COMPUTER VISION

Dr. Saúl Alonso-Monsalve

saul.alonso.monsalve@cern.ch

ETH Zurich

51st SLAC Summer Institute

August 9, 2023

About me

- BS and MS in Computer Science (Madrid, Spain).
- PhD at CERN (Geneva, Switzerland).
  - Deep learning for neutrino event reconstruction (DUNE and T2K experiments).
- Currently: senior researcher at ETH Zurich (Switzerland).
  - Leading the deep-learning efforts at the near detector of the T2K neutrino experiment in Japan.
- Expertise in deep-learning projects: developing CNNs, GNNs, RNNs, Transformers… applied to various domains: neutrinos, medical diagnosis, and cryptocurrencies.

Overview

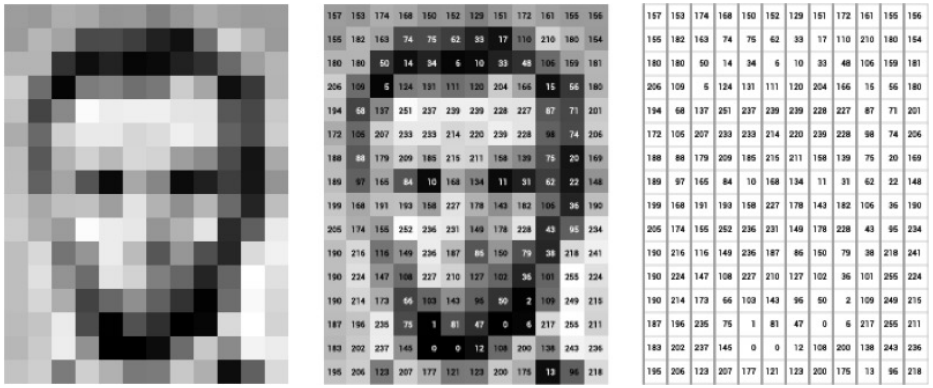
1. **Introduction to computer vision.**
2. Convolutional neural networks (CNNs).
3. Beyond CNNs: exploring other current methods.
4. Generative models.
5. Challenges and future directions.
6. Conclusion.

Computer vision

- Study of visual data.
- Massive amount of visual data produced every day.
- Origin in the late 50s.
  - Started with edge detection, line labeling…
  - Relied on "hand-engineering."
  - Some key breakthroughs: Hough transform [Duda, R.O. & Hart, P.E., 1972], Convolutional neural networks [LeCun, Y. et al., 1989], Viola-Jones object detection [Viola, P. and Jones, M. (2001)].
- Deep learning has been the dominant approach in computer vision research for the past decade.
- Applications in many areas: automotive, healthcare, robotics, media, agriculture, security, physics…

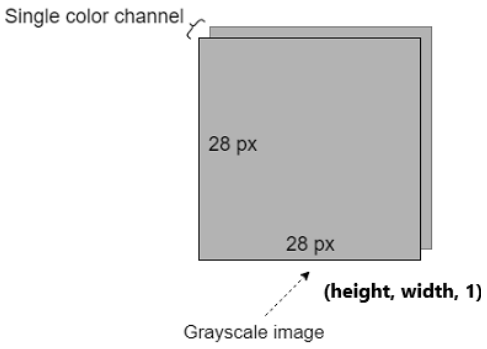
Digital images

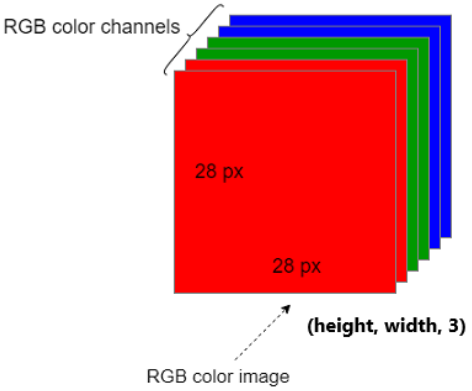
- A computer "sees" a grid of numbers.



[Source: Openframeworks]

- Grayscale images:
  - Pixel values from 0 (black) to 255 (white).
- RGB images:
  - Each color represents a "channel".





Deep learning in computer vision

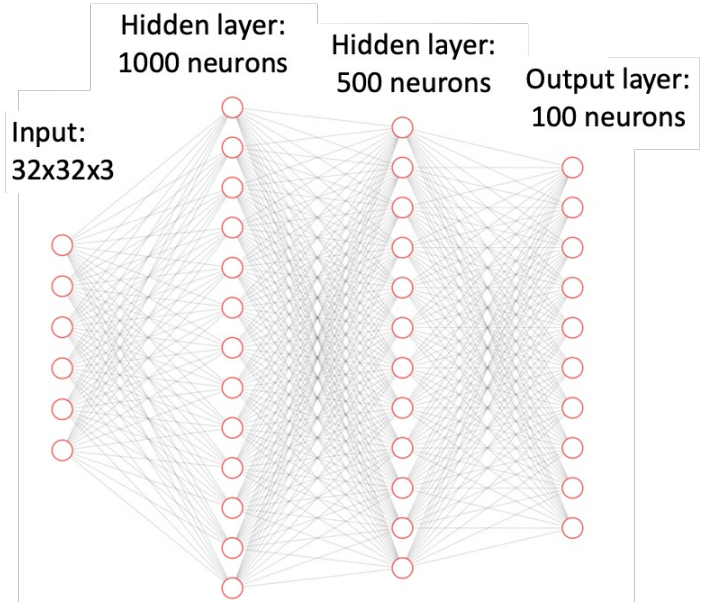
- Deep learning (DL) refers to neural networks with multiple layers (deep neural networks) aimed to solve complex problems (see Monday and Tuesday lectures).
  - Although neural networks have been around for decades, it was not until recently that they became feasible to run on large datasets using available hardware.

- The DL revolution began in 2012, when a convolutional neural network achieved a milestone in image classification by significantly reducing the classification error of a dataset with 10,000 categories and 10 million images [Krizhevsky A. et al., 2012].
  - Classical machine learning techniques, have become less relevant due to the impressive performance of deep neural nets [O'Mahony et al., 2019].

Why not MLPs?

- Fully-connected neural networks (FCNNs), also known as dense neural networks or multi-layer perceptrons (MLPs), are a type of artificial neural network where each neuron in one layer is connected to every neuron in the next layer.
  - MLPs have been used in the early days of computer vision research.
  - However, they are not commonly used for modern computer vision tasks.

- MLPs require a fixed-size one-dimensional input, meaning grids of pixels need to be flattened before being fed to the network, resulting in an extremely large number of parameters .
  - In the example, the input (flattened) has 3,072 values. If the first layer has 1,000 neurons, that's 3,072,000 parameters (without bias) for only the first layer!
  - They are not translation invariant.



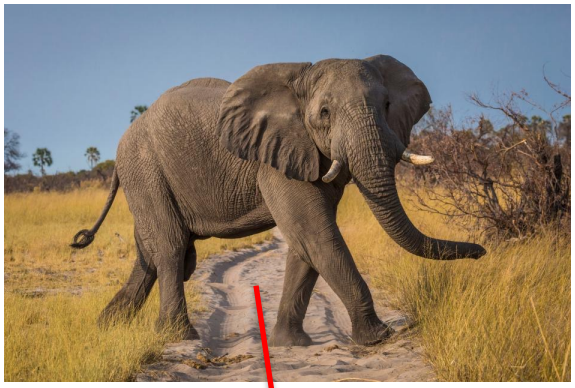
Overview

1. Introduction to computer vision.
2. **Convolutional neural networks (CNNs).**
3. Beyond CNNs: exploring other current methods.
4. Generative models.
5. Challenges and future directions.
6. Conclusion.

