Enhancing generalization in high energy physics using white-box adversarial attacks

ROTHEN, Franck[1]
franck.rothen@unige.ch

KLEIN, Samuel[1]
samuel.klein@unige.ch

LEIGH, Matthew[1]
matthew.leigh@unige.ch

GOLLING, Tobias[1]
tobias.golling@unige.ch

06. November 2024

[1]University of Geneva, Faculty of science, DPNC

Introduction and Motivation

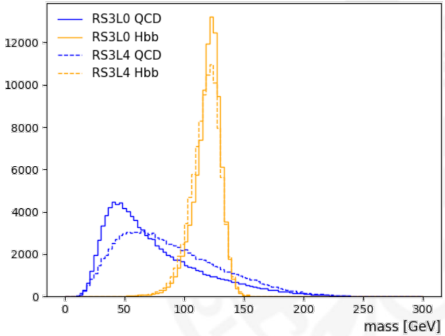
- Uncover new fundamental physics at the LHC through advanced reconstruction and classification algorithms.
- Machine learning is a key tool for background discrimination, e.g., in rare Higgs or SUSY decay.
- Supervised models are trained on Monte Carlo data and tested on real data.

- This study warns against over-reliance on simulation artifacts and poor generalization to real data.
- The key target of this study is improving generalization performance.

Table of Contents

Experimental setup

- Classification task
- Higgs decay $H \to b\bar{b}$ as signal
- QCD jets as background
- Re-simulation based dataset (RS3L) arXiv:2403.07066, Harris et al.
- Physical processes are generated and re-showered using different simulators
- Both dense and transformer models are used



Augmentation	Description
RS3L0	Jet showered with Pythia8 (Nominal scenario)
RS3L4	Use of Herwig7 as parton shower

Generalization performance of default models

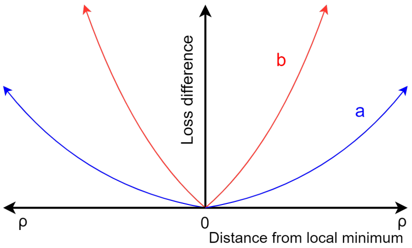
- Models are trained and cross-evaluated on both Herwig and Pythia

Table: Inverse of the FPR at 0.85 signal efficiency

	Evaluation sets	
Training sets	Pythia	Herwig
Pythia	$\mathbf{24.2 \pm 0.4}$	11.3 ± 0.2
Herwig	15.0 ± 0.2	$\mathbf{21.2 \pm 0.4}$

- Poor generalization between Pythia and Herwig datasets.
- Models overfit to simulation.
- Need for cross-evaluation performance improvement.

Sharpness definition and relation with generalization



- Simpler models generalize better (MDL principle).

  Neural Computation, 9(1):1–42, Hochreiter et al.

- Sharpness as a proxy for model complexity.

  arXiv:1609.04836, Keskar et al.

Definition 1. Sharpness

A minimum b is sharper than a minimum a if,

$$\mathop{\mathbb{E}}_{||\delta||=\epsilon}\left[\Delta\mathcal{L}_a\left(\delta\right)\right] \leq \mathop{\mathbb{E}}_{||\delta||=\epsilon}\left[\Delta\mathcal{L}_b\left(\delta\right)\right], \forall\epsilon \in \mathbb{R}_+,$$

where $\Delta\mathcal{L}_i\left(\delta\right) := \mathcal{L}_i\left(x+\delta\right) - \mathcal{L}_i\left(x\right)$ is the increase in loss due to the perturbation δ for the local minimum i.

Adversarial attacks

Theoretical Adversarial Loss $\mathcal{L}_{\mathcal{A}}$

$$\mathcal{L}_{\mathsf{A}}\left(w, x, y\right) = \max_{\|\delta\|<\epsilon} \mathcal{L}\left(w, x + \delta, y\right),$$

where ϵ is the perturbation strength.

