From Higgless models to Transformers:

Breaking the cut-offs in Physics and Machine Learning

Riccardo^3 Torre

INFN - Sezione di Genova

SNS - 09 November 2024

About some future directions in fundamental physics

a speculative talk in honour of Riccardo Barbieri

Objective of this talk

Objective of this talk

- Build an analogy between the language of Fundamental Physics and the language of Machine Learning

Objective of this talk

- Build an analogy between the language of Fundamental Physics and the language of Machine Learning
- Explain (this you all know) how the Higgs discovery impacted our understanding of Particle Physics

Objective of this talk

- Build an analogy between the language of Fundamental Physics and the language of Machine Learning
- Explain (this you all know) how the Higgs discovery impacted our understanding of Particle Physics
- Explain (this you may not all know) how the discovery of "self-attention" revolutionized the field of Machine Learning

Objective of this talk

- Build an analogy between the language of Fundamental Physics and the language of Machine Learning
- Explain (this you all know) how the Higgs discovery impacted our understanding of Particle Physics
- Explain (this you may not all know) how the discovery of "self-attention" revolutionized the field of Machine Learning
- Explain why the latter is relevant for us

Objective of this talk

- Build an analogy between the language of Fundamental Physics and the language of Machine Learning
- Explain (this you all know) how the Higgs discovery impacted our understanding of Particle Physics
- Explain (this you may not all know) how the discovery of "self-attention" revolutionized the field of Machine Learning
- Explain why the latter is relevant for us
- Point out some challenges in ML that we need to face to use it in science

Objective of this talk

- Build an analogy between the language of Fundamental Physics and the language of Machine Learning
- Explain (this you all know) how the Higgs discovery impacted our understanding of Particle Physics
- Explain (this you may not all know) how the discovery of "self-attention" revolutionized the field of Machine Learning
- Explain why the latter is relevant for us
- Point out some challenges in ML that we need to face to use it in science

*Many of the analogies I will introduce are speculative, even though there exist ongoing research aiming to make them formal

Why do I start from Higgless models?

Why do I start from Higgless models?

Just because my journey in physics research started there, with the mentorship of Riccardo Barbieri

arXiv:0911.1942v3 [hep-ph] 31 Mar 2010

Composite Vectors at the Large Hadron Collider

R. Barbieri[a,b], A.E. Cárcamo Hernández[a,b], G. Corcella[a,b,c],
R. Torre[b,d] and E. Trincherini[a]

[a] *Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa, Italy*

[b] *INFN, Sezione di Pisa, Largo Fibonacci 3, I-56127 Pisa, Italy*

[c] *Museo Storico della Fisica e Centro Studi e Ricerche E. Fermi
Piazza del Viminale 1, I-00184 Roma, Italy*

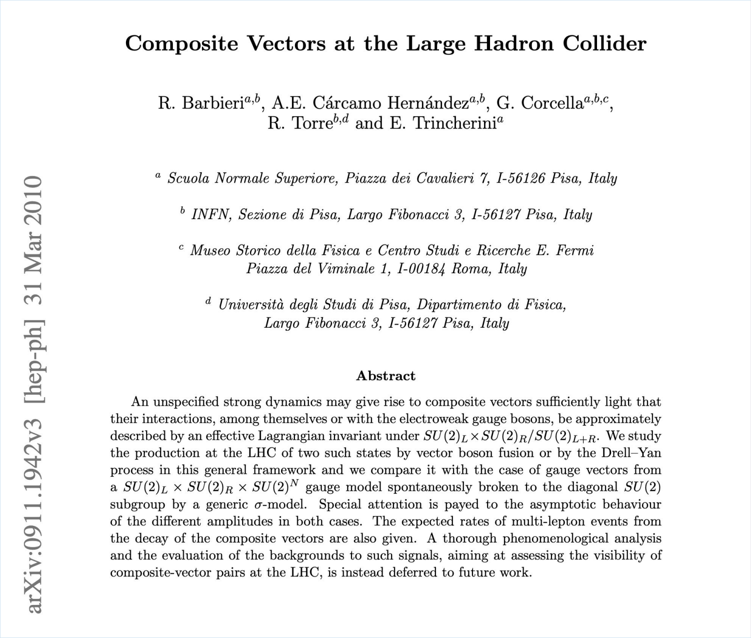
[d] *Università degli Studi di Pisa, Dipartimento di Fisica,
Largo Fibonacci 3, I-56127 Pisa, Italy*

Abstract

An unspecified strong dynamics may give rise to composite vectors sufficiently light that their interactions, among themselves or with the electroweak gauge bosons, be approximately described by an effective Lagrangian invariant under $SU(2)_L \times SU(2)_R / SU(2)_{L+R}$. We study the production at the LHC of two such states by vector boson fusion or by the Drell–Yan process in this general framework and we compare it with the case of gauge vectors from a $SU(2)_L \times SU(2)_R \times SU(2)^N$ gauge model spontaneously broken to the diagonal $SU(2)$ subgroup by a generic σ-model. Special attention is payed to the asymptotic behaviour of the different amplitudes in both cases. The expected rates of multi-lepton events from the decay of the composite vectors are also given. A thorough phenomenological analysis and the evaluation of the backgrounds to such signals, aiming at assessing the visibility of composite-vector pairs at the LHC, is instead deferred to future work.

Why do I start from Higgless models?

Just because my journey in physics research started there, with the mentorship of Riccardo Barbieri



arXiv:0911.1942v3 [hep-ph] 31 Mar 2010

Composite Vectors at the Large Hadron Collider

R. Barbieri[a,b], A.E. Cárcamo Hernández[a,b], G. Corcella[a,b,c], R. Torre[b,d] and E. Trincherini[a]

[a] *Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa, Italy*

[b] *INFN, Sezione di Pisa, Largo Fibonacci 3, I-56127 Pisa, Italy*

[c] *Museo Storico della Fisica e Centro Studi e Ricerche E. Fermi Piazza del Viminale 1, I-00184 Roma, Italy*

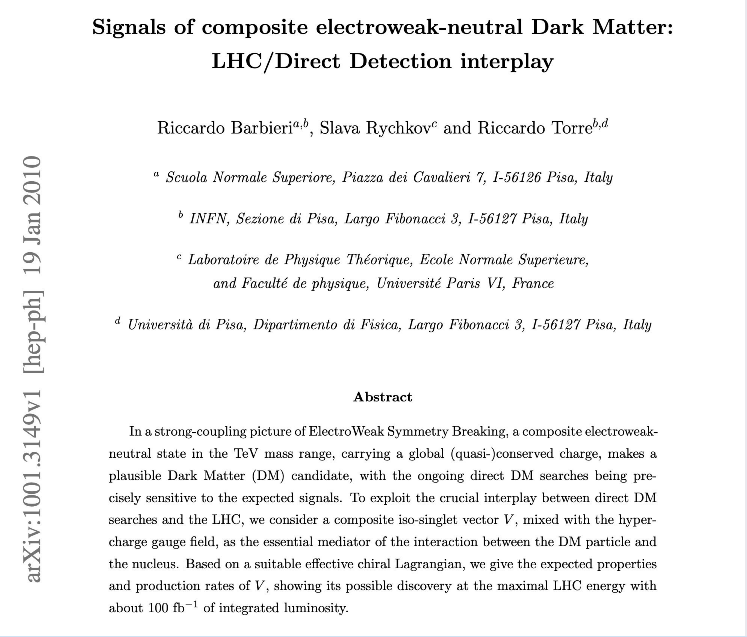
[d] *Università degli Studi di Pisa, Dipartimento di Fisica, Largo Fibonacci 3, I-56127 Pisa, Italy*

Abstract

An unspecified strong dynamics may give rise to composite vectors sufficiently light that their interactions, among themselves or with the electroweak gauge bosons, be approximately described by an effective Lagrangian invariant under $SU(2)_L \times SU(2)_R / SU(2)_{L+R}$. We study the production at the LHC of two such states by vector boson fusion or by the Drell–Yan process in this general framework and we compare it with the case of gauge vectors from a $SU(2)_L \times SU(2)_R \times SU(2)^N$ gauge model spontaneously broken to the diagonal $SU(2)$ subgroup by a generic σ-model. Special attention is payed to the asymptotic behaviour of the different amplitudes in both cases. The expected rates of multi-lepton events from the decay of the composite vectors are also given. A thorough phenomenological analysis and the evaluation of the backgrounds to such signals, aiming at assessing the visibility of composite-vector pairs at the LHC, is instead deferred to future work.

Why do I start from Higgless models?

Just because my journey in physics research started there, with the mentorship of Riccardo Barbieri



arXiv:1001.3149v1 [hep-ph] 19 Jan 2010

Signals of composite electroweak-neutral Dark Matter: LHC/Direct Detection interplay

Riccardo Barbieri[a,b], Slava Rychkov[c] and Riccardo Torre[b,d]

[a] *Scuola Normale Superiore, Piazza dei Cavalieri 7, I-56126 Pisa, Italy*

[b] *INFN, Sezione di Pisa, Largo Fibonacci 3, I-56127 Pisa, Italy*

[c] *Laboratoire de Physique Théorique, Ecole Normale Superieure, and Faculté de physique, Université Paris VI, France*

[d] *Università di Pisa, Dipartimento di Fisica, Largo Fibonacci 3, I-56127 Pisa, Italy*

Abstract

In a strong-coupling picture of ElectroWeak Symmetry Breaking, a composite electroweak-neutral state in the TeV mass range, carrying a global (quasi-)conserved charge, makes a plausible Dark Matter (DM) candidate, with the ongoing direct DM searches being precisely sensitive to the expected signals. To exploit the crucial interplay between direct DM searches and the LHC, we consider a composite iso-singlet vector V, mixed with the hypercharge gauge field, as the essential mediator of the interaction between the DM particle and the nucleus. Based on a suitable effective chiral Lagrangian, we give the expected properties and production rates of V, showing its possible discovery at the maximal LHC energy with about 100 fb^{-1} of integrated luminosity.

A driving principle

<u>scientific revolutions are more often driven by new tools than by new concepts.</u>

Freeman J. Dyson, Birds and frogs, Selected Papers 1990-2014

A driving principle

The second theme that **George Green's work exemplifies is the** **historical fact that**

<u>scientific revolutions are more often driven by new tools than by new concepts</u>.