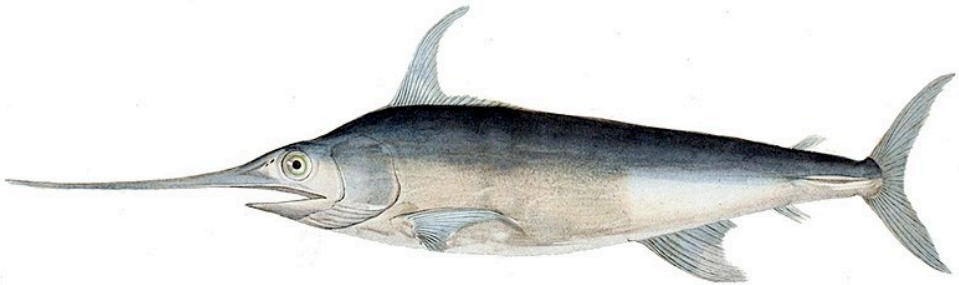
Image recognition (intuition)

- Humans break down images into different parts before assembling the information back together.

It is orange with black stripes: it is a tiger.



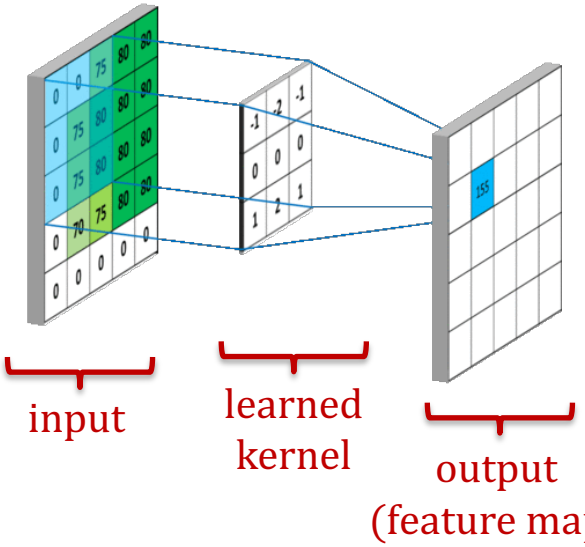
It is gray with a long trunk: it is an elephant.

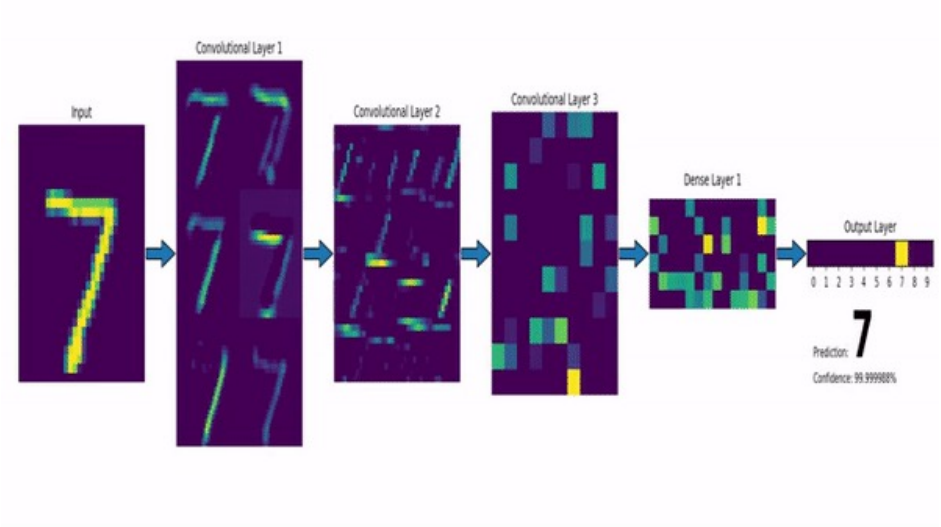


It is gray with a long trunk: it is an elephant.

Convolutional neural networks (CNNs)

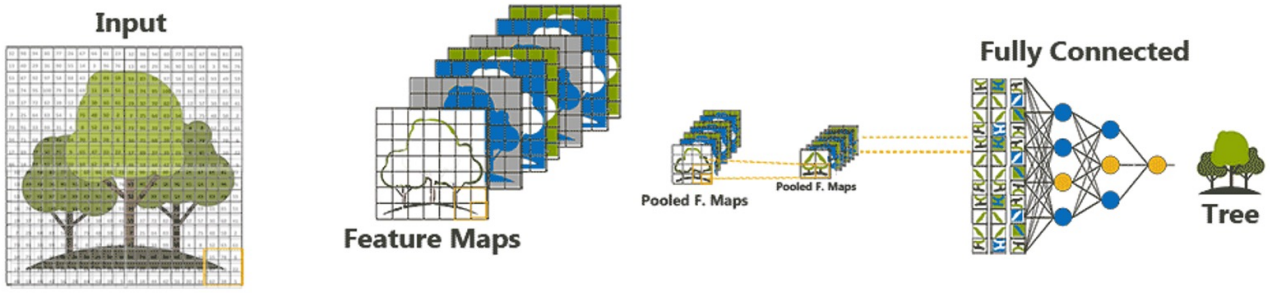
- Convolutional neural networks, or CNNs, are a type of neural network architecture specifically designed for image recognition tasks in computer vision.
  - Convolution: element-wise multiplication and sum of the overlapping elements between the kernel and the input.
  - CNNs use a series of convolutional layers to extract hierarchical features from images.
  - CNNs have achieved state-of-the-art performance in various computer vision tasks, including object detection, image segmentation, and facial recognition.
  - They are translation invariant!





Convolutional neural networks (CNNs)

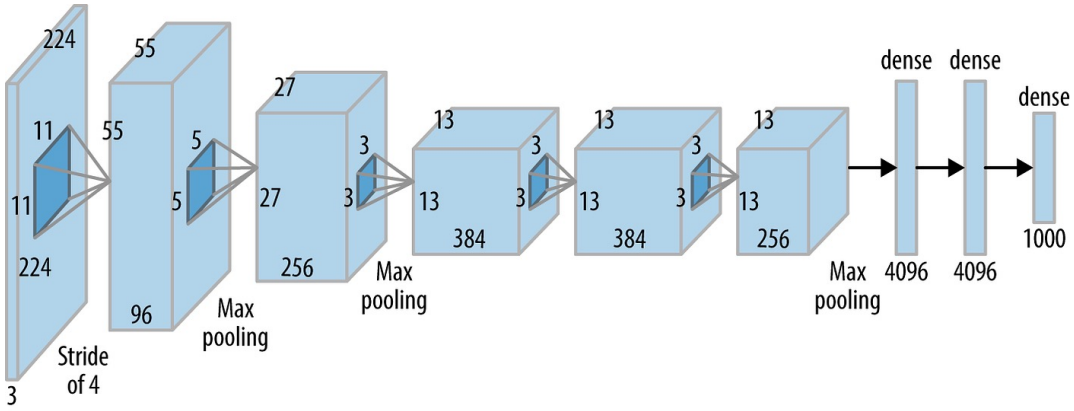
- Convolutional layers, apart from extracting features, also tend to downsample the height and width of the input image/feature map while they increase the number of channels (after concatenating multiple output feature maps).
- Pooling layers, which do not have learnable weights, are used to apply an even stronger downsampling to their input.



[Source: Adatis]

AlexNet (2012)

- Krizhevsky, A. et al., 2012.
- Winner of the ImageNet LSVRC-2012 challenge.
  - Accuracy of 84.7% compared with a 73.8% accuracy of the runner-up.
- ~62M parameters.
  - 5 convolutional layers, max pooling, 3 fully-connected layers.
  - ReLU as an activation function (sigmoid and tanh were the most common activation functions back then for the hidden units).
  - Dropout of 0.5 as regularization.



