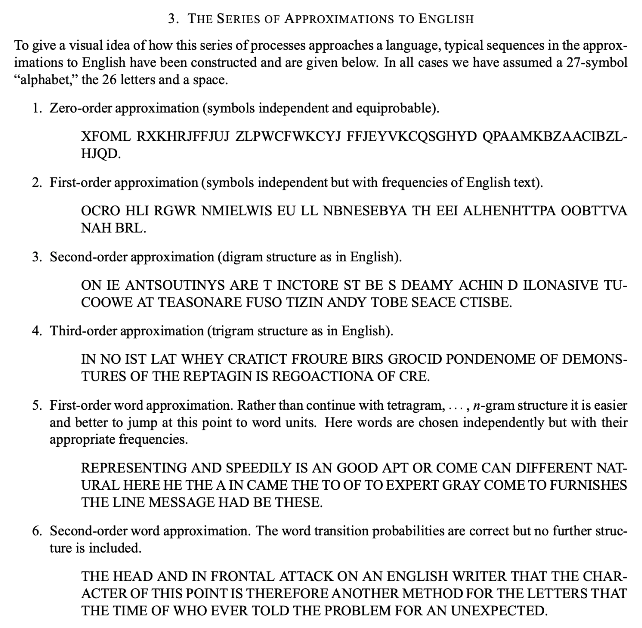
Generative Methods

Webinaire Inteligence Artificelle Générative Lyon 3

Aurélien Garivier

January 25th, 2024

Language Models date back to Shannon (1948)!



3. THE SERIES OF APPROXIMATIONS TO ENGLISH

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27-symbol "alphabet," the 26 letters and a space.

1. Zero-order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGHYD QPAAMKBZAACIBZLHJQD.

2. First-order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA OOBTTVA NAH BRL.

3. Second-order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE SEACE CTISBE.

4. Third-order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE.

5. First-order word approximation. Rather than continue with tetragram, . . . , *n*-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate frequencies.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

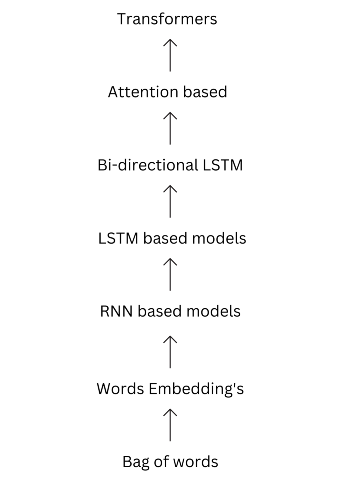
6. Second-order word approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.

Language models have constantly evolved

Shannon: "It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source."

- n-gram language models
- Variable-length Markov models
- grammar-based methods



Src: Wikipedia

Stochastic Parrots

"On the Dangers of **Stochastic Parrots**: Can Language Models Be Too Big?" by Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell (2021)

= entity "for haphazardly stitching together sequences of linguistic forms . . . according to probabilistic information about how they combine, but without any reference to meaning."

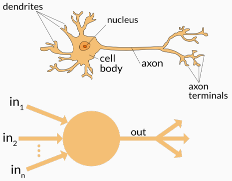
- environmental and financial costs
- inscrutability leading to unknown dangerous biases
- inability of the models to understand the concepts underlying what they learn
- the potential for using them to deceive people.

Feedforward Neural Networks: Mimicking Brains?