Thomas Kuhn in his famous **book, "The Structure of Scientific Revolutions",** talked almost exclusively about concepts and hardly at all about tools. His idea of a scientific revolution is based on a single example, the revolution in theoretical physics that occurred in the 1920s with the advent of quantum mechanics. This was a prime example of a concept-driven revolution. Kuhn's book **was so brilliantly** written **that** it became an instant classic. It **misled a whole generation of students and historians of science into believing that all scientific revolutions are concept driven.** The concept-driven revolutions are the ones that attract the most attention and have the greatest impact on public awareness of science, but in fact they are comparatively rare. **In the last five hundred years we have had six major concept driven revolutions, associated with the names of Copernicus, Newton, Darwin, Maxwell, Einstein and Freud, besides the quantum-mechanical revolution that Kuhn took as his model. During the same period there have been about twenty tool-driven revolutions, not so impressive to the general public but of equal importance to the progress of science.** I will not attempt to make a complete list of tool-driven revolutions. Two prime examples are the Galilean revolution resulting from the use of the telescope in astronomy, and the Crick–Watson revolution resulting from the use of X-ray diffraction to determine the structure of big molecules in biology. **The effect of a concept-driven revolution is to explain old things in new ways. The effect of a tool-driven revolution is to discover new things that have to be explained.** In physics there has been a preponderance of tool-driven revolutions. We have been more successful in discovering new things than in explaining old ones. **George Green's great discovery, the Green's function, is a mathematical tool rather than a physical concept. It did not give the world a new theory of electricity and magnetism or a new picture of physical reality. It gave the world a new bag of mathematical tricks**, useful for exploring the consequences of theories and for predicting the existence of new phenomena that experimenters could search for. The Green's function was **a tool of discovery, like the telescope and the microscope, but aimed at mathematical models and theories instead of being aimed at the sky and the microbe.**

Freeman J. Dyson, Birds and frogs, Selected Papers 1990-2014

Preview

Preview

Fundamental Physics

Machine Learning

Preview

Fundamental Physics

Objective: describe all known particles, their interactions, and at all scales

Machine Learning

Objective: emulate human intelligence to allow machines to perform tasks without explicitly programming them

Preview

Fundamental Physics

Objective: describe all known particles, their interactions, and at all scales

Machine Learning

Objective: describe all features, their interactions (correlations), and at all scales (context)

Preview

Fundamental Physics

Objective: describe all known particles, their interactions, and at all scales

Energy

G_N^{-2} EFT close to the Planck scale (quantum gravity, string theory, bla bla bla)

EFT above the EW scale (SMEFT)

EFT at the EW scale (dim-4 Standard Model)

G_F^{-2}

EFT below EW scale (QED+QCD+Fermi Theory)

$\Lambda_{\rm QCD}$ EFT below QCD scale (QED + Chiral QCD+Fermi Theory)

m_e

EFT below electron mass (Euler-Heisenberg)

Machine Learning

Objective: describe all features, their interactions (correlations), and at all scales (context)

Preview

Fundamental Physics

Objective: describe all known particles, their interactions, and at all scales

Energy

G_N^{-2} EFT close to the Planck scale (quantum gravity, string theory, bla bla bla)

EFT above the EW scale (SMEFT)

G_F^{-2} EFT at the EW scale (dim-4 Standard Model)

EFT below EW scale (QED+QCD+Fermi Theory)

Λ_{QCD} EFT below QCD scale (QED + Chiral QCD+Fermi Theory)

m_e

EFT below electron mass (Euler-Heisenberg)

Machine Learning

Objective: describe all features, their interactions (correlations), and at all scales (context)

Context Complexity / Correlation length

Fully connected graph data / Unlimited range
Transformers, Self-Attention, GATs: Maximum flexibility, allowing global dependencies across all points in the input

Sparse graph data - Extended range
Graph NN (GNN): Variable-length, structured data with complex dependencies and structured relationships

Grids/sequences - Finite range
Convolutional NN (CNNs), Recursive NNs (RNN), etc: structured spatial or sequential data, capturing local patterns but with limited range

Vector data - Moderate Range
Feedforward Neural Networks (FFNNs): Fixed-length vectors, capturing shallow dependencies within small data sets

Scalar Data - Short Range
Associative Memory Models, Hopfield Networks, Perceptrons: minimal complexity, point-wise or isolated associations

The Higgs in particle physics

The Higgs in particle physics

- It is often said that the role of the Higgs is to "give masses" to fundamental particles

The Higgs in particle physics

- It is often said that the role of the Higgs is to "give masses" to fundamental particles
- I believe that, in a more modern view, one can say that the role of the Higgs is to extend the validity of the Standard Model "far above" the weak scale

The Higgs in particle physics

- It is often said that the role of the Higgs is to "give masses" to fundamental particles
- I believe that, in a more modern view, one can say that the role of the Higgs is to extend the validity of the Standard Model "far above" the weak scale

SM + non-linear EWSB

$$\mathscr{L}_{\mathrm{EWSB}} = \frac{v^2}{4}\langle D_\mu \Sigma (D^\mu \Sigma)^\dagger \rangle$$

$$\Sigma = e^{i\pi^a \sigma^a / v}$$

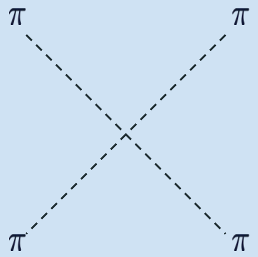
The Higgs in particle physics

- It is often said that the role of the Higgs is to "give masses" to fundamental particles
- I believe that, in a more modern view, one can say that the role of the Higgs is to extend the validity of the Standard Model "far above" the weak scale

SM + non-linear EWSB

$$\mathscr{L}_{\mathrm{EWSB}} = \frac{v^2}{4}\langle D_\mu \Sigma (D^\mu \Sigma)^\dagger \rangle$$

$$\Sigma = e^{i\pi^a \sigma^a / v}$$



$$\propto \frac{s}{v^2}$$

$$\Lambda \sim 4\pi v$$

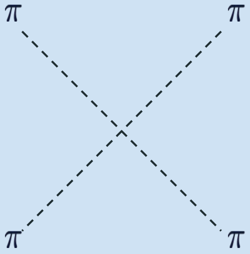
The Higgs in particle physics

- It is often said that the role of the Higgs is to "give masses" to fundamental particles
- I believe that, in a more modern view, one can say that the role of the Higgs is to extend the validity of the Standard Model "far above" the weak scale

SM + non-linear EWSB

$$\mathscr{L}_{\text{EWSB}} = \frac{v^2}{4}\langle D_\mu \Sigma (D^\mu \Sigma)^\dagger \rangle$$

$$\Sigma = e^{i\pi^a \sigma^a / v}$$



$$\propto \frac{s}{v^2}$$

$$\Lambda \sim 4\pi v$$

SM + linear EWSB (Higgs)

$$\mathscr{L}_{\text{EWSB}} = \langle D_\mu \mathcal{H} (D^\mu \mathcal{H})^\dagger \rangle - V(\mathcal{H})$$

$$\mathcal{H} = \frac{1}{\sqrt{2}}(v+h)\Sigma, \quad \Sigma = e^{i\pi^a \sigma^a / v}$$

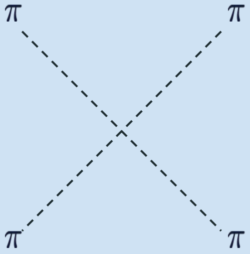
The Higgs in particle physics

- It is often said that the role of the Higgs is to "give masses" to fundamental particles
- I believe that, in a more modern view, one can say that the role of the Higgs is to extend the validity of the Standard Model "far above" the weak scale

SM + non-linear EWSB

$$\mathscr{L}_{\text{EWSB}} = \frac{v^2}{4}\langle D_\mu \Sigma (D^\mu \Sigma)^\dagger \rangle$$

$$\Sigma = e^{i\pi^a \sigma^a / v}$$



$$\propto \frac{s}{v^2}$$

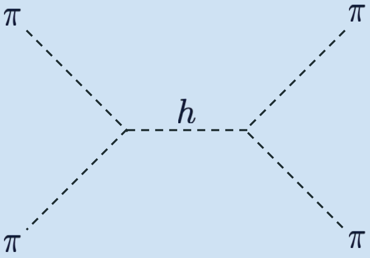
$$\Lambda \sim 4\pi v$$

SM + linear EWSB (Higgs)

$$\mathscr{L}_{\text{EWSB}} = \langle D_\mu \mathcal{H} (D^\mu \mathcal{H})^\dagger \rangle - V(\mathcal{H})$$

$$\mathcal{H} = \frac{1}{\sqrt{2}}(v+h)\Sigma, \quad \Sigma = e^{i\pi^a \sigma^a / v}$$

$$+$$

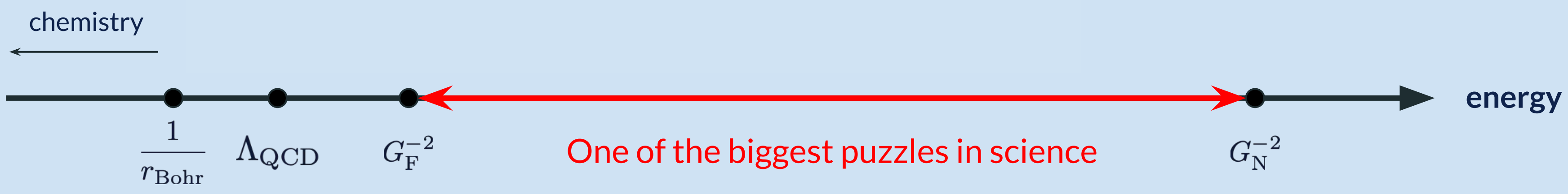


$$+\ \text{crossing} \propto \lambda$$

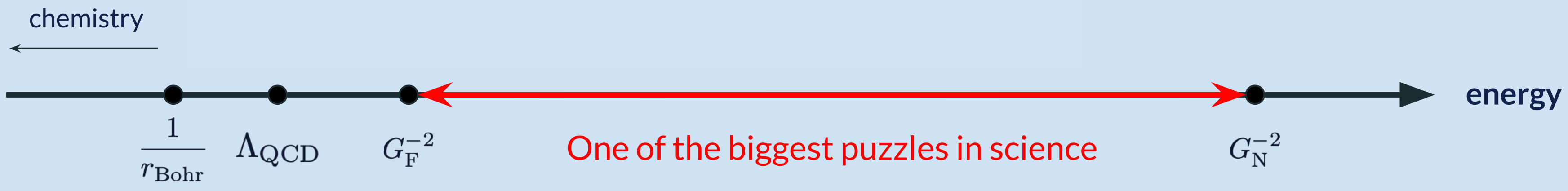
$$\Lambda \gg 4\pi v$$

How far above the weak scale?

How far above the weak scale?