[Source: oreilly.com]

AlexNet Network - Structural Details													
Input			Output			Layer	Stride	Pad	Kernel size		in	out	# of Param
227	227	3	55	55	96	conv1	4	0	11	11	3	96	34944
55	55	96	27	27	96	maxpool1	2	0	3	3	96	96	0
27	27	96	27	27	256	conv2	1	2	5	5	96	256	614656
27	27	256	13	13	256	maxpool2	2	0	3	3	256	256	0
13	13	256	13	13	384	conv3	1	1	3	3	256	384	885120
13	13	384	13	13	384	conv4	1	1	3	3	384	384	1327488
13	13	384	13	13	256	conv5	1	1	3	3	384	256	884992
13	13	256	6	6	256	maxpool5	2	0	3	3	256	256	0
						fc6			1	1	9216	4096	37752832
						fc7			1	1	4096	4096	16781312
						fc8			1	1	4096	1000	4097000
Total													**62,378,344**

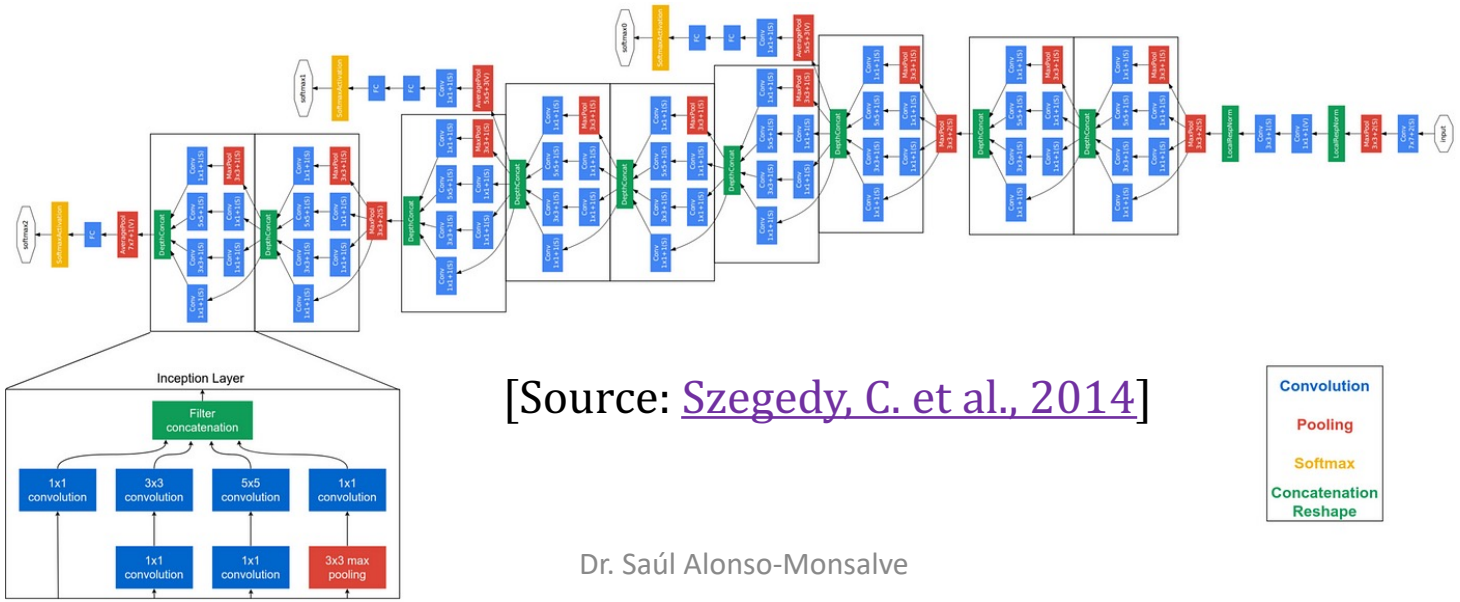
GoogLeNet (2014)

- Szegedy, C. et al., 2014. Also known as *Inception v1*.

- Winner of the ImageNet LSVRC-2014 challenge.
  - Accuracy of 93.3% in classification tasks.

- ~6.4M parameters.
  - Inception module: 1x1 conv, 3x3 conv, 5x5 conv, max pooling (outputs are concatenated).
  - The network "decides" the best kernel size in in each module!

- First CNN used in a neutrino experiment! (NOvA, see L. Whitehead lecture later!)



[Source: Szegedy, C. et al., 2014]

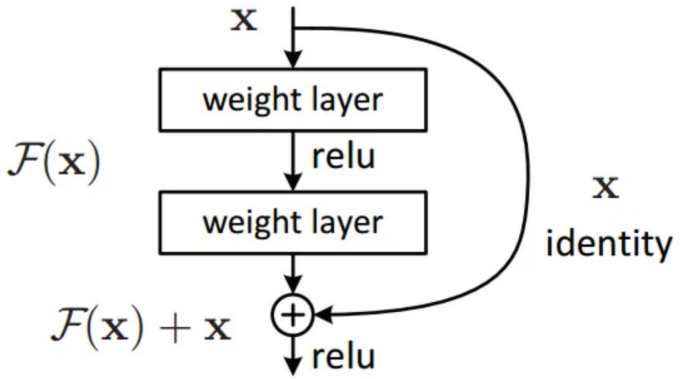
ResNet (2015)

- He, K. et al., 2015.

- Winner of the ImageNet LSVRC-2015 challenge.
  - Accuracy of 95.51% in classification tasks.
  - Outperforming humans!
  - Several versions depending on the number of convolutional layers: ResNet-18, ResNet-34, ResNet-50, ResNet-152.

- ~60.3M parameters (ResNet-152).

- Back then, making a CNN deeper usually decreased its classification accuracy (vanishing gradients).



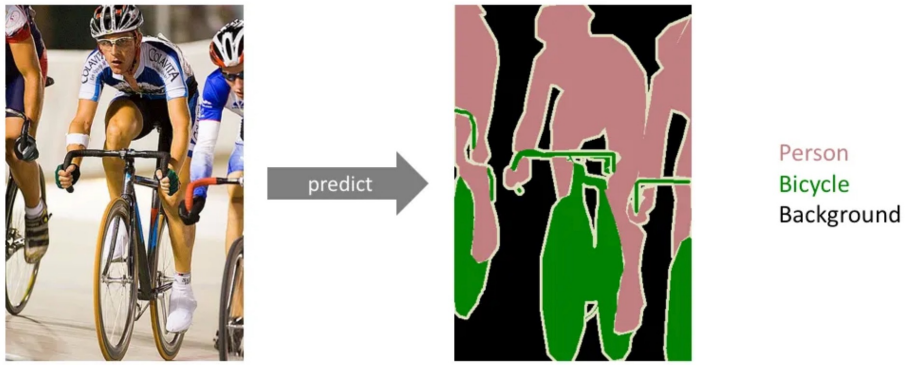
[Source: He, K. et al., 2015]

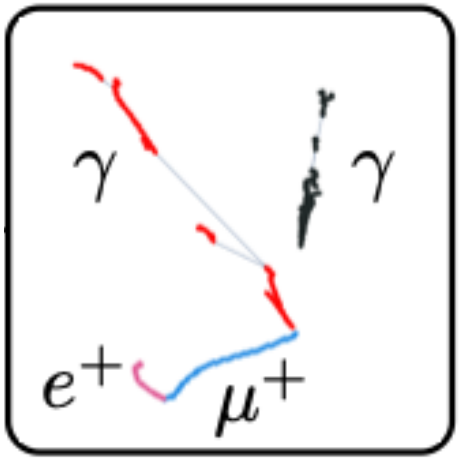
- The "residual blocks" allow the network to apply the identify function (f(x) = x) when needed (zeroing out the weights of the intermediate layers, avoiding vanishing gradients).

