and images into a mathematical vector space that allows comparing how close objects are; then, CLIP tries to maximize the similarity between texts and their corresponding images.

CLIP may be used without fine-tuning, as semantics extracted from the text is used to add value to the images. CLIP also learns from unfiltered and noisy data, which is both adding to its flexibility and robustness and increases its accuracy on real-world data.

Dually, although CLIP a multimodal neuron responsible for abstract thinking, it still struggles with more abstract and systematic tasks (e.g., counting the objects on an image) and has difficulties with industry-specific classification (e.g., determining the car model).

Computer vision models

SEER

SEER (SElf-supERvised) is a self-supervised computer vision model developed by Facebook AI [4] that can learn from any random set of images, w/o needing any preprocessing or labeling. It was pre-trained on a billion random public images from Instagram, reaching 84.2% accuracy on ImageNet.

Its new algorithm SwAV works with a large number of unlabeled pictures quickly clustering related visual concepts by their similarities. Its second major component is a convolutional

neural network (based on RegNet models) that works with large and complex data without losing accuracy. Moreover, for the development of SEER, Facebook AI also released an all-purpose library for self-supervised learning.

SEER allows directly using data already existing in the world rather than specifically preparing it. Obviously, training models on real-life data increases accuracy and generalizing ability, while saving time and resources that would otherwise be wasted on manual data preparation and labeling. Moreover, self-supervised learning also mitigates biases arising during data annotation.

Facebook uses SEER on their platforms for safety concerns (e.g., rapidly identifying and removing hate and racist images) and automatic generation of descriptions for images, as well as better items categorization. Moreover, Facebook hopes that SEER's efficiency and speed make it suited for medical purposes too, including disease diagnosis.

DINO and PAWS

Simultaneously released by Facebook AI, DINO and PAWS [5] are two new methods for model training. DINO trains Vision Transformers (ViT) w/o supervision, which is a powerful combination of self-supervised learning and transformers that discovers and segments objects in images or videos w/o segmentation-targeted objectives. PAWS is a semi-supervised approach, which optimizes model training and produces state-of-the-art results using much less computing power. When combined, DINO and PAWS significantly enhance computer vision systems, making them more efficient and less dependent on labeled data.

DINO's self-supervision is based on label-free self-distillation and two identical networks - a student one and a teacher one. Both take a same image as input, but in different vector representations: the teacher gets the global idea of the image (obtained from two great dimensions partially overlapped patches), while the student receives a local representation of the image (acquired by a series of smaller patches).

During training, the student matches local views to the global ones, trying to understand whether they represent a same image, after which the teacher performs classification based only on the global views, trying to match the output obtained by the student.

PAWS needs a small amount of labeled data: given an unlabeled training image, several views are generated using random data augmentations and transformations. PAWS is trained to make the representations of these views as similar as possible; its algorithm then uses a random subsample of labeled images for assigning a pseudo-label to the unlabeled views; such assignments are conducted by comparing the representations of the unlabeled views and the labeled samples; finally, the model is updated by minimizing a standard classification loss between the pseudo-labels of pairs of views of the same unlabeled image.

DINO's self-supervised learning process allows training of highly accurate models with unlabeled data. PAWS dramatically reduces training time using a small set of labeled examples, also solving the common issue for self-supervised methods, i.e. collapsing representations when all images get mapped to a same one. PAWS is focused on efficiency rather than performance; for example, a standard ResNet-50 model needs only one percent of the labels in ImageNet and 10 times fewer steps to reach the same accuracy as previous models when trained with PAWS. PAWS is an excellent tool for domains with few annotated images, which includes medical imagery.

DINO produces easily interpretable features and is also one of the best at identifying image copies, even though it was not trained to do that.