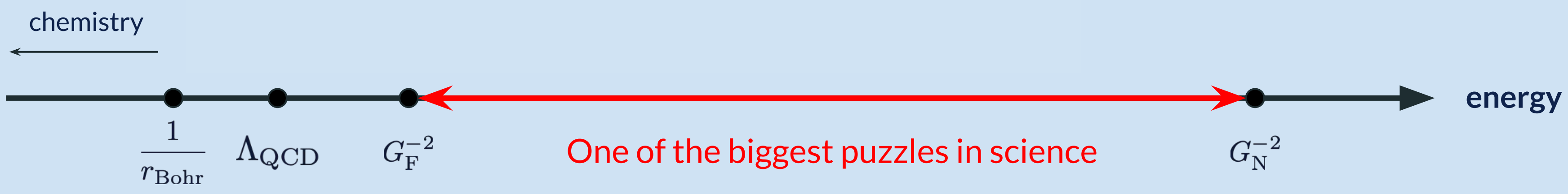


How far above the weak scale?



Large separations of scales in QFT are **unnatural!**

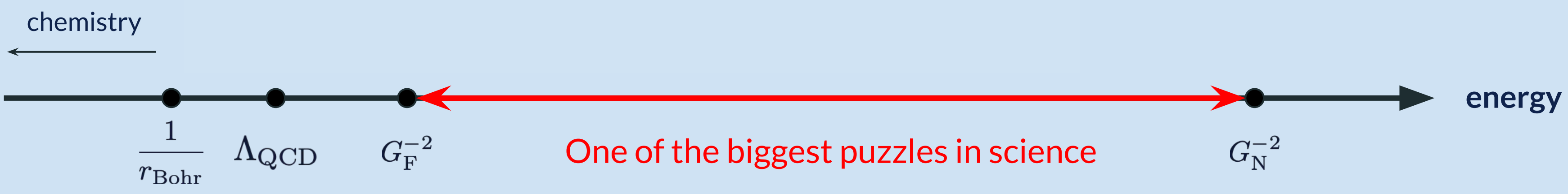
How far above the weak scale?



Large separations of scales in QFT are **unnatural!**

$$g_{\mathcal{O}}(\mu_{\rm IR}) = g_{\mathcal{O}}(\mu_{\rm UV}) \left(\frac{\mu_{\rm IR}}{\mu_{\rm UV}}\right)^{\Delta_{\mathcal{O}}-4}$$

How far above the weak scale?

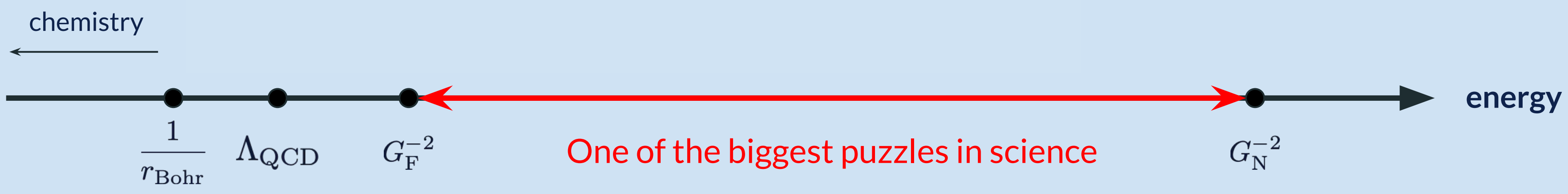


Large separations of scales in QFT are **unnatural!**

- Irrelevant $g_{\mathcal{O}}(\Lambda_{\mathrm{UV}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\mathrm{IR}}) \approx \left(\frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)^{c>0}$

$$g_{\mathcal{O}}(\mu_{\mathrm{IR}}) = g_{\mathcal{O}}(\mu_{\mathrm{UV}}) \left(\frac{\mu_{\mathrm{IR}}}{\mu_{\mathrm{UV}}}\right)^{\Delta_{\mathcal{O}}-4}$$

How far above the weak scale?

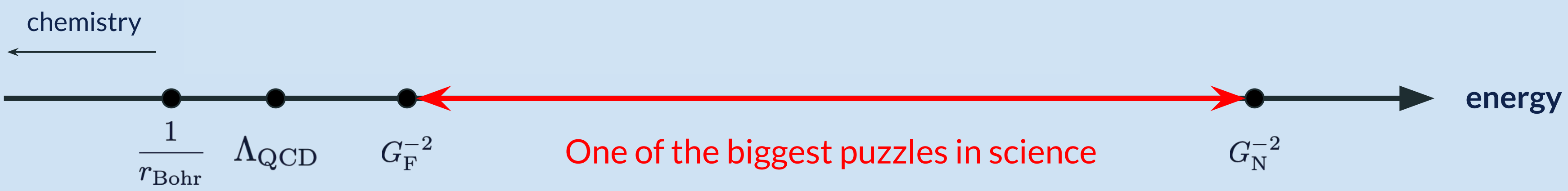


Large separations of scales in QFT are **unnatural!**

$$g_{\mathcal{O}}(\mu_{\mathrm{IR}}) = g_{\mathcal{O}}(\mu_{\mathrm{UV}}) \left(\frac{\mu_{\mathrm{IR}}}{\mu_{\mathrm{UV}}}\right)^{\Delta_{\mathcal{O}}-4}$$

- Irrelevant $\quad g_{\mathcal{O}}(\Lambda_{\mathrm{UV}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\mathrm{IR}}) \approx \left(\frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)^{c>0}$
- Marginal $\quad g_{\mathcal{O}}(\Lambda_{\mathrm{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\mathrm{UV}}) \approx \left(\frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)^{\epsilon>0} \approx \left(1 + \epsilon \log \frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)$

How far above the weak scale?

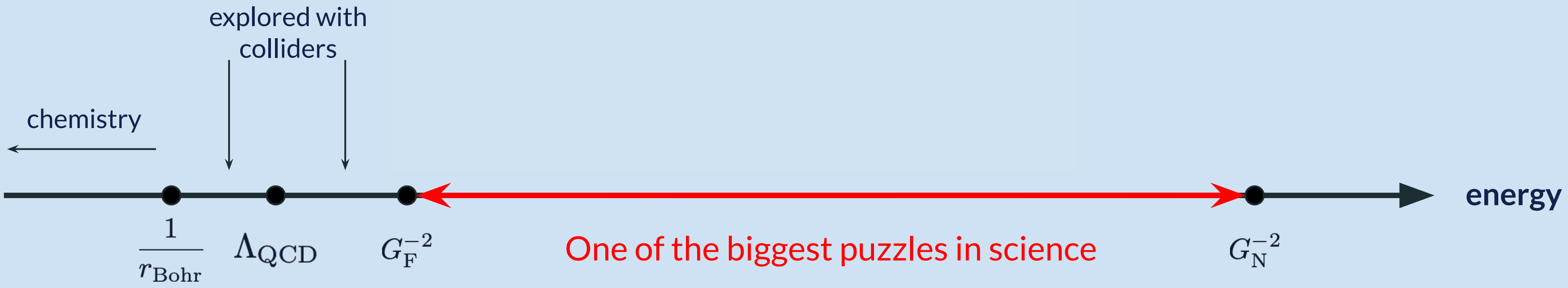


Large separations of scales in QFT are **unnatural!**

$$g_{\mathcal{O}}(\mu_{\text{IR}}) = g_{\mathcal{O}}(\mu_{\text{UV}}) \left(\frac{\mu_{\text{IR}}}{\mu_{\text{UV}}}\right)^{\Delta_{\mathcal{O}}-4}$$

- Irrelevant $\quad g_{\mathcal{O}}(\Lambda_{\text{UV}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\text{IR}}) \approx \left(\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)^{c>0}$
- Marginal $\quad g_{\mathcal{O}}(\Lambda_{\text{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\text{UV}}) \approx \left(\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)^{\epsilon>0} \approx \left(1+\epsilon \log \frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)$
- Relevant $\quad g_{\mathcal{O}}(\Lambda_{\text{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\text{UV}}) \approx \left(\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)^{c>0}$

How far above the weak scale?

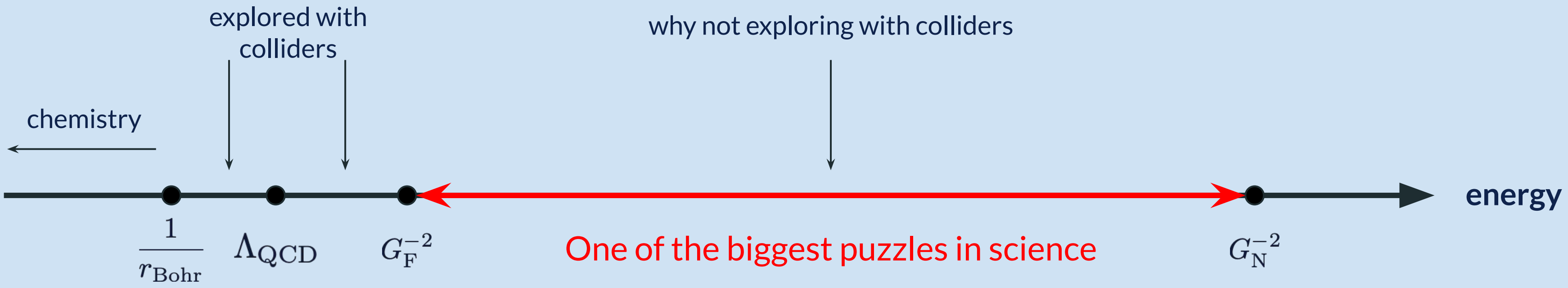


Large separations of scales in QFT are **unnatural!**

$$g_{\mathcal{O}}(\mu_{\text{IR}}) = g_{\mathcal{O}}(\mu_{\text{UV}}) \left(\frac{\mu_{\text{IR}}}{\mu_{\text{UV}}}\right)^{\Delta_{\mathcal{O}}-4}$$

- Irrelevant $\quad g_{\mathcal{O}}(\Lambda_{\text{UV}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\text{IR}}) \approx \left(\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)^{c>0}$
- Marginal $\quad g_{\mathcal{O}}(\Lambda_{\text{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\text{UV}}) \approx \left(\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)^{\epsilon>0} \approx \left(1+\epsilon\log\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)$
- Relevant $\quad g_{\mathcal{O}}(\Lambda_{\text{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\text{UV}}) \approx \left(\frac{\Lambda_{\text{IR}}}{\Lambda_{\text{UV}}}\right)^{c>0}$

How far above the weak scale?