- Widely used in particle physics experiments: DUNE, CMS, MicroBooNE… (again, see L. Whitehead lecture later!).

Computer vision tasks

- Image classification – identifying the class an object belongs to – is a core task in computer vision and is the primary goal of the architectures shown so far.
- However, variations of the presented CNN architectures can approach different kinds of problems, such as face recognition or object detection.
- Current architectures tend to be formed from building blocks of successful models (e.g., inception block from GoogleNet, skip connections from ResNets).
- Semantic segmentation tries to identify objects at pixel level belonging to the same class.
  - It is key for pattern recognition in fundamental physics.

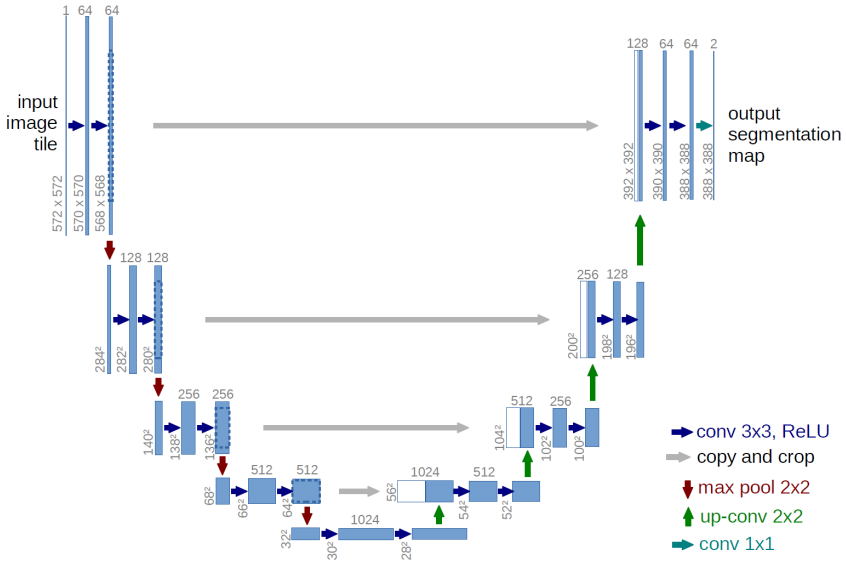




[Source: F. Drielsma et al. 2021]

U-Net

- A successful neural network for semantic segmentation.
- "U" shape (symmetric):
  - Encoder: learns to extract relevant information from the image by applying convolutions.
  - Decoder: learns to recover the location of the target information by upsampling (usually through transposed convolutions).
  - Skip connections between encoder and decoder layers to improve the output locations.
- The output is a high-resolution image rather than a single value.
- Applications in autonomous driving, medical diagnosis, particle physics, etc.



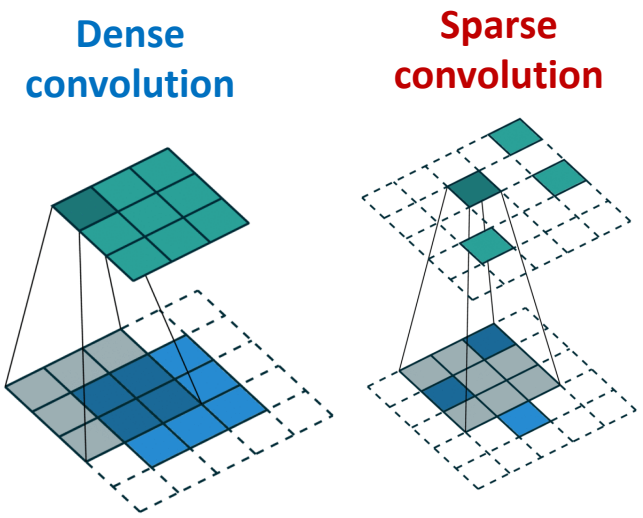
[Source: Ronneberger, O. et al., 2015]

Overview

1. Introduction to computer vision.
2. Convolutional neural networks (CNNs).
3. **Beyond CNNs: exploring other current methods.**
4. Generative models.
5. Challenges and future directions.
6. Conclusion.

Handling sparse data

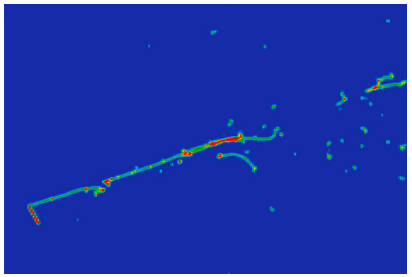
- In particle physics and astrophysics, data is often sparse due to the nature of the objects being studied or the particles detected.
- This poses a challenge for computer vision, as standard CNNs are designed to work with dense data. To address this, researchers are developing new algorithms and techniques specifically tailored to sparse data.
  - For example, one approach is to use Submanifold Sparse Convolutional Networks (SSCN), where the convolution operation is performed only on the non-zero elements of the sparse data, resulting in an efficient and accurate representation of the data.



"Dense" image

- All pixels might be helpful for the classification.
- Ideal for standard CNNs.

"Sparse" images

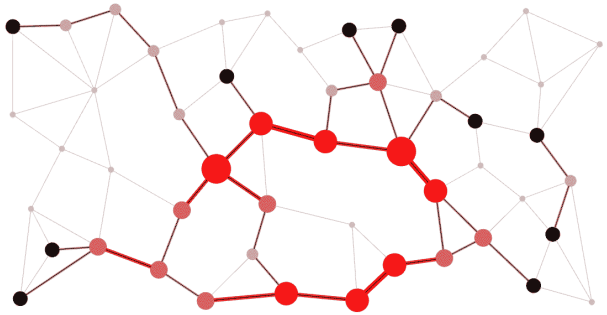


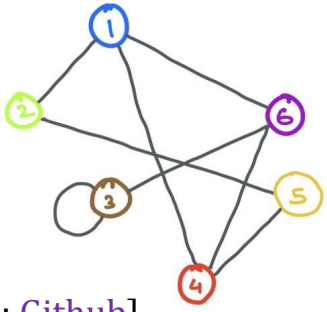
- Most pixels are background.
- A standard CNN would perform loads of useless computations.

[Source: britannica.com]

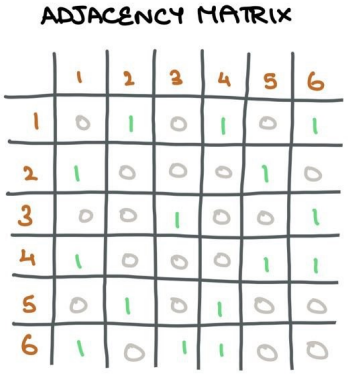
Graph neural networks (GNNs)

- Graph Neural Networks (GNNs) are a type of deep learning model that can learn and process information from the complex structure of graphs, which makes them suitable for tasks such as node classification, link prediction, or graph classification.





[Source: Github]



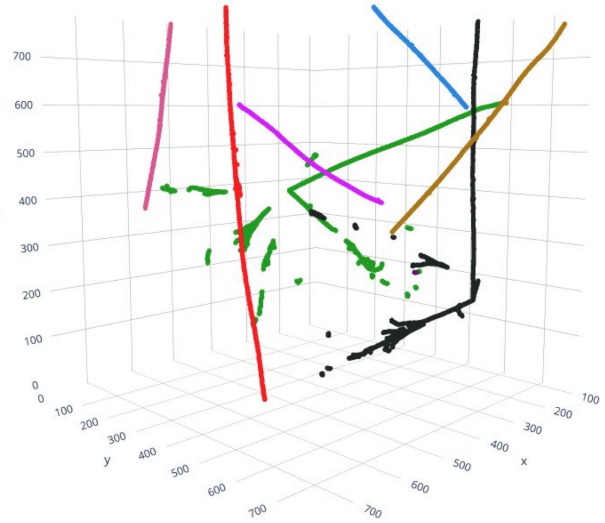
- Compared to CNNs, GNNs can handle graph data with variable size and structure, making them more suitable for relational data applications.