Transformer-based models

Perceiver

Developed by DeepMind, Perceiver is a new state-of-the-art transformer-based model [6] that works with multimodal data: just as human brains simultaneously analyze data received from all our sense organs, Perceiver processes data received in different formats. To reduce time complexity, its self-attention layer was replaced by a cross-attention one and all its inputs (be them image, audio, or sensor data) are converted into bytes. Moreover (which was borrowed from the Set Transformer), to dramatically reduce the training time, Perceiver creates a summary version of each data sample. Jeff Dean, Google AI's Lead, has described the Perceiver as "the model that can handle any task, and learn faster, with fewer data".

Perceiver scales to hundreds of thousands of inputs of different formats, thus opening new possibilities for general perception

architectures. Even overfitting, unavoidable in such big compelling models, is greatly reduced.

T0

T0 [7] is a series of publicly available encoder-decoder models developed by Hugging Face. The zero in its name stands for "zero-shot task generalization", as this model works even with data and problems it has never seen before. Hugging Face claims that T0 outperforms GTP-3 on many tasks, although being 16 times smaller.

Based on Google's T5 transformer-based language model, T0 contains 11 billion parameters and was fine-tuned by thousands of additional training steps on new problems with previously unseen inputs on multiple English supervised datasets converted into prompts. Each such prompt has several templates constructed by various formulations, which allowed T0 in the end to perform well on entirely new tasks.

The encoder part receives input text, while the decoder produces the target text. T0 is fine-tuned for autoregressively generating the target text through standard maximum likelihood training.

T0 is a multi-task language model, i.e. it performs inference on many natural language processing tasks (e.g., sentiment analysis, question answering, text summarization, topic classification, paraphrase identification, etc.) outperforming the GPT-3 model (considered one of the best in the field) in many tasks, although being much smaller.

However, T0 requires non-trivial computational resources and cannot work with computer code (unlike Codex) or non-English texts.

Speech recognition and generation

HuBERT

HuBERT [8] is another Facebook AI remarkable achievement: an approach for self-supervised learning speech representations. To learn the spoken input structure, HuBERT uses an offline k-means clustering step for predicting a proper cluster for masked audio segments. HuBERT learns both acoustic and language models from its inputs.

HuBERT first builds meaningful representations from its unmasked audio inputs; then, for reducing the prediction error, it uses masked prediction to learn representations by capturing the long-run temporal relationships between them. HuBERT focuses on modeling its input data sequential structure using K-means mapping. Only the masked regions are subject to the predictive loss, which makes HuBERT learn representations of unmasked inputs for inferring the targets of masked ones.

HuBERT allows developing systems and models trained on audio only, w/o having to translate it into text. HuBERT's self-supervision technology eliminates the need for large data volumes, thus making it possible to quickly and accurately develop solutions for new languages as well.

HuBERT is used for profounder speech understanding, improving existing audio, simplifying development of new ones, and, as a preprocessing tool, for further machine learning models (especially in natural language processing tasks). HuBERT is also great at audio compressing Speech representations from HuBERT are also used for synthesizing speech.

GSLM

Facebook AI claims that their GSLM (Generative Spoken Language Model) [9] is the first high-performing NLP (neuro-linguistic programming) model that solely relies on audio (i.e. is textless).

GSLM gathered researchers from various teams across Facebook AI: signal processing specialists, speech processing engineers, data scientists, computer linguists, psycholinguists, etc.

GSLM has three components:

- A HuBERT encoder that converts speech into discrete units (called pseudo-text) representing sounds.
- An auto-regressive language model that predicts the following units of text based on the previously ones (a multistream transformer with multiple heads being used for this purpose).
- A Tacotron 2 decoder that converts units back into sounds.

GSLM has been trained on more than 6,000 hours of audiobooks with no text or labels.

To increase the speed and accuracy of NLP applications, GSLM removes automatic speech recognition (ASR) systems from the

language processing pipeline (as ASR systems are known for their resource-intensive operations and poor performance).

GSLM also allows working with rare languages that have little or no written texts.

Sentiment analysis, translation, summarization, question-answering, etc. are possible with GSLM. Moreover, GSLM also captures other significant features like tone, emotions, or intonations, hopefully making it possible to develop a first universal translator.