Neuron: $x \mapsto \sigma\big(\langle w, x\rangle + b\big)$ with

- parameter $w \in \mathbb{R}^p, b \in R$
- (non-linear) activation function $\sigma : \mathbb{R} \to \mathbb{R}$

  typically $\sigma(x) = \frac{1}{1+\exp(-x)}$ or $\sigma(x) = \max(x, 0)$ called ReLU



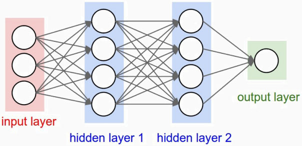
Layer: $x \mapsto \boldsymbol{\sigma}\big(Mx + \mathbf{b}\big)$ with

- parameter $M \in M_{q,p}(\mathbb{R}), \mathbf{b} \in \mathbb{R}^q$
- component-wise activation function $\boldsymbol{\sigma} = \sigma^{\otimes q}$

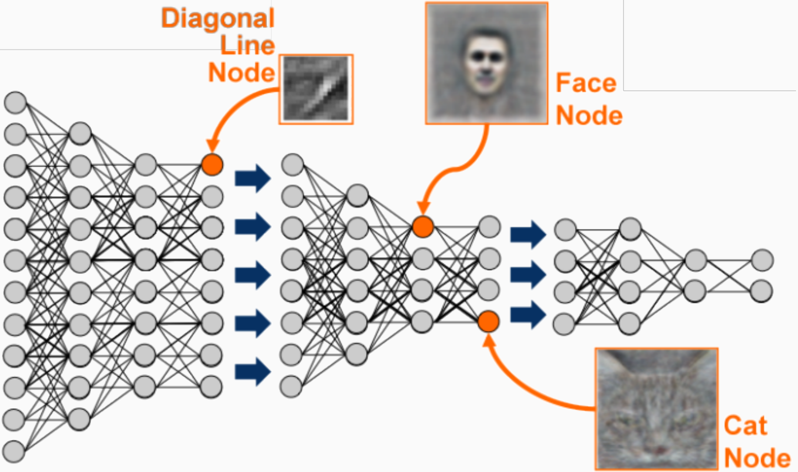
Network: composition of layers $f_\theta = \boldsymbol{\sigma}_D \circ T_D \circ \cdots \circ \boldsymbol{\sigma}_1 \circ T_1$ with

- architecture $A = \big(D, (p_1, \ldots, p_{D-1})\big)$
- $x_0 = x, \quad x_d = \boldsymbol{\sigma}_d\big(T_d x_{d-1}\big) \in \mathbb{R}^{p_d}$
- $T_d x = M_d x + \mathbf{b}_d$
- parameter $\theta = (M_1, \mathbf{b}_1, \ldots, \ldots, M_D, \mathbf{b}_D)$

  $\theta \in \Theta_A = \prod_{d=1}^{D} \mathcal{M}_{p_{d-1},p_d}(\mathbb{R}) \times \mathbb{R}^{p_d}$
- *depth* D (⚠st. nb layers), *width* $\max_{1 \leq d \leq D} p_d$



Why deep learning, then? The dream

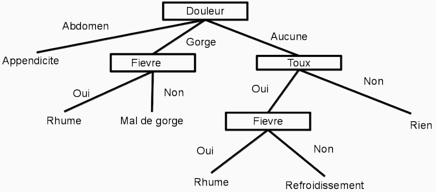


Explainability vs Interpretability

Two distinct notions (but the vocabulary is misleading: we flllow here https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf).

A decision rule is said to be:

interpretable if we understand how a prediction is associated to an observation; typical example: decision tree



http://www.up2.fr/

explainable if we understand what feature values led to the prediction, possibly by a counterfactual analysis; for example: "if variable X_3 had taken that other value, then the prediction would have been different".

Explainability vs Interpretability

Two distinct notions (but the vocabulary is misleading: we flllow here https://www2.eecs.berkeley.edu/Pubs/TechRpts/2017/EECS-2017-159.pdf).