- See J. Duarte and F. Drielsma's lectures tomorrow!

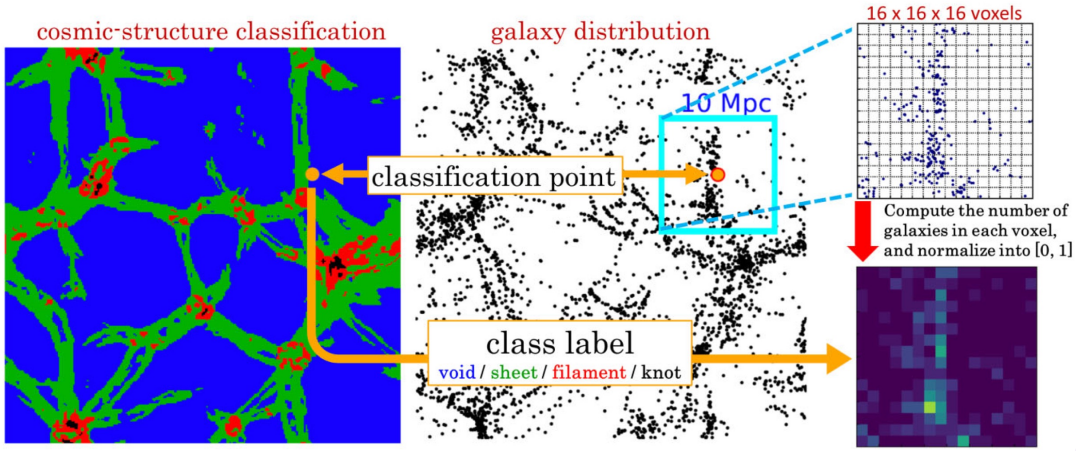
- Some applications of GNNs include social network analysis, recommendation systems, or bioinformatics. GNNs can also be used to model and reason about physical and biological systems, such as predicting the behavior of proteins or designing new molecules.

SSCNs and GNNs: applications in physics

- In physics, SSCNs and GNNs have been used for a variety of applications in physics, including:
  - Anomaly detection.
  - Signal vs. background discrimination.
  - Galaxy identification and classification.
  - Neutrino interaction classification.
  - Pileup mitigation.
  - Event energy reconstruction.
  - Track vs. shower separation.
  - Particle tracking.
  - Etc.



[Source: K. Terao, 2020]



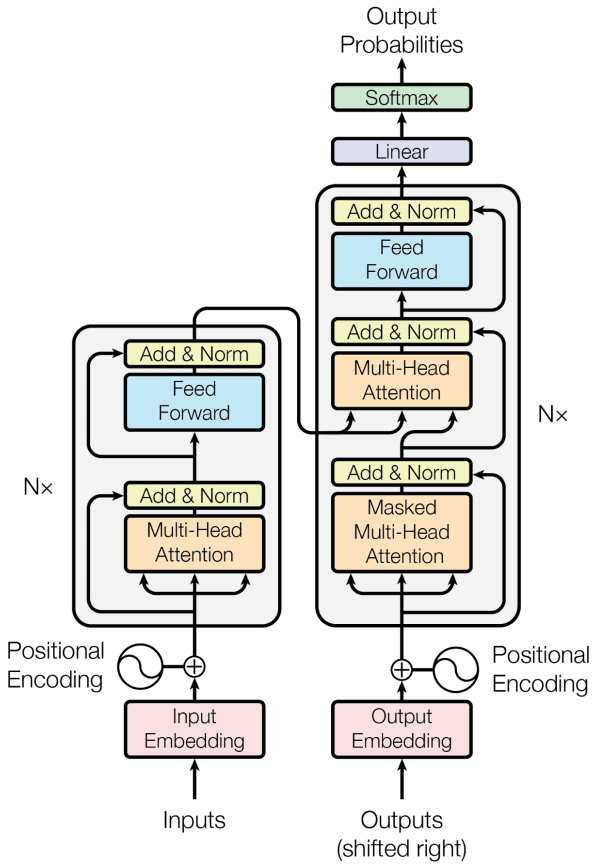
[Source: S. Inoue et al., 2022]

Transformers

- Transformers are a type of deep neural network architecture that has revolutionized natural language processing (NLP) and other sequence modeling tasks.

- They were first introduced in the 2017 paper "Attention is All You Need" by Vaswani et al. (arXiv:1706.03762) and have since become one of the most popular deep learning models.

- Transformers have been successfully applied to a wide range of NLP tasks, including machine translation, text summarization, sentiment analysis, and named entity recognition.
  - ChatGPT is a Transformer.



Transformers: self-attention

- Self-attention is a mechanism that allows each token in the input sequence to attend to all other tokens and learn context-specific representations.

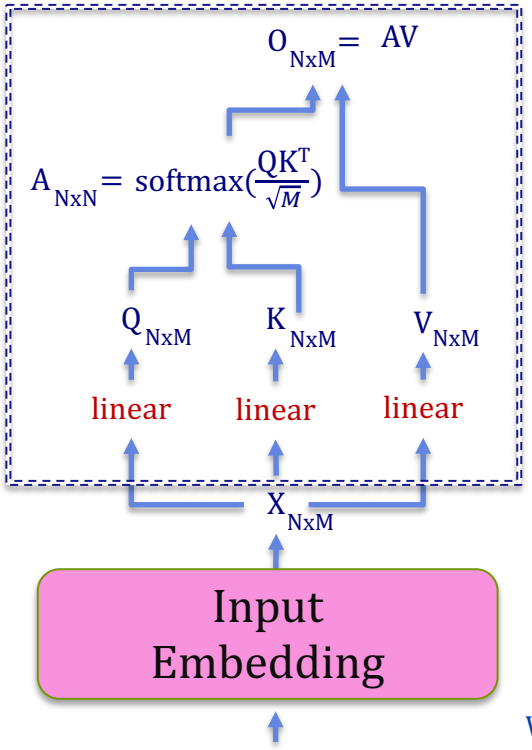
	I	am	a	student
I	0.4	0.1	0.2	0.3
am	0.3	0.6	0.0	0.1
a	0.2	0.1	0.5	0.2
student	0.4	0.1	0.1	0.4

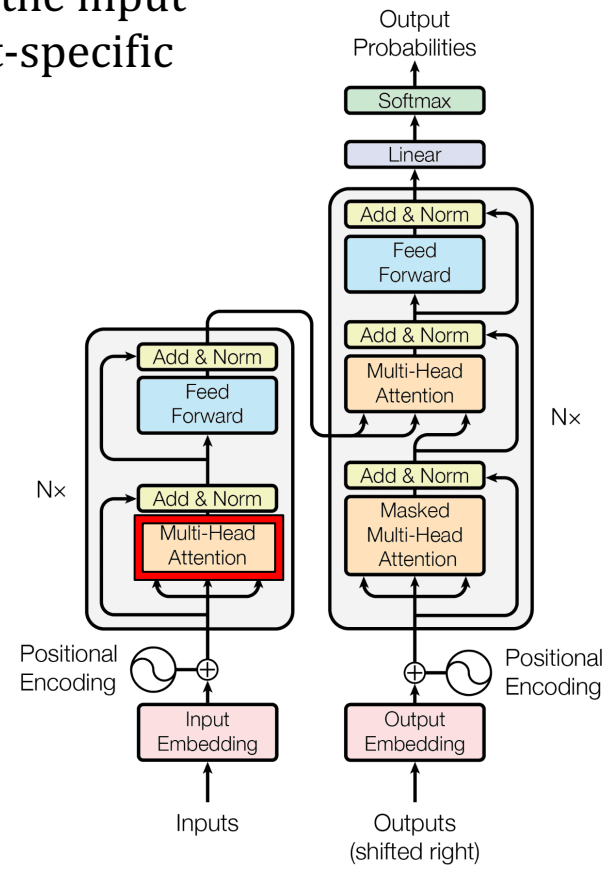
A_{NxN}

- The self-attention mechanism computes a weighted sum of the input embeddings, where the weights are learned based on the similarity between the tokens.
- Unlike memory mechanisms in RNNs, self-attention enables the Transformer model to capture long-range dependencies and handle variable-length input sequences.