Large separations of scales in QFT are **unnatural!**

$$g_{\mathcal{O}}(\mu_{\mathrm{IR}}) = g_{\mathcal{O}}(\mu_{\mathrm{UV}}) \left(\frac{\mu_{\mathrm{IR}}}{\mu_{\mathrm{UV}}}\right)^{\Delta_{\mathcal{O}}-4}$$

- Irrelevant $\quad g_{\mathcal{O}}(\Lambda_{\mathrm{UV}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\mathrm{IR}}) \approx \left(\frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)^{c>0}$
- Marginal $\quad g_{\mathcal{O}}(\Lambda_{\mathrm{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\mathrm{UV}}) \approx \left(\frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)^{\epsilon>0} \approx \left(1+\epsilon \log \frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)$
- Relevant $\quad g_{\mathcal{O}}(\Lambda_{\mathrm{IR}}) \approx O(4\pi) \implies g_{\mathcal{O}}(\Lambda_{\mathrm{UV}}) \approx \left(\frac{\Lambda_{\mathrm{IR}}}{\Lambda_{\mathrm{UV}}}\right)^{c>0}$

Machine Learning (vs Fitting)

Machine Learning (vs Fitting)

ML is about predicting output from input

Machine Learning (vs Fitting)

ML is about predicting output from input

- It is like fitting, but with a complicated non-linear function of many many parameters

Machine Learning (vs Fitting)

ML is about predicting output from input

- It is like fitting, but with a complicated non-linear function of many many parameters
- To prevent overfitting one needs to generalize well on unseen data

Machine Learning (vs Fitting)

ML is about predicting output from input

- It is like fitting, but with a complicated non-linear function of many many parameters
- To prevent overfitting one needs to generalize well on unseen data
- Validating on unseen data is what distinguishes learning from fitting

Machine Learning (vs Fitting)

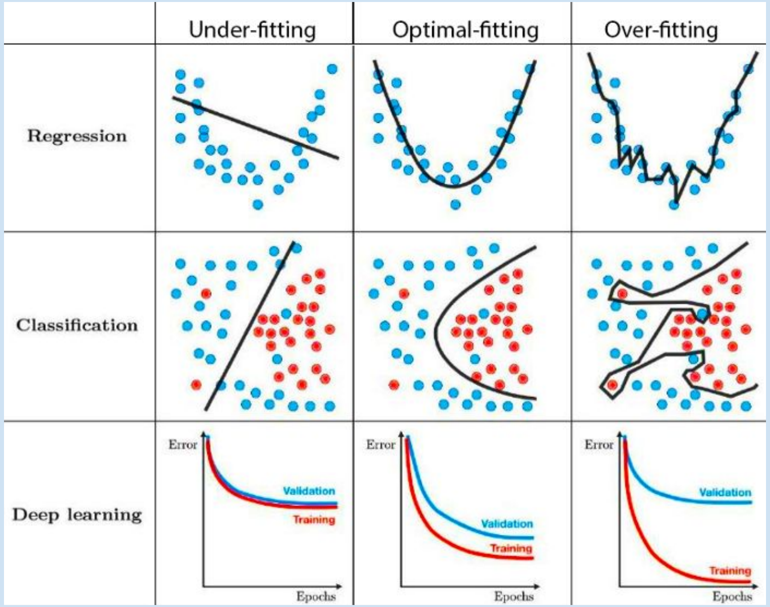
ML is about predicting output from input

- It is like fitting, but with a complicated non-linear function of many many parameters
- To prevent overfitting one needs to generalize well on unseen data
- Validating on unseen data is what distinguishes learning from fitting
- Several regularization techniques exist that help preventing overfitting

Machine Learning (vs Fitting)

ML is about predicting output from input

- It is like fitting, but with a complicated non-linear function of many many parameters
- To prevent overfitting one needs to generalize well on unseen data
- Validating on unseen data is what distinguishes learning from fitting
- Several regularization techniques exist that help preventing overfitting



Tasks vs complexity in ML

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Examples are:

- Scalar/vector to discrete scalar: this is a typical **classification** problem

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Examples are:

- Scalar/vector to discrete scalar: this is a typical **classification** problem
- Scalar/vector to continuous scalar: this is a typical **regression** problem

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Examples are:

- Scalar/vector to discrete scalar: this is a typical **classification** problem
- Scalar/vector to continuous scalar: this is a typical **regression** problem
- Vector to vector: **multivariate regression** with fixed length vectors

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Examples are:

- Scalar/vector to discrete scalar: this is a typical **classification** problem
- Scalar/vector to continuous scalar: this is a typical **regression** problem
- Vector to vector: **multivariate regression** with fixed length vectors
- Sequence to sequence: example are language translation, speech to text, etc.

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Examples are:

- Scalar/vector to discrete scalar: this is a typical **classification** problem
- Scalar/vector to continuous scalar: this is a typical **regression** problem
- Vector to vector: **multivariate regression** with fixed length vectors
- Sequence to sequence: example are language translation, speech to text, etc.
- Structured to structured: mapping between complex representations like images, graphs, text, etc.

Tasks vs complexity in ML

The complexity of the problem depends on the "representation" of input and output

Examples are:

- Scalar/vector to discrete scalar: this is a typical **classification** problem
- Scalar/vector to continuous scalar: this is a typical **regression** problem
- Vector to vector: **multivariate regression** with fixed length vectors
- Sequence to sequence: example are language translation, speech to text, etc.
- Structured to structured: mapping between complex representations like images, graphs, text, etc.

All the difficulty of ML is handling complicated dependencies that allow to capture higher-order, long-range correlations in data of arbitrary representation and dimension

The "QFT" of ML: Neural Networks

The "QFT" of ML: Neural Networks

The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

The "QFT" of ML: Neural Networks

The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

Perceptron vs neuron

Input vector of features

Weights vector

Bias

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b)$$

Perceptron: Heaviside activation
Neuron: more general activation

The "QFT" of ML: Neural Networks

The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

Perceptron vs neuron

Input vector of features

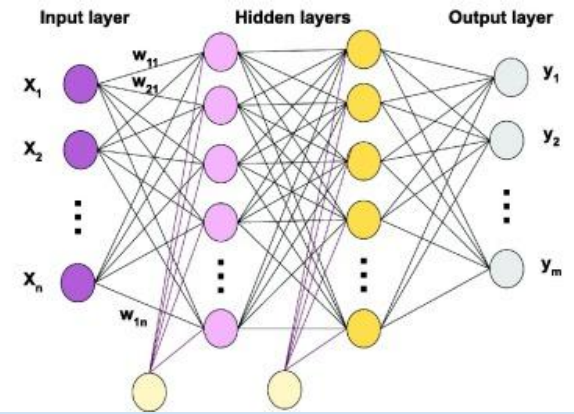
Weights vector

Bias

$$f(\boldsymbol{x}) = \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b)$$

Perceptron: Heaviside activation
Neuron: more general activation

Multilayer perceptron (feedforward Neural Network)



The "QFT" of ML: Neural Networks

The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

Cost (loss) function

$$f(X, W) = d(\boldsymbol{y}_{\mathrm{true}} - \boldsymbol{y}_{\mathrm{pred}}(X, W))$$

The "QFT" of ML: Neural Networks

The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

Cost (loss) function

$$f(X,W) = d(\boldsymbol{y}_{\text{true}} - \boldsymbol{y}_{\text{pred}}(X,W))$$

Optimization problem

$$\hat{W} = \arg\min_{W} f(X,W)$$

The "QFT" of ML: Neural Networks

The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

Cost (loss) function

$$f(X,W) = d(\boldsymbol{y}_{\text{true}} - \boldsymbol{y}_{\text{pred}}(X,W))$$

Optimization problem

$$\hat{W} = \arg\min_{W} f(X,W)$$

Gradient Descent with Backpropagation

$$W_{t+1} = W_t - \underbrace{\eta}_{\text{Learning rate}} \nabla_{W_t} f$$

$$\underbrace{\frac{\partial f}{\partial w_i} = \frac{\partial f}{\partial y_k} \cdots \frac{\partial z_j}{\partial w_i}}_{\text{Chain rule for derivatives}}$$

The "QFT" of ML: Neural Networks

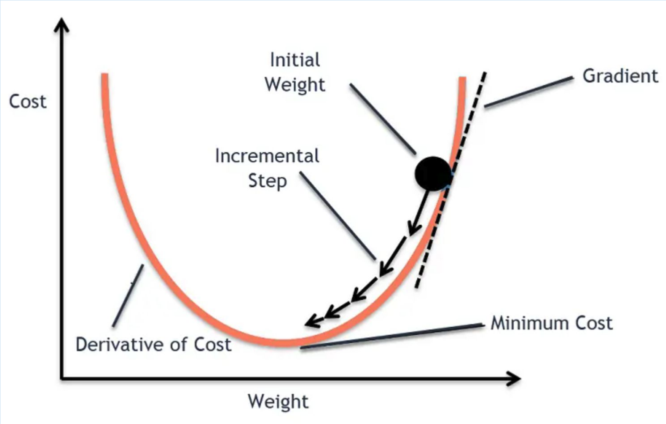
The framework/language of Machine Learning is provided by feed-forward Neural Networks (inspired by the perceptron model) trained with backpropagation using Gradient Descent techniques

Cost (loss) function

$$f(X, W) = d(\boldsymbol{y}_{\text{true}} - \boldsymbol{y}_{\text{pred}}(X, W))$$

Optimization problem

$$\hat{W} = \arg\min_{W} f(X, W)$$



Gradient Descent with Backpropagation

$$W_{t+1} = W_t - \eta \nabla_{W_t} f$$

Learning rate

$$\underbrace{\frac{\partial f}{\partial w_i} = \frac{\partial f}{\partial y_k} \cdots \frac{\partial z_j}{\partial w_i}}_{\text{Chain rule for derivatives}}$$

QED: Convolutional NN (CNN)

QED: Convolutional NN (CNN)

- CNN can be seen as the QED of Neural Networks

QED: Convolutional NN (CNN)

- CNN can be seen as the QED of Neural Networks
- They provide the basic building block to handle structured data (like images)

QED: Convolutional NN (CNN)