Translating natural language to code: Codex

Codex [10] is another remarkable system created by Open AI, which is the model that powers GitHub Copilot, being capable of translating natural language to code. Codex is already proficient with some popular programming languages (e.g., Python, JavaScript, Go).

Codex is a descendant of GPT-3 (thus based on transformers), which was trained not only on English, but also on billions of lines of publicly available computer code (including those stored in the public GitHub repositories). Codex produces working code from prompts in natural language. Codex has 14KB of memory (i.e., 3.5 times more than the 4KB of GPT-3), so it can take into account three times more contextual information than GPT-3.

Codex is not just a simple autocomplete tool, but a real AI-based powerful partner that significantly speeds up the writing code process. For example, instead of browsing through Stack Overflow or alike searching for correct syntax, programmers can simply type what they want in natural language and get code that works, which is significantly timesaving.

Codex was designed to map simple problems to existing code, such as conducting calculations and text preprocessing, creating tests and simple applications, building graphs, processing images, etc. Although it cannot solve complex problems or write entire applications, Codex is the first step towards such full functionality.

Machine learning models: Pathways

Google's Pathways [11] is a next-generation AI architecture that can simultaneously handle multiple tasks. Pathways never does forget what it has learned so far; instead, it always uses its knowledge to learn new tasks faster and more effectively.

Pathways has not been yet released: Google is still developing it, so the math behind it is unknown. What has been announced is that the model will be trained on multiple datasets with different inputs. When working on a problem, Pathways will use only the necessary parts (so not all its neurons will be activated).

Pathways addresses many of the weaknesses of existing systems and aggregates their strengths. For example, instead of doing only one thing, a single model will perform well on thousands of different tasks.

Pathways will enable building multimodal models able to simultaneously perceive different types of inputs (e.g., images, texts, sensor data, etc.).. Such multimodality will clearly provide more insights, making the results less biased.

Pathways will be a huge sparse model, i.e. only a fraction of it will be activated as needed, while current models are dense, i.e. all neurons must be activated for accomplishing a task.

Conclusion

Even if only considering the narrow AI achievements of 2021 briefly discussed in this paper (and there are very many others, e.g., the Boston Dynamics' Stretch robot, the Facebook AI's TimeSformer and Wave2vec, the Google AI's FLAN, the Google Brain's Switch Transformer, the NVIDIA and Microsoft's MT-NLG, etc.), Elon Musk is right on the acceleration of innovations in this field.

Elon Musk's fears are not absurd: he does not think that there will come a day when digital superintelligence surpasses the human one. He only fears that bad guys (be them dictatorships, companies, or individuals) would acquire it (and it doesn't matter whether through cooperation, buying, or stealing) and use it against humans. Moreover, he's also worried that someday the total amount of digital superintelligence will surpass the human one.

Consequently, he advocates for the need to urgently regulate AI as well, before being too late (and he surely knows what he is talking about: for example, yes, a couple a days ago a woman gave birth in the front seat of a Tesla that was driving on Autopilot, but, unfortunately, other Tesla cars on Autopilot crashed, some of them even killing people…). Will he be understood by politicians before too late?

Bibliography

1. Wayne C. "Warning from Elon Musk" (2021).
2. Ramesh A., *et al*. "DALL-E: Creating Images from Text" (2021).
3. Radford A., *et al*. "CLIP: Connecting Text and Images" (2021).
4. Goyal P., *et al*. "SEER: The start of a more powerful, flexible, and accessible era for computer vision" (2021).
5. Bojanowski P., *et al*. "Advancing the state of the art in computer vision with self-supervised Transformers and 10x more efficient training" (2021).
6. Jaegle A., *et al*. "Perceiver: General Perception with Iterative Attention" (2021).
7. "Hugging Face". T0pp (2021).
8. Hsu WN., *et al*. "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units" (2021).
9. Adi Y., *et al*. "Textless NLP: Generating expressive speech from raw audio" (2021).
10. Chen M., *et al*. "Evaluating Large Language Models Trained on Code" (2021).
11. Dean J. "Introducing Pathways: A next-generation AI architecture" (2021).