I	0.59	0.23	0.02
am	0.11	0.28	0.11
a	0.40	0.15	0.99
student	0.12	0.35	0.61

X_{NxM}





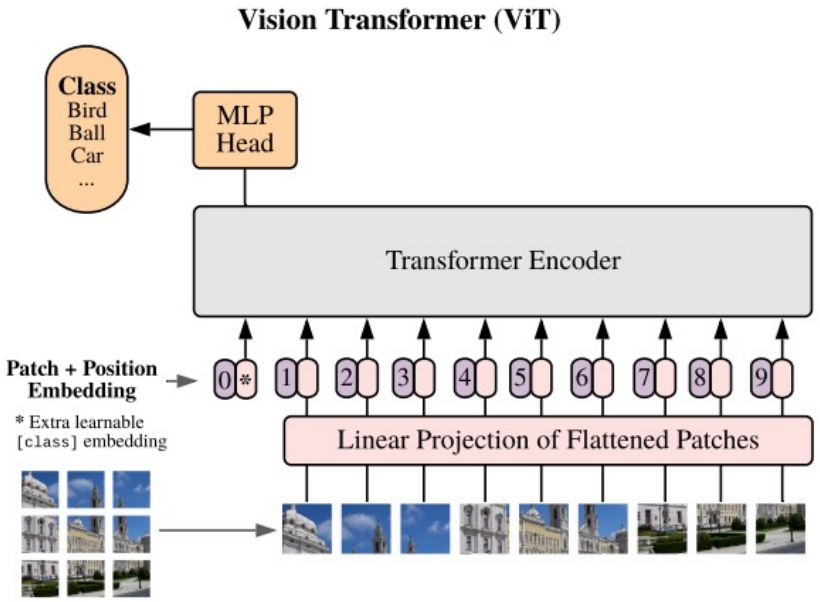
While self-attention is a key component of the Transformer architecture, it is important to note that Transformers use **multi-head attention**, which allows the model to attend to information from different representation subspaces.

Vision Transformer (ViT)

- Dosovitskiy, A. et al., 2020.

- Although Transformers were initially developed for natural language processing (NLP) tasks, they have found applications in a wide range of domains beyond NLP as well.
  - Transformers have been applied to computer vision too.
  - Vision Transformers (ViT) is one such example that can achieve state-of-the-art results on several benchmark datasets.
  - The input sequence consists of squared "patches" of the image.

- In physics, Transformers have been used for a variety of applications, including:
  - Particle decay prediction.
  - Particle track fitting.
  - Vertex finding.
  - Jet identification.
  - Etc.



[Source: Dosovitskiy, A. et al., 2020]

Choosing the right architecture

- When choosing a neural network architecture, consider the following factors:
  - Data type and task complexity: different architectures are designed to handle different types of data and tasks. For example, SSCNs are best for sparse data classification, while U-Nets are best for semantic segmentation.
  - Amount of training data: some architectures require large amounts of data to train effectively, while others can achieve good results with smaller amounts of data.
  - Network capacity and computing resources: having more model parameters can potentially improve a model's performance, allowing the model to learn more complex representations of the data. However:
    - Larger models require more computational resources to train and inference, which can be a practical limitation in some applications.
    - As the number of parameters increases, so does the risk of overfitting the training data, which can lead to poor performance on new, unseen data.
    - Optimisation algorithms can also struggle with larger models due to increased computation time and the possibility of getting stuck in local minima.

- **Overall, the best architecture for a neural network depends on various factors and requires experimentation and iteration to find the optimal solution.**

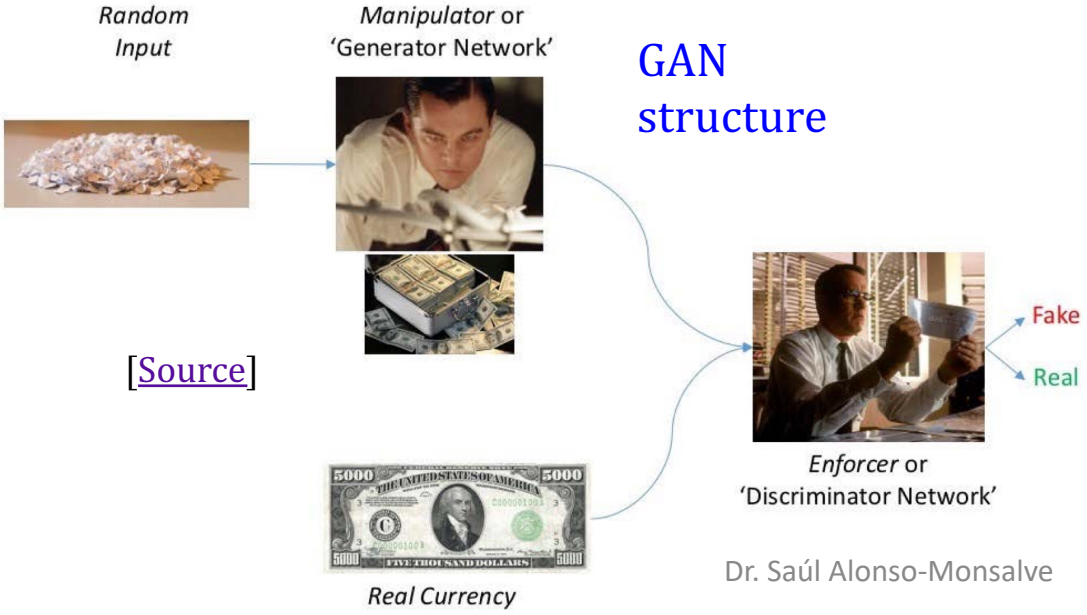
Overview

1. Introduction to computer vision.
2. Convolutional neural networks (CNNs).
3. Beyond CNNs: exploring other current methods.
4. **Generative models.**
5. Challenges and future directions.
6. Conclusion.

Generative models

- Generative models can create new data samples that resemble the input data distribution.
  - See G. Louppe, F. Lanusse, D. Shih, and T. Wongjirad's lectures next week!

- Two main types of generative models are Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs).
  - GANs consist of a generator and discriminator networks that are trained together to generate realistic samples.
  - VAEs encode input data into a latent space and generate new samples by sampling from this latent space and decoding the samples back into the original input space.



[Source]

[Source]

Generative models

- Particle Flows and Stable Diffusion are two newer types of generative models that have shown promising results.
  - Particle Flows transform an initial distribution of particles to a target distribution through a series of continuous transformations.
  - Stable Diffusion uses a multi-step diffusion process with controlled noise levels, allowing the algorithm to produce high-quality and diverse images.