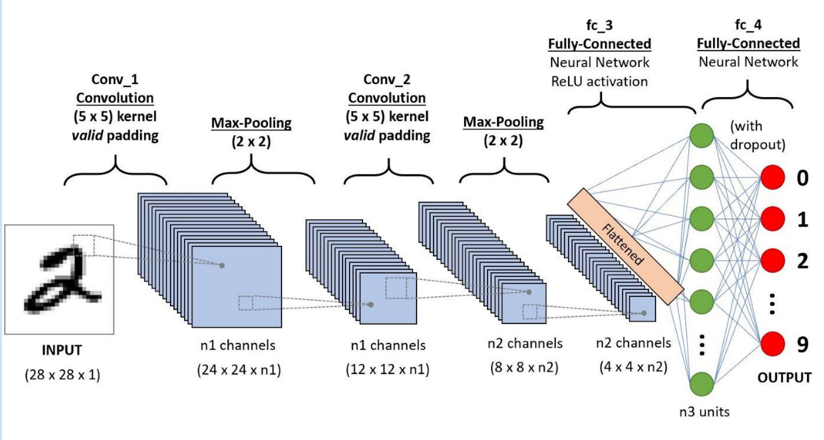
- CNN can be seen as the QED of Neural Networks
- They provide the basic building block to handle structured data (like images)
- Convolutional layers are designed to capture **local** correlations (analog to nearest neighbors dependencies) and extract relevant features

QED: Convolutional NN (CNN)

- CNN can be seen as the QED of Neural Networks
- They provide the basic building block to handle structured data (like images)
- Convolutional layers are designed to capture **local** correlations (analog to nearest neighbors dependencies) and extract relevant features
- CNN are highly scalable and almost ubiquitous in modern ML applications

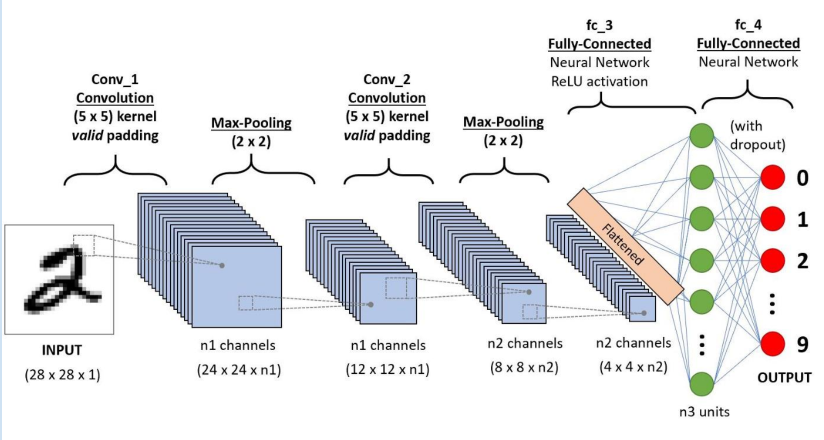
QED: Convolutional NN (CNN)

- CNN can be seen as the QED of Neural Networks
- They provide the basic building block to handle structured data (like images)
- Convolutional layers are designed to capture **local** correlations (analog to nearest neighbors dependencies) and extract relevant features
- CNN are highly scalable and almost ubiquitous in modern ML applications



QED: Convolutional NN (CNN)

- CNN can be seen as the QED of Neural Networks
- They provide the basic building block to handle structured data (like images)
- Convolutional layers are designed to capture **local** correlations (analog to nearest neighbors dependencies) and extract relevant features
- CNN are highly scalable and almost ubiquitous in modern ML applications



Convolutional filter

Convolution: $$Y_{i,j} = \sum_m \sum_n X_{i+m,j+n} \cdot K_{m,n} + b$$

Activation: $$Y'_{i,j} = \sigma(Y_{i,j})$$

Pooling (2x2): $$P_{i,j} = \max(Y'_{i,j}, Y'_{i+1,j}, Y'_{i,j+1}, Y'_{i+1,j+1})$$

Fermi Theory: Recursive NN (RNN)

Fermi Theory: Recursive NN (RNN)

- Sequences are lists of "tokens" (e.g. words in a sentence)

Fermi Theory: Recursive NN (RNN)

- Sequences are lists of "tokens" (e.g. words in a sentence)
- RNN are designed to handle sequences in which the order is essential

Fermi Theory: Recursive NN (RNN)

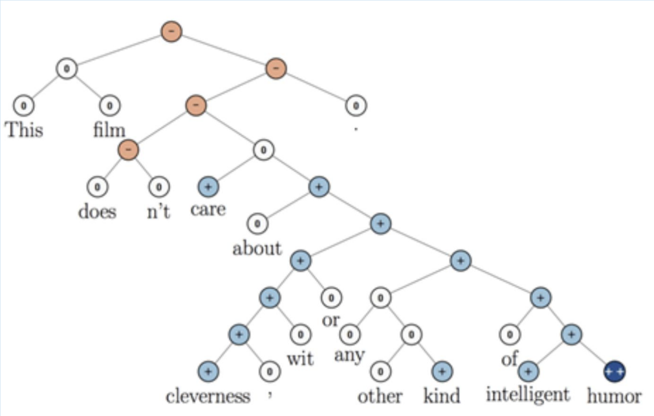
- Sequences are lists of "tokens" (e.g. words in a sentence)
- RNN are designed to handle sequences in which the order is essential
- RNN process data sequentially and can capture short- and medium-range correlations

Fermi Theory: Recursive NN (RNN)

- Sequences are lists of "tokens" (e.g. words in a sentence)
- RNN are designed to handle sequences in which the order is essential
- RNN process data sequentially and can capture short- and medium-range correlations
- As the Fermi theory has a limited range of validity, RNN cannot capture long-range correlation and become ineffective with very large sequences

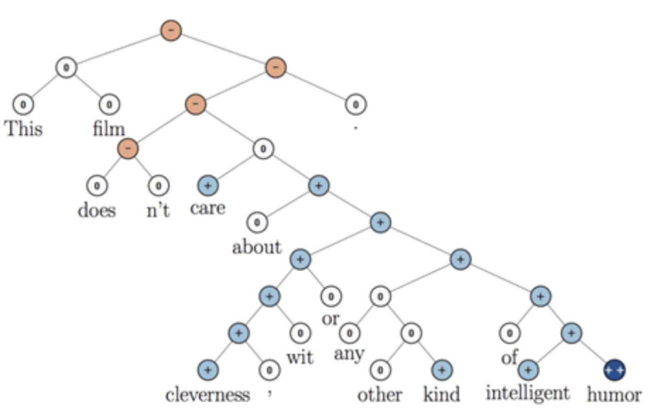
Fermi Theory: Recursive NN (RNN)

- Sequences are lists of "tokens" (e.g. words in a sentence)
- RNN are designed to handle sequences in which the order is essential
- RNN process data sequentially and can capture short- and medium-range correlations
- As the Fermi theory has a limited range of validity, RNN cannot capture long-range correlation and become ineffective with very large sequences



Fermi Theory: Recursive NN (RNN)

- Sequences are lists of "tokens" (e.g. words in a sentence)
- RNN are designed to handle sequences in which the order is essential
- RNN process data sequentially and can capture short- and medium-range correlations
- As the Fermi theory has a limited range of validity, RNN cannot capture long-range correlation and become ineffective with very large sequences



Hidden state update: $\boldsymbol{h}_t = \sigma(W_h \boldsymbol{h}_{t-1} + W_x \boldsymbol{x}_t + \boldsymbol{b})$

Output update: $\boldsymbol{y}_t = f(\boldsymbol{h}_t)$

Over long sequences it suffers from vanishing or exploding gradients, which limits the correlation length

QCD: Graph NN (GNN)

QCD: Graph NN (GNN)

- GNN go beyond grid-structured data and towards graph-structured ones (graphs can have arbitrary **non-local** correlations)

QCD: Graph NN (GNN)

- GNN go beyond grid-structured data and towards graph-structured ones (graphs can have arbitrary **non-local** correlations)
- GNN captures both local and global features

QCD: Graph NN (GNN)

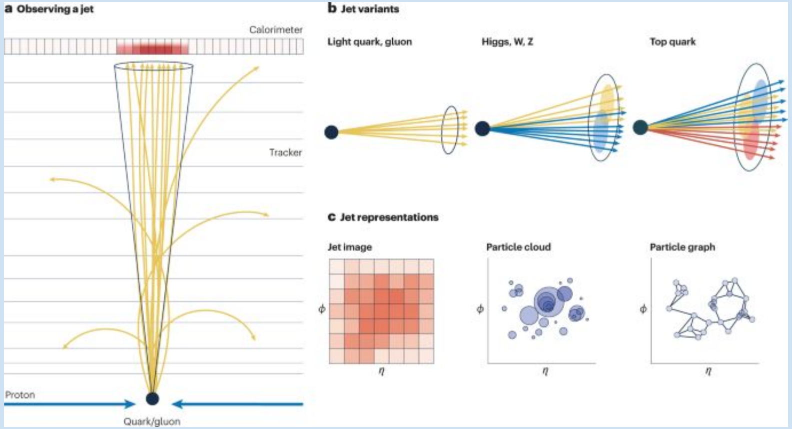
- GNN go beyond grid-structured data and towards graph-structured ones (graphs can have arbitrary **non-local** correlations)
- GNN captures both local and global features
- Just like QCD is an evolution of QED with stronger interactions, GNN are an evolution of CNN describing more sophisticated interactions

QCD: Graph NN (GNN)

- GNN go beyond grid-structured data and towards graph-structured ones (graphs can have arbitrary **non-local** correlations)
- GNN captures both local and global features
- Just like QCD is an evolution of QED with stronger interactions, GNN are an evolution of CNN describing more sophisticated interactions
- As QCD with respect to QED, GNN are more computationally intensive

QCD: Graph NN (GNN)

- GNN go beyond grid-structured data and towards graph-structured ones (graphs can have arbitrary **non-local** correlations)
- GNN captures both local and global features
- Just like QCD is an evolution of QED with stronger interactions, GNN are an evolution of CNN describing more sophisticated interactions
- As QCD with respect to QED, GNN are more computationally intensive



- GNN particularly suited for particle physics applications
- Clouds of particles with related observables naturally represented as graphs
- Great improvements in tagging and reconstruction

SM: Self-attention (Transformers)

SM: Self-attention (Transformers)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

~109K citations

(comparable to the total number of papers mentioning the Higgs in all Inspire)

SM: Self-attention (Transformers)

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

~109K citations

(comparable to the total number of papers mentioning the Higgs in all Inspire)

- Self-attention transforms a sequence into a another sequence (Transformer) passing through a fully connected graph with weighted edges (whose weights are learned)

SM: Self-attention (Transformers)

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

~109K citations

(comparable to the total number of papers mentioning the Higgs in all Inspire)

- Self-attention transforms a sequence into a another sequence (Transformer) passing through a fully connected graph with weighted edges (whose weights are learned)
- Differently from GNN, the graph structure is not fixed a priori and is always fully connected

SM: Self-attention (Transformers)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

~109K citations

(comparable to the total number of papers mentioning the Higgs in all Inspire)

- Self-attention transforms a sequence into a another sequence (Transformer) passing through a fully connected graph with weighted edges (whose weights are learned)
- Differently from GNN, the graph structure is not fixed a priori and is always fully connected
- Attention "heads" can implement overlapping graphs with edges of different nature

SM: Self-attention (Transformers)

Provided proper attribution is provided, Google hereby grants permission to reproduce the tables and figures in this paper solely for use in journalistic or scholarly works.