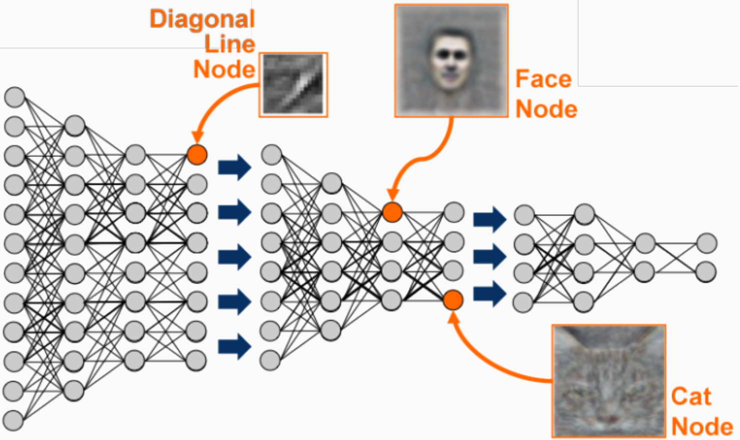
A decision rule is said to be:

interpretable if we understand how a prediction is associated to an observation; typical example: decision tree

explainable if we understand what feature values led to the prediction, possibly by a counterfactual analysis; for example: "if variable X_3 had taken that other value, then the prediction would have been different".

Expainability relates to the statistical notions of *causal inference* and *sensibility analysis*

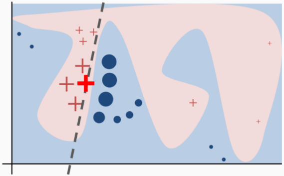
Interpreting a deep Neural Work : the Founding Dream



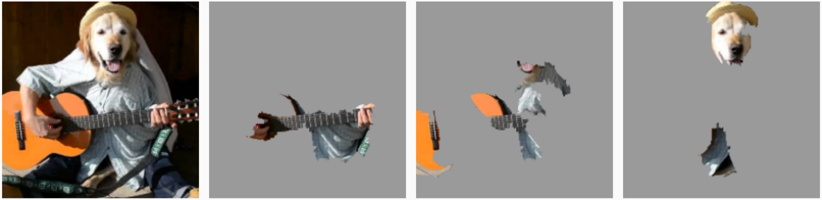
http://aiehive.com

An audacious scientific bet...

Local Interpretable Model-Agnostic Explanations: LIME



Linear model with feature selection on local subset of data



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Src: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin.

Local Interpretable Model-Agnostic Explanations: LIME

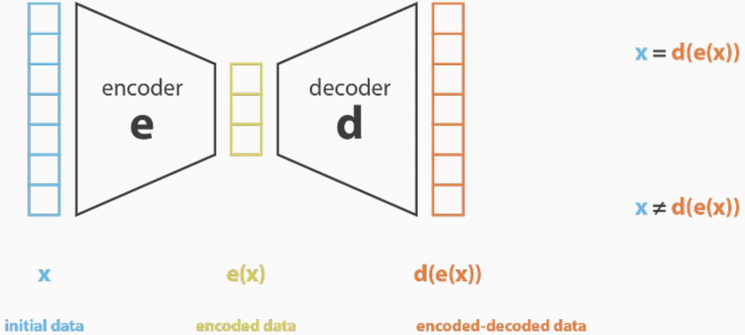
(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

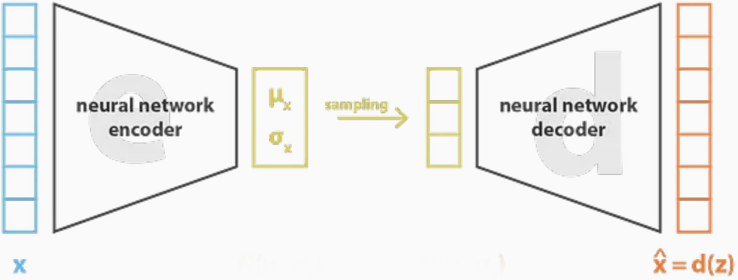
Src: "Why Should I Trust You?" Explaining the Predictions of Any Classifier, by Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin.