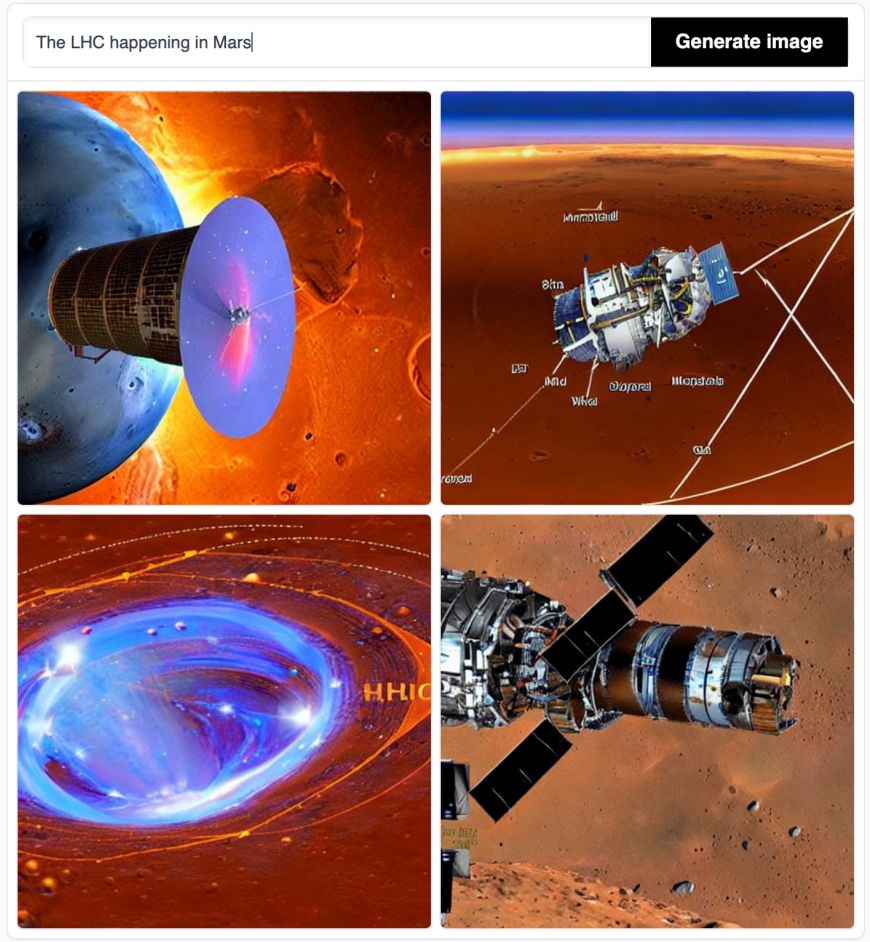
- Generative models have applications in various areas, such as data augmentation, super-resolution, or style transfer.

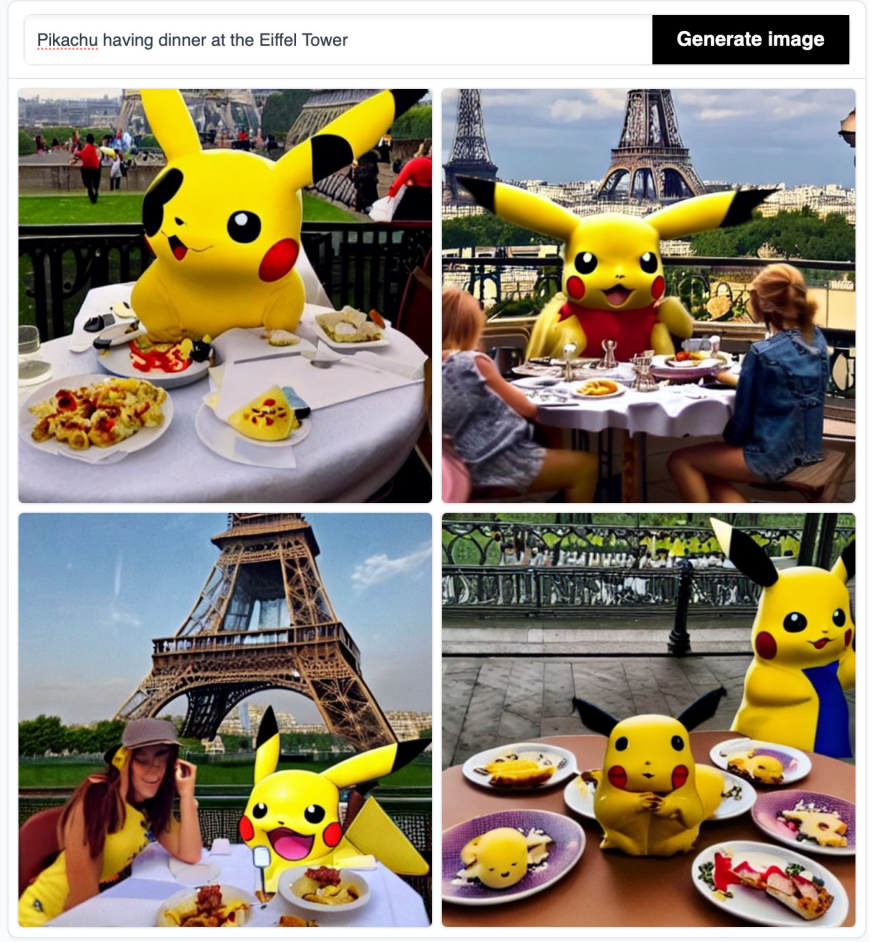
- In particle physics, generative models can be used to simulate particle interactions and generate new data samples for analysis.
  - Generative models are, in general, much faster than Montecarlo simulations.

- In astrophysics, generative models can be used to generate simulations of the universe and the distribution of dark matter.

[Source: Example of Stable Diffusion]