Attention Is All You Need

arXiv:1706.03762v7 [cs.CL] 2 Aug 2023

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

~109K citations

(comparable to the total number of papers mentioning the Higgs in all Inspire)

- Self-attention transforms a sequence into a another sequence (Transformer) passing through a fully connected graph with weighted edges (whose weights are learned)
- Differently from GNN, the graph structure is not fixed a priori and is always fully connected
- Attention "heads" can implement overlapping graphs with edges of different nature
- Computationally intensive, but scalable to any correlation length

SM: Self-attention (Transformers)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

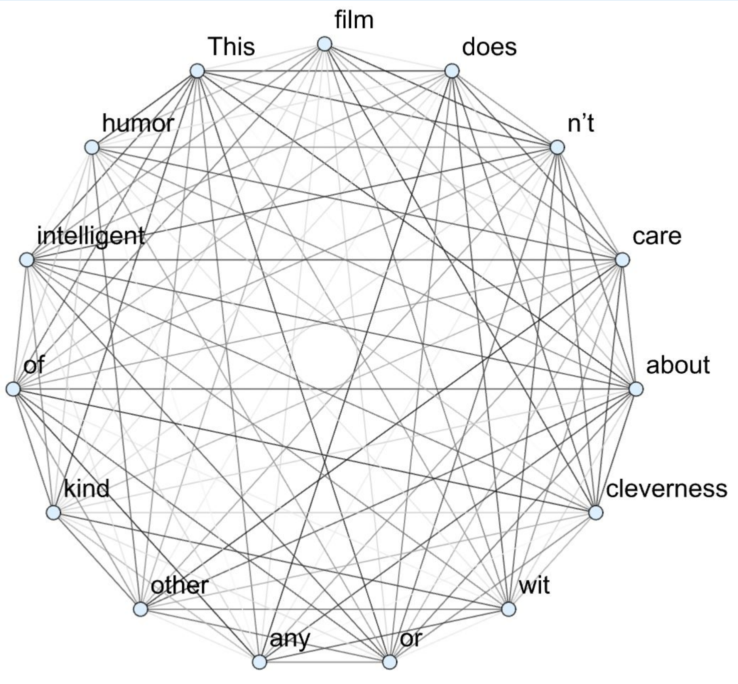
~109K citations

(comparable to the total number of papers mentioning the Higgs in all Inspire)

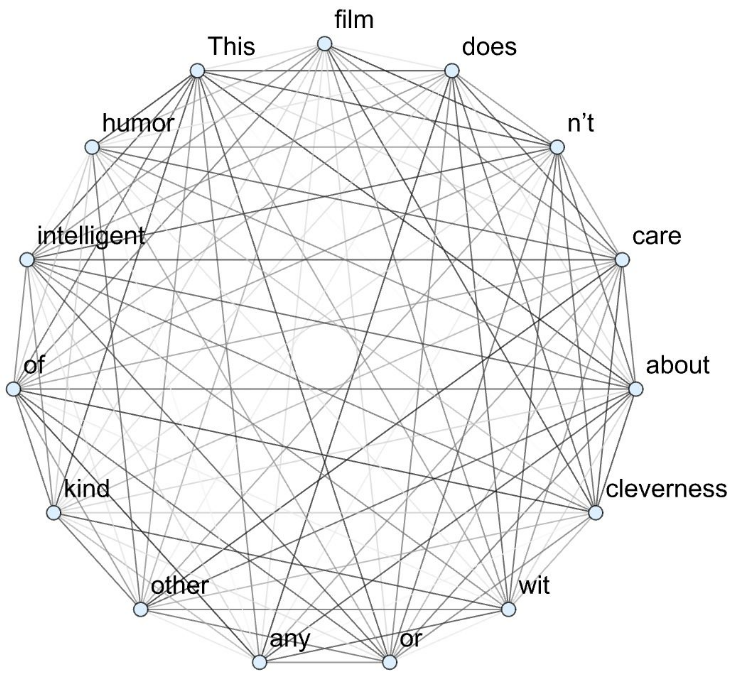
- Self-attention transforms a sequence into a another sequence (Transformer) passing through a fully connected graph with weighted edges (whose weights are learned)
- Differently from GNN, the graph structure is not fixed a priori and is always fully connected
- Attention "heads" can implement overlapping graphs with edges of different nature
- Computationally intensive, but scalable to any correlation length
- Only model able to catch long-range correlations (e.g. ChatGPT)

From sequences to weighted graphs

From sequences to weighted graphs

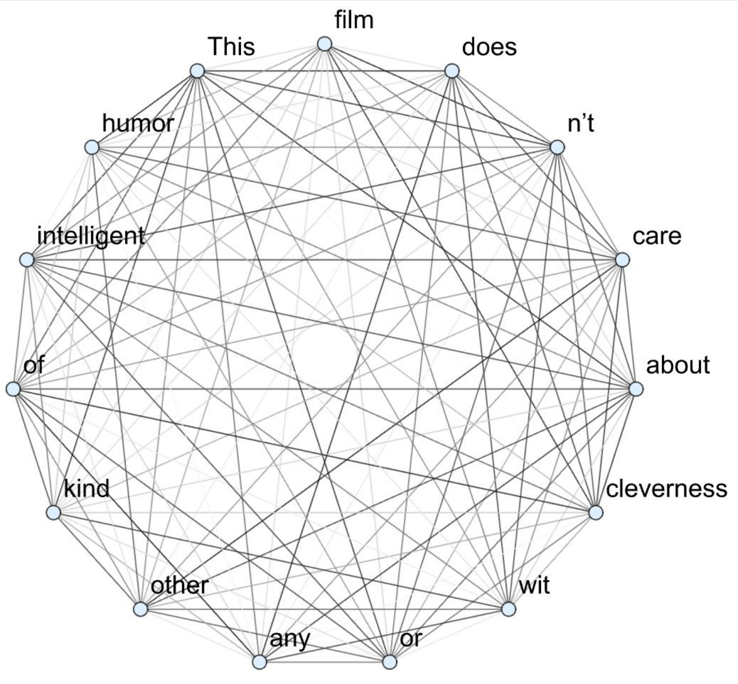


From sequences to weighted graphs



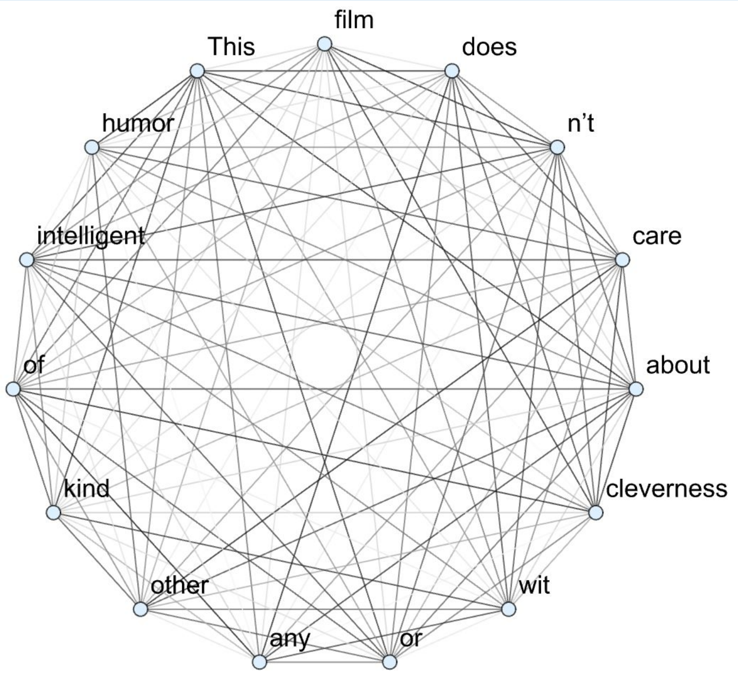
- The self-attention mechanism (layers) transform sequential data into "attention graphs" in which all weights are learned

From sequences to weighted graphs



- The self-attention mechanism (layers) transform sequential data into "attention graphs" in which all weights are learned
- In this way, instead of enforcing a graph structure a-priori, the graph is based, for each attention head, on learned attention of each token on any other token

From sequences to weighted graphs



- The self-attention mechanism (layers) transform sequential data into "attention graphs" in which all weights are learned
- In this way, instead of enforcing a graph structure a-priori, the graph is based, for each attention head, on learned attention of each token on any other token
- This idea, at the basis of Transformers, has revolutionized ML, allowing to extend the correlation range that NNs are able to capture

Why HEP should care about ML

Why HEP should care about ML

- HEP is the scientific field of science that produces and needs to analyse the largest amount of data, both real and synthetic (Monte Carlo)

Why HEP should care about ML

- HEP is the scientific field of science that produces and needs to analyse the largest amount of data, both real and synthetic (Monte Carlo)
- ML can find application in each step of the HEP Workflow

Why HEP should care about ML

- HEP is the scientific field of science that produces and needs to analyse the largest amount of data, both real and synthetic (Monte Carlo)
- ML can find application in each step of the HEP Workflow
  - Precision calculations
  - Monte Carlo simulations
  - Detector design
  - Detector operation
  - Data collection
  - Data analysis
  - Physics interpretation and hypothesis testing

Why HEP should care about ML

- HEP is the scientific field of science that produces and needs to analyse the largest amount of data, both real and synthetic (Monte Carlo)
- ML can find application in each step of the HEP Workflow
  - Precision calculations
  - Monte Carlo simulations
  - Detector design
  - Detector operation
  - Data collection
  - Data analysis
  - Physics interpretation and hypothesis testing
- Thanks to the technologies I discussed earlier, ML has been shown to be able to improve over traditional techniques in all of these tasks

Why ML should care about HEP

Why ML should care about HEP

- On the other hand, HEP researchers can actively contribute to the ongoing ML revolution

Why ML should care about HEP

- On the other hand, HEP researchers can actively contribute to the ongoing ML revolution
- Indeed HEP represents the natural playground to better understand ML
    - Large amount of clean data

Why ML should care about HEP

- On the other hand, HEP researchers can actively contribute to the ongoing ML revolution
- Indeed HEP represents the natural playground to better understand ML
  - Large amount of clean data
  - Knowledge of the model underlying data