Dimensionality reduction

- Data: $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathcal{M}_{n,p}(\mathbb{R}),\ p \gg 1$
- Dimensionality reduction: replace x_i with $y_i = \mathrm{enc}(x_i)$, where $\mathrm{enc} : \mathbb{R}^p \to \mathbb{R}_d,\ d \ll p$
- Hopefully, we do not loose too much by replacing x_i by y_i: there exists a recovering mapping $\mathrm{dec} : \mathbb{R}^d \to \mathbb{R}^p$ such that for all $i \in \{1, \ldots, n\}$, $\mathrm{dec}(\mathrm{enc}(x_i)) \approx x_i$

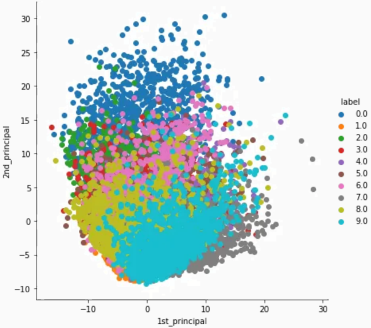


Auto-encoders

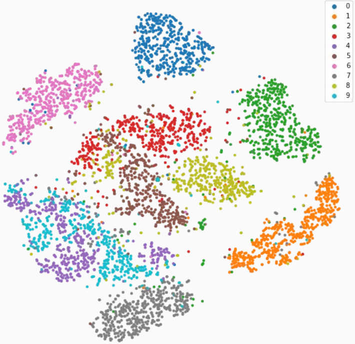


Src: https://towardsdatascience.com/

Auto-encoders



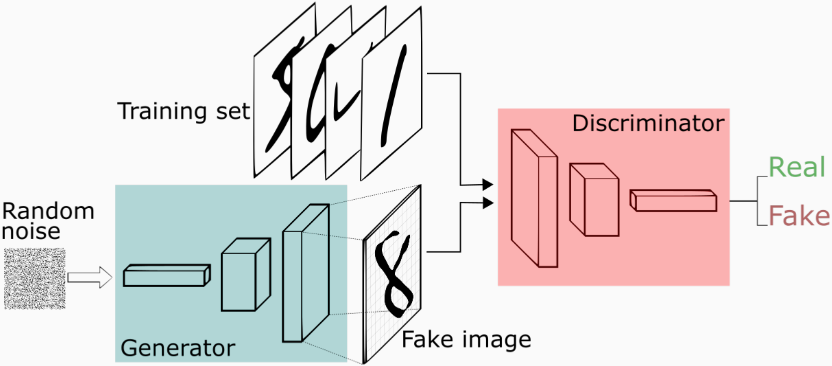
PCA



Auto-encoder

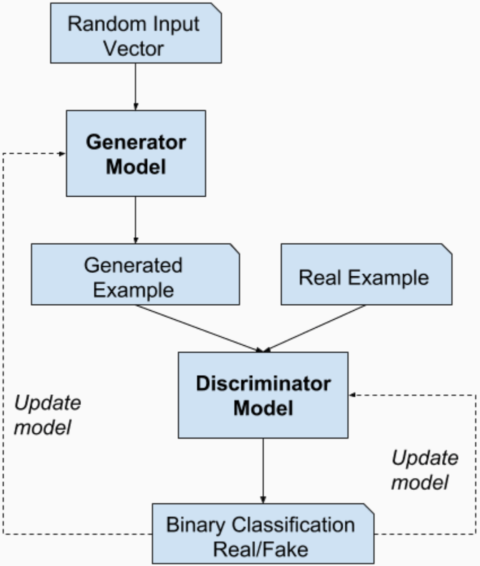
Src: `https://medium.com/`

Generative Adversarial Networks



Src: `https://sthalles.github.io/`

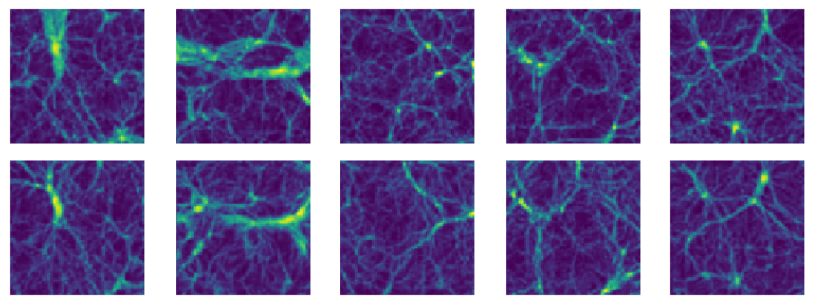
Generator / Discriminator