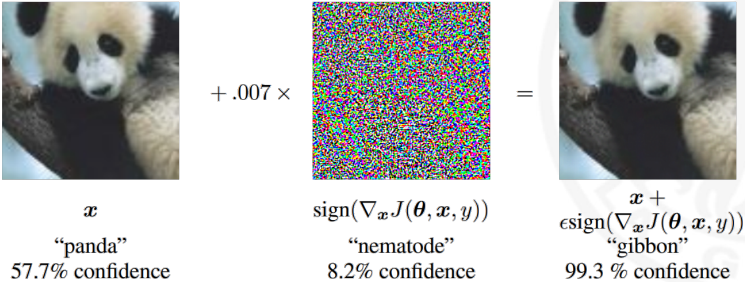


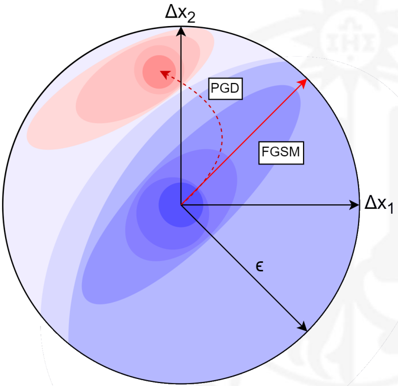
Figure: arXiv:1412.6572, Goodfellow et al.

Feature Space Perturbation: FGSM and PGD

Fast Gradient Sign Method (FGSM)

$$x \to x' = x + \epsilon \cdot \text{sign}\nabla_x \mathcal{L}(w, x, y).$$

- FGSM is a first-order Taylor expansion.
- Backpropagation needs to be performed twice.
- Projected Gradient Descent (PGD) is obtained by iterating FGSM.
- PGD is more effective but computationally expensive.



arXiv:1412.6572, Goodfellow et al.
arXiv:1706.06083, Madry et al.

Weight Space Perturbation: SAM and SSAM-D

$$\mathcal{L}_{\mathsf{A}} = \max_{\|\epsilon\| \leq \rho} \mathcal{L}\left(w + \epsilon, x, y\right)$$

- Sharpness Aware Minimization (SAM) is a first-order Taylor expansion.

Sharpness Aware Minimization (SAM)

$$\epsilon_{SAM} := \rho \cdot \mathrm{sign}\left(\nabla_w \mathcal{L}\left(w\right)\right)$$

$$\nabla_w \mathcal{L}_{\mathsf{SAM}} \approx \nabla_w \mathcal{L}(w)|_{w+\epsilon}$$

arXiv:2010.01412, Foret et al.
arXiv:2210.05177, Mi et al.

Dynamical Sparse SAM (SSAM-D)

$$\mathcal{L}_{\mathsf{SSAM}} := \max_{\|\epsilon\| \leq \rho} \mathcal{L}\left(w + \epsilon \odot \mathbf{m}_w\right),$$

where $\mathbf{m}_w$ is the mask.

- Only 5% of the weights exhibit sharp behavior.
- The aim is to reduce the training penalty by focusing on these weights.
- The mask is dynamically updated during training.

Sharpness analysis: Sampling and Gradient Ascent

- A way to quantify loss sharpness is desired.
- Direct computation results in:

$$\mathbb{E}\left[\mathcal{L}_\rho\right] = \frac{1}{S_\epsilon} \oint_{\|\epsilon\|=\rho} \mathcal{L}\left(w+\epsilon, x, y\right) d^n\epsilon \overset{\text{M.C}}{\approx} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\left(w+\epsilon_i, x, y\right),$$

- In practice, this requires too many random perturbations to obtain an accurate average of the landscape.
- Instead, let's consider the sharpness upper bound as a proxy:

$$\max_{\|\epsilon\|\leq\rho} \mathcal{L}^a\left(w+\epsilon\right) \leq \max_{\|\epsilon\|\leq\rho} \mathcal{L}^b\left(w+\epsilon\right)$$
$$\overset{\approx}{\Rightarrow} \mathbb{E}\left[\mathcal{L}^a_\rho\right] \leq \mathbb{E}\left[\mathcal{L}^b_\rho\right].$$

Gradient ascent path results

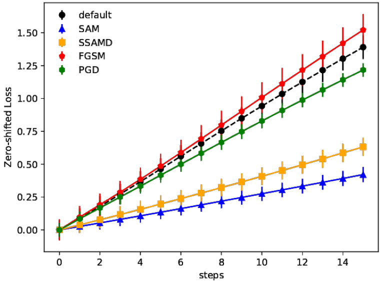


Figure: Weight space perturbation

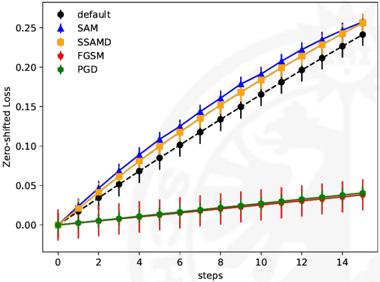


Figure: Feature space perturbation

- Adversarial training respectively reduces loss sharpness in their own spaces.
- Loss sharpness reduction in one space doesn't imply the same in the other.