Why ML should care about HEP

- On the other hand, HEP researchers can actively contribute to the ongoing ML revolution
- Indeed HEP represents the natural playground to better understand ML
  - Large amount of clean data
  - Knowledge of the model underlying data
  - Advanced/robust statistical workflows

Why ML should care about HEP

- On the other hand, HEP researchers can actively contribute to the ongoing ML revolution
- Indeed HEP represents the natural playground to better understand ML
    - Large amount of clean data
    - Knowledge of the model underlying data
    - Advanced/robust statistical workflows
    - Knowledge of the physical behavior of complex systems (there is an entire field trying to explain NNs in terms of concepts coming from physics, from the Ising model, to the RG)

Why ML should care about HEP

- On the other hand, HEP researchers can actively contribute to the ongoing ML revolution
- Indeed HEP represents the natural playground to better understand ML
  - Large amount of clean data
  - Knowledge of the model underlying data
  - Advanced/robust statistical workflows
  - Knowledge of the physical behavior of complex systems (there is an entire field trying to explain NNs in terms of concepts coming from physics, from the Ising model, to the RG)
  - Analogy of the faced problems as stressed in this talk

Challenges

Challenges

ML is mostly developed "by companies for companies"

This raises some issues that also represent the biggest challenges

Challenges

ML is mostly developed "by companies for companies"

This raises some issues that also represent the biggest challenges

- Data representation

Challenges

ML is mostly developed "by companies for companies"

This raises some issues that also represent the biggest challenges

- Data representation
- Precision

Challenges

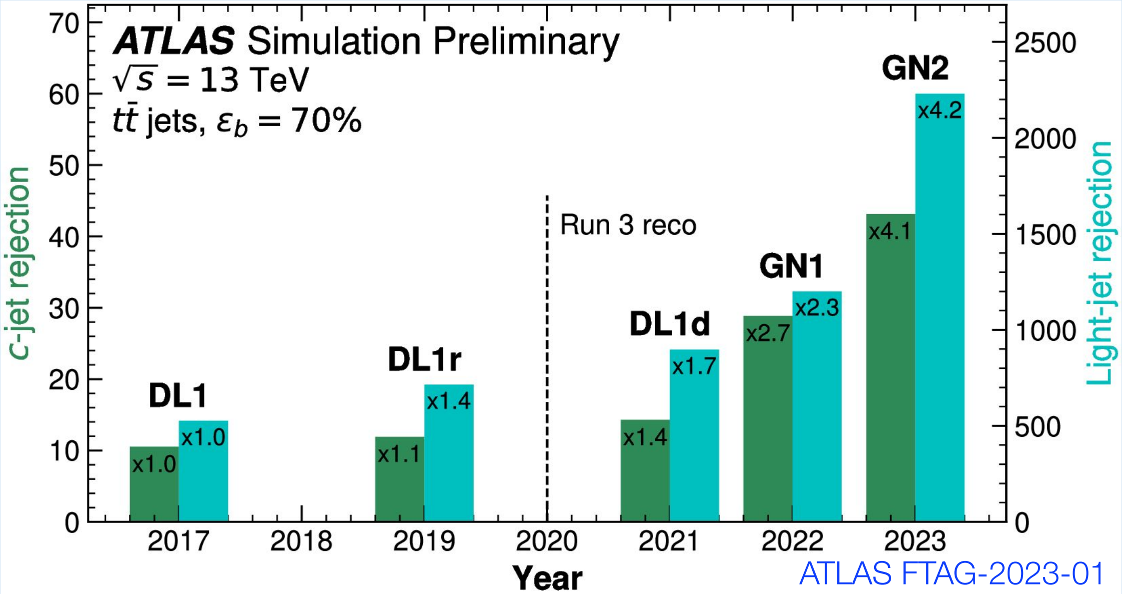
ML is mostly developed "by companies for companies"

This raises some issues that also represent the biggest challenges

- Data representation
- Precision
- Evaluation

Choosing the right representation

Choosing the right representation



Factor of 4x in 4 years, a tool-induced "revolution" in HEP!

Choosing the right representation (graphs) together with implementing "attention" let to an impressive improvement

The quest for precision

The quest for precision

- Including uncertainties into ML models is essential for their use in HEP

The quest for precision

- Including uncertainties into ML models is essential for their use in HEP
- Physics information may also be beneficial (physics-informed ML)

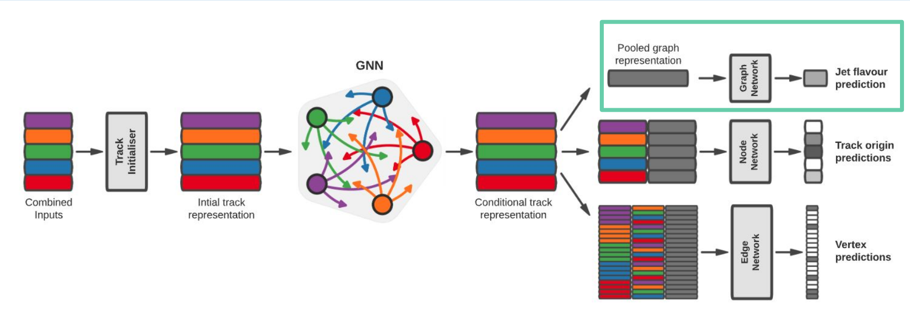
The quest for precision

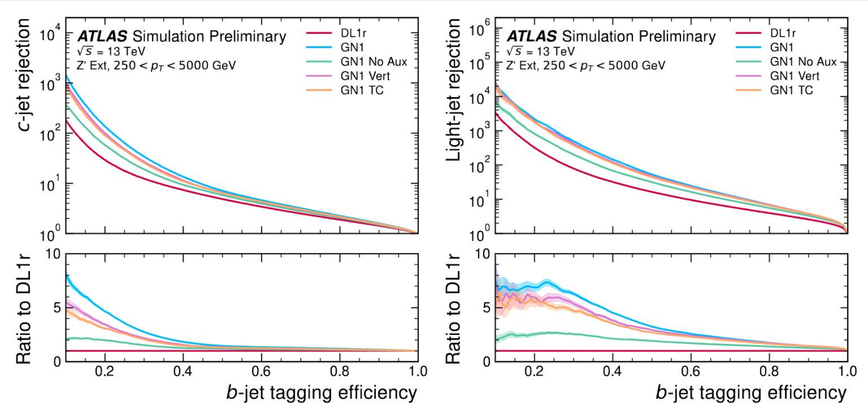
- Including uncertainties into ML models is essential for their use in HEP
- Physics information may also be beneficial (physics-informed ML)
- Physics information can also be included through "auxiliary tasks"

The quest for precision

- Including uncertainties into ML models is essential for their use in HEP
- Physics information may also be beneficial (physics-informed ML)
- Physics information can also be included through "auxiliary tasks"

ATLAS PHYS-PUB-2022-027

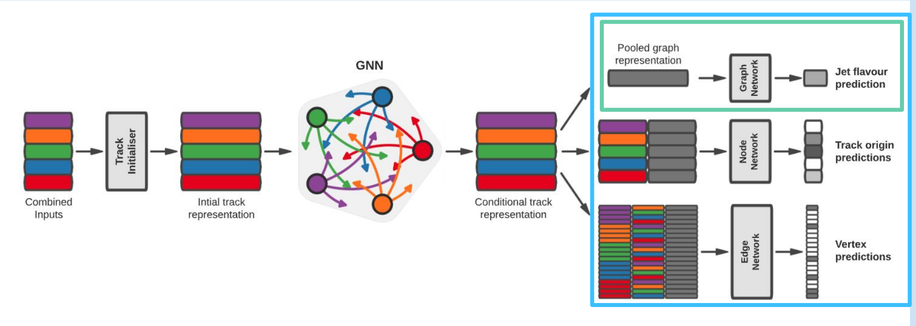


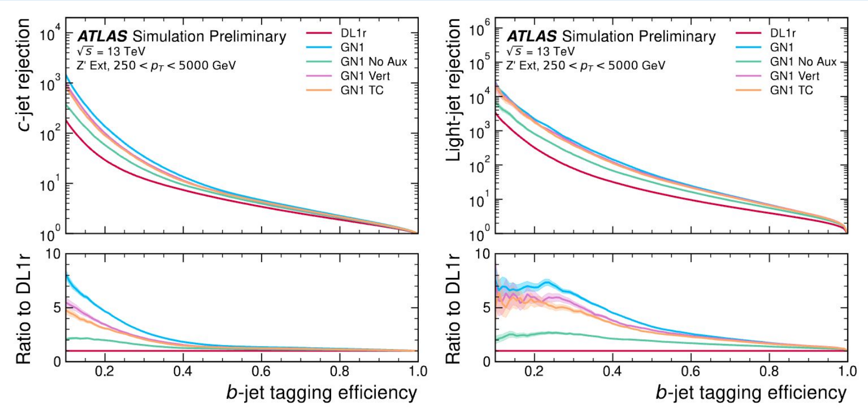


The quest for precision

- Including uncertainties into ML models is essential for their use in HEP
- Physics information may also be beneficial (physics-informed ML)
- Physics information can also be included through "auxiliary tasks"

ATLAS PHYS-PUB-2022-027





Developing robust evaluation

Developing robust evaluation

Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences

Samuele Grossi[a,b], Marco Letizia[b,c], and Riccardo Torre[a,b]

[a] *Department of Physics, University of Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[b] *INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[c] *MaLGa-DIBRIS, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy*

September 26, 2024

arXiv:2409.16336v1 [stat.ML] 24 Sep 2024

Abstract

We propose a robust methodology to evaluate the performance and computational efficiency of non-parametric two-sample tests, specifically designed for high-dimensional generative models in scientific applications such as in particle physics. The study focuses on tests built from univariate integral probability measures: the sliced Wasserstein distance and the mean of the Kolmogorov-Smirnov statistics, already discussed in the literature, and the novel sliced Kolmogorov-Smirnov statistic. These metrics can be evaluated in parallel, allowing for fast and reliable estimates of their distribution under the null hypothesis. We also compare these metrics with the recently proposed unbiased Fréchet Gaussian Distance and the unbiased quadratic Maximum Mean Discrepancy, computed with a quartic polynomial kernel. We evaluate the proposed tests on various distributions, focusing on their sensitivity to deformations parameterized by a single parameter ϵ. Our experiments include correlated Gaussians and mixtures of Gaussians in 5, 20, and 100 dimensions, and a particle physics dataset of gluon jets from the JetNet dataset, considering both jet- and particle-level features. Our results demonstrate that one-dimensional-based tests provide a level of sensitivity comparable to other multivariate metrics, but with significantly lower computational cost, making them ideal for evaluating generative models in high-dimensional settings. This methodology offers an efficient, standardized tool for model comparison and can serve as a benchmark for more advanced tests, including machine-learning-based approaches.