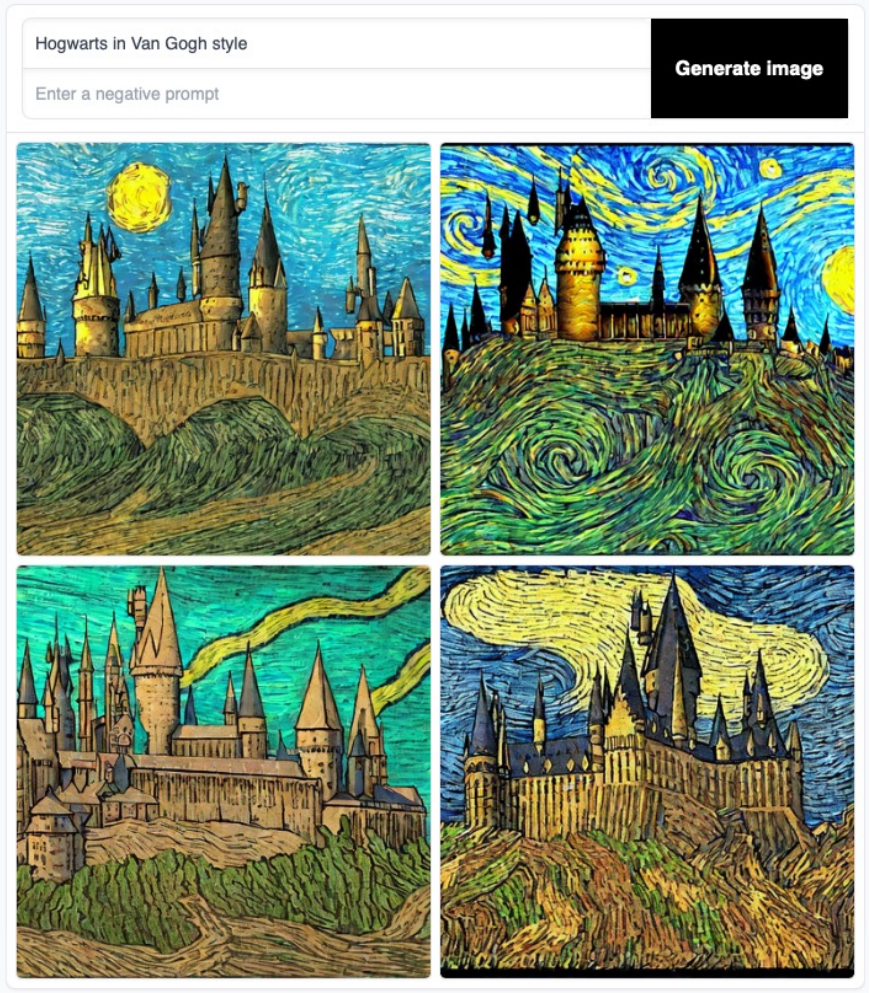
Examples of Stable Diffusion





Credit: https://stablediffusionweb.com

Examples of Stable Diffusion



Credit: https://stablediffusionweb.com

Overview

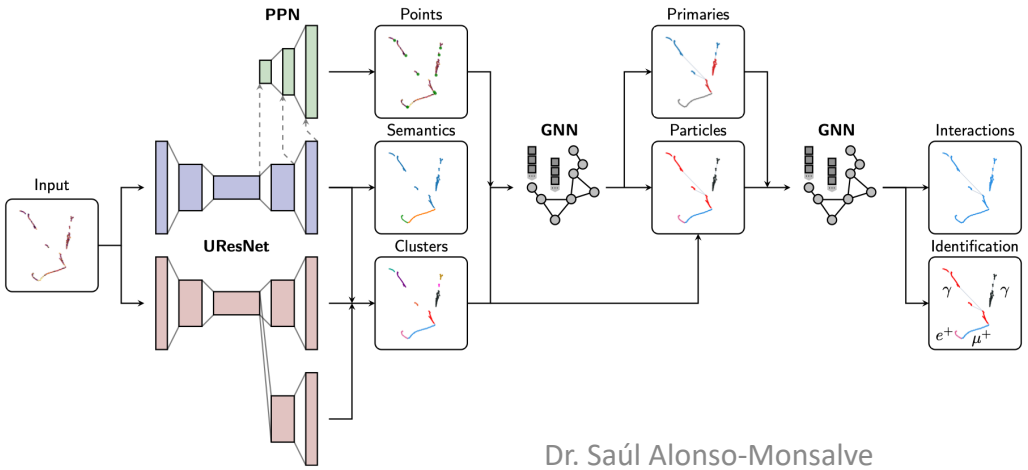
1. Introduction to computer vision.
2. Convolutional neural networks (CNNs).
3. Beyond CNNs: exploring other current methods.
4. Generative models.
5. **Challenges and future directions in fundamental physics.**
6. Conclusion.

Automated physics analyses

- Deep learning and computer vision can be used to automate certain aspects of physics analyses, such as data preprocessing, event selection, reconstruction, etc.

- This saves significant time and resources and can also help ensure that analyses are reproducible and consistent.
  - For example, computer vision can automatically detect and remove background events in particle physics experiments or identify and classify different types of galaxies in astrophysics.
  - It can also help reduce human bias in the analysis process.

- There are many remarkable advances in this regard.
  - Despite promising advances in this area, integrating deep-learning-based computer-vision techniques into the analysis flow of physics experiments can be challenging due to technical, logistical, and sometimes skeptical barriers.



"Scalable, End-to-End, Deep-Learning-Based Data Reconstruction Chain for Particle Imaging Detectors" - F. Drielsma et al. 2021

Robustness against systematic uncertainties and simulation mismodelings

- In particle physics and astrophysics, there are often systematic uncertainties related to the measurements, as well as mismodelings in the simulations.
  - These uncertainties can arise from a variety of sources and can affect the accuracy and precision of the measurements and simulations in these fields.

- Deep learning models can be biased or inaccurate as a result.
  - To address this, researchers are developing methods to make machine learning models more robust against these uncertainties and mismodelings.

- One approach is to use adversarial training, where the model is trained to be robust against adversarial examples that are specifically designed to trick the model.
  - Another approach is to incorporate physics-based constraints or priors into the model (e.g., penalty terms in the loss function), to help ensure that the model is consistent with known physics.
  - Adversarial trainings can also be used with detector data to refine the ML models in an unsupervised way.

Generative models to replace simulations

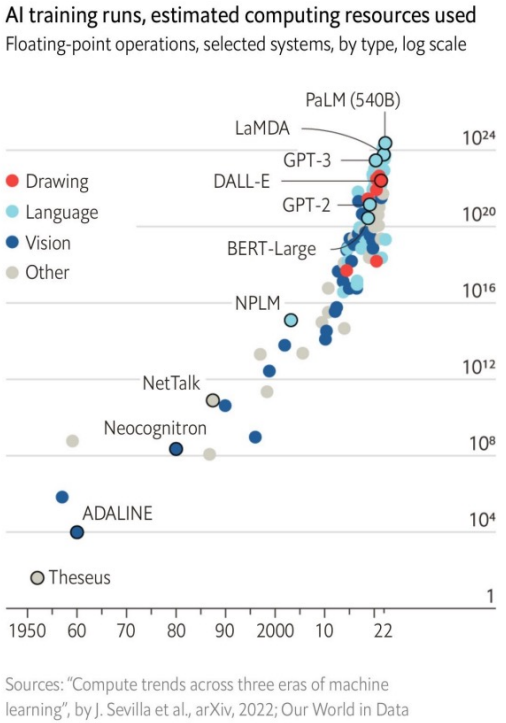
- Generative models are machine learning models that can generate new data that is similar to the training data.

- In particle physics and astrophysics, generative models can be used to generate new simulated data, which can be used to supplement or eventually replace existing simulations.
  - This can save significant time and resources and can also help address uncertainties and mismodelings in the simulations.
  - Current work cannot fully replace current simulations yet, but are more suited for fast prototyping.

- Despite the limitations, generative models are a promising area of research in HEP, and have the potential to revolutionize the way simulations are performed in the field.
  - Although Stable Diffusion shows promise for replacing simulations in HEP experiments, its current computational cost remains challenging.

Large models and infrastructure

- Particle physics and astrophysics generate vast amounts of data, and machine learning models trained on this data can be very large and complex.
  - This requires significant computational resources and infrastructure to train and deploy these models.
  - Investing in large-scale infrastructure and end-to-end systems for machine learning in particle physics and astrophysics is an important future direction.
- We are very far away from state-of-the-art applications:
  - A typical deep learning model in physics usually has never more than a few million parameters.
  - GPT-3.5 (the model behind ChatGPT) was trained for ~12-18 months on a supercomputer with ~10,000 GPUs and ~285,000 CPU cores (~1 billion dollars to rent) and has 175 billion parameters [Source].
- Beware of the significant environmental impact caused by the large carbon footprint of deep learning models.



Overview

1. Introduction to computer vision.
2. Convolutional neural networks (CNNs).
3. Beyond CNNs: exploring other current methods.
4. Generative models.
5. Challenges and future directions.
6. **Conclusion.**

Summary and conclusion

- Computer vision is an essential field in fundamental physics research.
- We have explored the state-of-the-art of deep learning for computer vision.
  - This includes CNNs, GNNs, Transformers…
  - Direct application in fundamental physics.
- Challenges for future research include dealing with sparse data, addressing uncertainties, and creating generative models.
- Developing large models, infrastructure, and real-time models is also crucial for future research.
- Overall, computer vision has opened up new avenues for fundamental physics research, and addressing its challenges can lead to a deeper understanding of the universe.

Recommended links

- "*Machine learning at the energy and intensity frontiers of particle physics*", A. Radovic et al., Nature (2018): https://doi.org/10.1038/s41586-018-0361-2.

- "*A Living Review of Machine Learning for Particle and Nuclear Physics*" (2021): https://iml-wg.github.io/HEPML-LivingReview/review/hepml-review.pdf.

- "*Physics-based Deep Learning Book*", N. Thuerey et al. (2021): https://physicsbaseddeeplearning.org.

- Computer vision tool for anyone to use! : https://landing.ai.

COMPUTER VISION

Dr. Saúl Alonso-Monsalve

saul.alonso.monsalve@cern.ch

ETH Zurich

51st SLAC Summer Institute

August 9, 2023