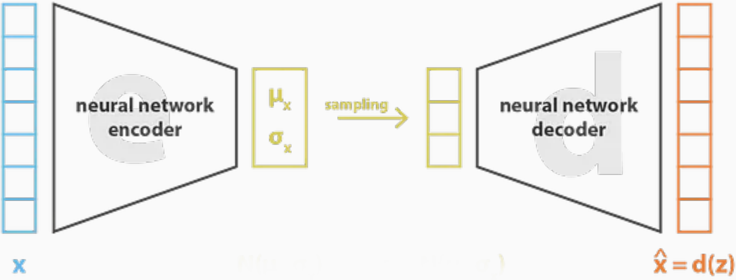
Src: https://machinelearningmastery.com

Example

[Encoding large scale cosmological structure with Generative Adversarial Networks, *Marion Ullmo, Aurélien Decelle and Nabila Aghanim*, Astronomy & Astrophysics]

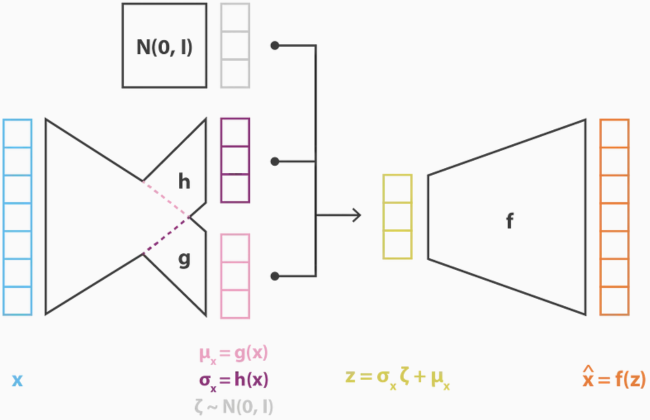


Using auto-encoders for data generation?



Src: `https://towardsdatascience.com/`

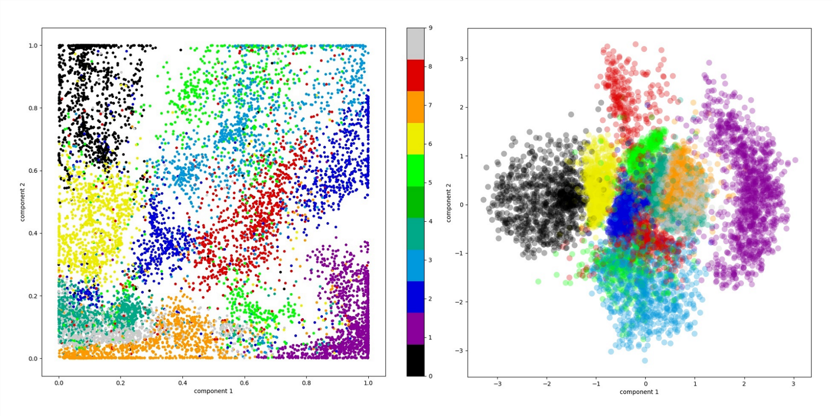
Variational Auto-Encoders



$$\text{loss} = C\,\|x - \hat{x}\|^2 + KL[N(\mu_x, \sigma_x), N(0, I)] = C\,\|x - f(z)\|^2 + KL[N(g(x), h(x)), N(0, I)]$$

Src: `https://towardsdatascience.com/`

Variational Auto-Encoders: example



AE **variational AE**

Src: `https://pureai.com/`

Example

[Geophysical Inversion Using a Variational Autoencoder to Model an Assembled Spatial Prior Uncertainty, *Jorge Lopez-Alvis, Frederic Nguyen, M. C. Looms, Thomas Hermans*, Journal of Geophysical Research: Solid Earth]