Keywords: Non-Parametric Two-Sample Tests, Multivariate Hypothesis Testing, Integral Probability Measure, Generative Models, Generative Models Evaluation

Developing robust evaluation

Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences

Samuele Grossi[a,b], Marco Letizia[b,c], and Riccardo Torre[a,b]

[a] *Department of Physics, University of Genova, Via Dodecaneso 33, I-16146 Genova, Italy*

[b] *INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy*

[c] *MaLGa-DIBRIS, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy*

September 26, 2024

arXiv:2409.16336v1 [stat.ML] 24 Sep 2024

Abstract

We propose a robust methodology to evaluate the performance and computational efficiency of non-parametric two-sample tests, specifically designed for high-dimensional generative models in scientific applications such as in particle physics. The study focuses on tests built from univariate integral probability measures: the sliced Wasserstein distance and the mean of the Kolmogorov-Smirnov statistics, already discussed in the literature, and the novel sliced Kolmogorov-Smirnov statistic. These metrics can be evaluated in parallel, allowing for fast and reliable estimates of their distribution under the null hypothesis. We also compare these metrics with the recently proposed unbiased Fréchet Gaussian Distance and the unbiased quadratic Maximum Mean Discrepancy, computed with a quartic polynomial kernel. We evaluate the proposed tests on various distributions, focusing on their sensitivity to deformations parameterized by a single parameter ϵ. Our experiments include correlated Gaussians and mixtures of Gaussians in 5, 20, and 100 dimensions, and a particle physics dataset of gluon jets from the JetNet dataset, considering both jet- and particle-level features. Our results demonstrate that one-dimensional-based tests provide a level of sensitivity comparable to other multivariate metrics, but with significantly lower computational cost, making them ideal for evaluating generative models in high-dimensional settings. This methodology offers an efficient, standardized tool for model comparison and can serve as a benchmark for more advanced tests, including machine-learning-based approaches.

Keywords: Non-Parametric Two-Sample Tests, Multivariate Hypothesis Testing, Integral Probability Measure, Generative Models, Generative Models Evaluation

- Develop new "metrics" for comparing multivariate probability distributions

Developing robust evaluation

Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences

Samuele Grossi[a,b], Marco Letizia[b,c], and Riccardo Torre[a,b]

[a] *Department of Physics, University of Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[b] *INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[c] *MaLGa-DIBRIS, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy*

September 26, 2024

Abstract

We propose a robust methodology to evaluate the performance and computational efficiency of non-parametric two-sample tests, specifically designed for high-dimensional generative models in scientific applications such as in particle physics. The study focuses on tests built from univariate integral probability measures: the sliced Wasserstein distance and the mean of the Kolmogorov-Smirnov statistics, already discussed in the literature, and the novel sliced Kolmogorov-Smirnov statistic. These metrics can be evaluated in parallel, allowing for fast and reliable estimates of their distribution under the null hypothesis. We also compare these metrics with the recently proposed unbiased Fréchet Gaussian Distance and the unbiased quadratic Maximum Mean Discrepancy, computed with a quartic polynomial kernel. We evaluate the proposed tests on various distributions, focusing on their sensitivity to deformations parameterized by a single parameter ϵ. Our experiments include correlated Gaussians and mixtures of Gaussians in 5, 20, and 100 dimensions, and a particle physics dataset of gluon jets from the JetNet dataset, considering both jet- and particle-level features. Our results demonstrate that one-dimensional-based tests provide a level of sensitivity comparable to other multivariate metrics, but with significantly lower computational cost, making them ideal for evaluating generative models in high-dimensional settings. This methodology offers an efficient, standardized tool for model comparison and can serve as a benchmark for more advanced tests, including machine-learning-based approaches.

Keywords: Non-Parametric Two-Sample Tests, Multivariate Hypothesis Testing, Integral Probability Measure, Generative Models, Generative Models Evaluation

- Develop new "metrics" for comparing multivariate probability distributions
- Introduce robust statistical framework for comparing the performances of such metrics

Developing robust evaluation

Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences

Samuele Grossi[a,b], Marco Letizia[b,c], and Riccardo Torre[a,b]

[a] *Department of Physics, University of Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[b] *INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[c] *MaLGa-DIBRIS, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy*

September 26, 2024

arXiv:2409.16336v1 [stat.ML] 24 Sep 2024

Abstract

We propose a robust methodology to evaluate the performance and computational efficiency of non-parametric two-sample tests, specifically designed for high-dimensional generative models in scientific applications such as in particle physics. The study focuses on tests built from univariate integral probability measures: the sliced Wasserstein distance and the mean of the Kolmogorov-Smirnov statistics, already discussed in the literature, and the novel sliced Kolmogorov-Smirnov statistic. These metrics can be evaluated in parallel, allowing for fast and reliable estimates of their distribution under the null hypothesis. We also compare these metrics with the recently proposed unbiased Fréchet Gaussian Distance and the unbiased quadratic Maximum Mean Discrepancy, computed with a quartic polynomial kernel. We evaluate the proposed tests on various distributions, focusing on their sensitivity to deformations parameterized by a single parameter ϵ. Our experiments include correlated Gaussians and mixtures of Gaussians in 5, 20, and 100 dimensions, and a particle physics dataset of gluon jets from the JetNet dataset, considering both jet- and particle-level features. Our results demonstrate that one-dimensional-based tests provide a level of sensitivity comparable to other multivariate metrics, but with significantly lower computational cost, making them ideal for evaluating generative models in high-dimensional settings. This methodology offers an efficient, standardized tool for model comparison and can serve as a benchmark for more advanced tests, including machine-learning-based approaches.

Keywords: Non-Parametric Two-Sample Tests, Multivariate Hypothesis Testing, Integral Probability Measure, Generative Models, Generative Models Evaluation

- Develop new "metrics" for comparing multivariate probability distributions
- Introduce robust statistical framework for comparing the performances of such metrics
- Validate results on complicated and representative toy distributions

Developing robust evaluation

Refereeing the Referees: Evaluating Two-Sample Tests for Validating Generators in Precision Sciences

Samuele Grossi[a,b], Marco Letizia[b,c], and Riccardo Torre[a,b]

[a] *Department of Physics, University of Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[b] *INFN, Sezione di Genova, Via Dodecaneso 33, I-16146 Genova, Italy*
[c] *MaLGa-DIBRIS, University of Genova, Via Dodecaneso 35, I-16146 Genova, Italy*

September 26, 2024

Abstract

We propose a robust methodology to evaluate the performance and computational efficiency of non-parametric two-sample tests, specifically designed for high-dimensional generative models in scientific applications such as in particle physics. The study focuses on tests built from univariate integral probability measures: the sliced Wasserstein distance and the mean of the Kolmogorov-Smirnov statistics, already discussed in the literature, and the novel sliced Kolmogorov-Smirnov statistic. These metrics can be evaluated in parallel, allowing for fast and reliable estimates of their distribution under the null hypothesis. We also compare these metrics with the recently proposed unbiased Fréchet Gaussian Distance and the unbiased quadratic Maximum Mean Discrepancy, computed with a quartic polynomial kernel. We evaluate the proposed tests on various distributions, focusing on their sensitivity to deformations parameterized by a single parameter ϵ. Our experiments include correlated Gaussians and mixtures of Gaussians in 5, 20, and 100 dimensions, and a particle physics dataset of gluon jets from the JetNet dataset, considering both jet- and particle-level features. Our results demonstrate that one-dimensional-based tests provide a level of sensitivity comparable to other multivariate metrics, but with significantly lower computational cost, making them ideal for evaluating generative models in high-dimensional settings. This methodology offers an efficient, standardized tool for model comparison and can serve as a benchmark for more advanced tests, including machine-learning-based approaches.

Keywords: Non-Parametric Two-Sample Tests, Multivariate Hypothesis Testing, Integral Probability Measure, Generative Models, Generative Models Evaluation

- Develop new "metrics" for comparing multivariate probability distributions
- Introduce robust statistical framework for comparing the performances of such metrics
- Validate results on complicated and representative toy distributions
- Validate results on physics datasets (gluon jets with both jet and particle-level information)

Summary

Summary

- Revolutions driven by tools are at least as important as revolutions inspired by new concepts, and generally more frequent

Summary

- Revolutions driven by tools are at least as important as revolutions inspired by new concepts, and generally more frequent
- It seems, in HEP, we are struggling looking for a concept-induced revolution

Summary

- Revolutions driven by tools are at least as important as revolutions inspired by new concepts, and generally more frequent
- It seems, in HEP, we are struggling looking for a concept-induced revolution
- Maybe we should also strongly pursue the idea of a tool-induced one, remembering what Dyson said about the impact of George Green: tools are also mathematical tools, not only "screwdrivers"

Summary

- Revolutions driven by tools are at least as important as revolutions inspired by new concepts, and generally more frequent
- It seems, in HEP, we are struggling looking for a concept-induced revolution
- Maybe we should also strongly pursue the idea of a tool-induced one, remembering what Dyson said about the impact of George Green: tools are also mathematical tools, not only "screwdrivers"
- ML provides, together with the hardware revolution of GPUs, a "new bag of tools for discovery"

Summary

- Revolutions driven by tools are at least as important as revolutions inspired by new concepts, and generally more frequent
- It seems, in HEP, we are struggling looking for a concept-induced revolution
- Maybe we should also strongly pursue the idea of a tool-induced one, remembering what Dyson said about the impact of George Green: tools are also mathematical tools, not only "screwdrivers"
- ML provides, together with the hardware revolution of GPUs, a "new bag of tools for discovery"
- ML is not "just for experimentalists", as it is not just a practical tool

Summary

- Revolutions driven by tools are at least as important as revolutions inspired by new concepts, and generally more frequent
- It seems, in HEP, we are struggling looking for a concept-induced revolution
- Maybe we should also strongly pursue the idea of a tool-induced one, remembering what Dyson said about the impact of George Green: tools are also mathematical tools, not only "screwdrivers"
- ML provides, together with the hardware revolution of GPUs, a "new bag of tools for discovery"
- ML is not "just for experimentalists", as it is not just a practical tool
- The role of theorists is essential, as it was in understanding how to use Green's function to explain physical phenomena

Thank you for your attention!

Thank you for your attention!

And thanks Riccardo for his mentorship and inspiration!

Auguri!