Hessian analysis

Taylor expansion of perturbed loss landscape

$$\mathcal{L}(w+\epsilon) = \mathcal{L}(w) + \underbrace{\nabla\mathcal{L}(w)^T\epsilon}_{=0,\text{ local minimum}} + \frac{1}{2}\epsilon^T H_{\mathcal{L}}(w)\epsilon + \mathcal{O}(\epsilon^3),$$

- Hessian determinant and eigenvalues measure the Gaussian curvature of the loss landscape.
- Hessian matrix $(H_{\mathcal{L}})_{ij} = \partial_{w_i}\left(\partial_{w_j}\mathcal{L}\right)$ can be computed through n backpropagation steps.
- Computation is expensive, especially in weight space ($n \gg 1$).
- Reduction of parameter space. (Only classification layer and only top 10 constituents)
- Von Mises (Power iteration) method for eigenvalues computation.

Hessian analysis results

Table: Largest Hessian eigenvalues. Lower values correlate with wider minimas.

Methods	**Weight-space**		**Feature-space**	
	Hbb	**QCD**	**Hbb**	**QCD**
Default	0.31 ± 0.05	0.28 ± 0.07	0.84 ± 0.08	0.03 ± 0.01
SAM	$\mathbf{0.11 \pm 0.01}$	$\mathbf{0.12 \pm 0.01}$	0.82 ± 0.11	0.07 ± 0.04
SSAMD	0.22 ± 0.01	0.19 ± 0.03	0.98 ± 0.09	0.04 ± 0.01
FGSM	0.80 ± 0.09	0.49 ± 0.07	0.17 ± 0.01	0.024 ± 0.006
PGD	0.72 ± 0.07	0.42 ± 0.08	$\mathbf{0.056 \pm 0.004}$	$\mathbf{0.005 \pm 0.002}$

- Adversarial training respectively reduces hessian eigenvalues in their own spaces.
- PGD significantly outperforms FGSM in feature space.

Results

Fractional improvement score ΔS

$$\Delta S = \frac{S'^{b}_{a} - S^{b}_{a}}{S^{b}_{b} - S^{b}_{a}},$$

where S^j_i: score of default model trained on i and evaluated on j
S': score of considered method.

Table: Fractional generalization performance ΔS increase

Evaluation cases	SAM	SSAMD	FGSM	PGD
Pythia $\to$ Herwig	0.44 ± 0.02	$\mathbf{0.47 \pm 0.01}$	0.21 ± 0.01	0.46 ± 0.02
Herwig $\to$ Pythia	0.20 ± 0.01	0.23 ± 0.01	0.44 ± 0.02	$\mathbf{0.76 \pm 0.03}$

- Adversarial training significantly improves generalization performance.
- PGD boosts generalization performance the most.

Conclusion

- Monte-Carlo does not cross-generalize well.
- Highlighted the importance of sharpness in generalization.
- Reviewed adversarial attacks methods in the context of jet tagging.
- Introduced new sharpness analysis methods.
- Demonstrated that adversarial training significantly improves generalization performance.

Backup: Jet and constituents features list

Jet Features

Feature	Description
$\log p_T$	Logarithm of the jet transverse momentum
$\log m$	Logarithm of the jet mass

Particle Constituents Features

Feature	Description
$\log p_T$	Logarithm of the transverse momentum
$\log E$	Logarithm of the energy
$\Delta\eta$	Pseudorapidity difference relative to the jet
$\Delta\phi$	Azimuthal angle difference relative to the jet
ΔR	Distance from the from the jet axis in the $\eta - \phi$ plane
charge	Charge of the particle
$\tanh d_0$	Hyperbolic tangent of the transverse impact parameter
$\tanh dz$	Hyperbolic tangent of the longitudinal impact parameter
isPhoton	Binary indicator of whether the particle is a photon
isMuon	Binary indicator of whether the particle is a muon
isElectron	Binary indicator of whether the particle is an electron
isCH	Binary indicator of whether the particle is a charged hadron
isNH	Binary indicator of whether the particle is a neutral hadron