ML-Uncertainties and Bayesian Networks

Tilman Plehn

Universität Heidelberg

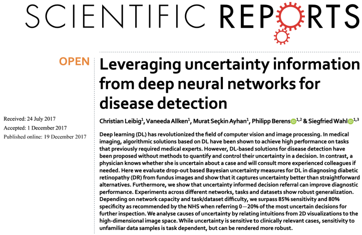
Stuttgart 7/2022

Neural networks and uncertainties

Neural networks

- nothing but numerically evaluated functions

  regression $x \to f(x)$
  classification $x \to p(x) \in [0, 1]$
  generation $x \to p_X(x)$ with sampled $x \sim \mathcal{N}$
- constructed through minimization of loss function
- Error bars making us scientists $x \to f(x) \pm \Delta f(x)$?



SCIENTIFIC REPORTS

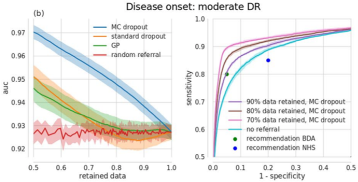
OPEN

Leveraging uncertainty information from deep neural networks for disease detection

Received: 24 July 2017
Accepted: 1 December 2017
Published online: 19 December 2017

Christian Leibig[1], Vaneeda Allken[1], Murat Seçkin Ayhan[1], Philipp Berens[1,2] & Siegfried Wahl[1,3]

Deep learning (DL) has revolutionized the field of computer vision and image processing. In medical imaging, algorithmic solutions based on DL have been shown to achieve high performance on tasks that previously required medical experts. However, DL-based solutions for disease detection have been proposed without methods to quantify and control their uncertainty in a decision. In contrast, a physician knows whether she is uncertain about a case and will consult more experienced colleagues if needed. Here we evaluate drop-out based Bayesian uncertainty measures for DL in diagnosing diabetic retinopathy (DR) from fundus images and show that it captures uncertainty better than straightforward alternatives. Furthermore, we show that uncertainty informed decision referral can improve diagnostic performance. Experiments across different networks, tasks and datasets show robust generalization. Depending on network capacity and task/dataset difficulty, we surpass 85% sensitivity and 80% specificity as recommended by the NHS when referring 0—20% of the most uncertain decisions for further inspection. We analyse causes of uncertainty by relating intuitions from 2D visualizations to the high-dimensional image space. While uncertainty is sensitive to clinically relevant cases, sensitivity to unfamiliar data samples is task dependent, but can be rendered more robust.



Uncertainties

Kinds of uncertainties

- statistical uncertainties [Poisson, Gauss, vanishing for large stats]
- systematic uncertainties [nuisance parameter]

  reference measurement elsewhere [Gauss, transferred statistical uncertainty]
  detector efficiency [distribution from simulations]
  unknown stuff [distribution unknown]
- theory: nuisance parameter

  no frequentist interpretation
  no transformation invariance, range [$\sigma \to 1/\sigma \to \log \sigma$]
- reduction of exclusive likelihood

  Bayesian: integrate out nuisance parameter
  likelihood/frequentist: profile over nuisance parameter

Uncertainties

Kinds of uncertainties

- statistical uncertainties [Poisson, Gauss, vanishing for large stats]
- systematic uncertainties [nuisance parameter]

  reference measurement elsewhere [Gauss, transferred statistical uncertainty]
  detector efficiency [distribution from simulations]
  unknown stuff [distribution unknown]
- theory: nuisance parameter

  no frequentist interpretation
  no transformation invariance, range [$\sigma \to 1/\sigma \to \log \sigma$]

NN with uncertainties

- regression: p_T of jet from constituents, error bar?
  classification: probability of Higgs event, error bar?
  generation: phase space density for large $\not{p}_T$, error bar?
- standard LHC approach

  train black box on Monte Carlo
  calibrate with reference data

$\to$ Try to do better...

A tale of four theses

David MacKay (1991)

- Bayesian methods [posterior=likelihood*prior/evidence]

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

- Bayesian networks for inference
  data modelling through parameters w

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)}$$

- Occam factor for model evidence [posterior/prior volume]
- technically: Gaussian weight distributions?

for Adaptive Models

Thesis by

David J.C. MacKay

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

(Submitted December 10, 1991)

Since the 1960's, the Bayesian minority has been steadily growing, especially in the fields of economics [89] and pattern processing [20]. At this time, the state of the art for the problem of speech recognition is a Bayesian technique (Hidden Markov Models), and the best image reconstruction algorithms are also based on Bayesian probability theory (Maximum Entropy), but Bayesian methods are still viewed with mistrust by the orthodox statistics community; the framework for model comparison is especially poorly known, even to most people who call themselves Bayesians. This thesis therefore takes some time to thoroughly review the flavour of Bayesianism that I am using. To some, the word Bayesian denotes

A tale of four theses

David MacKay (1991)

- Bayesian methods [posterior=likelihood*prior/evidence]

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

- Bayesian networks for inference
  data modelling through parameters w

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)}$$

- technically: Gaussian weight distributions?

Chapter 3

A Practical Bayesian Framework for Backpropagation Networks

Abstract

A quantitative and practical Bayesian framework is described for learning of mappings in feedforward networks. The framework makes possible: (1) objective comparisons between solutions using alternative network architectures; (2) objective stopping rules for network pruning or growing procedures; (3) objective choice of magnitude and type of weight decay terms or additive regularisers (for penalising large weights, etc.); (4) a measure of the effective number of well–determined parameters in a model; (5) quantified estimates of the error bars on network parameters and on network output; (6) objective comparisons with alternative learning and interpolation models such as splines and radial basis functions. The Bayesian 'evidence' automatically embodies 'Occam's razor', penalising over–flexible and over–complex models. The Bayesian approach helps detect poor underlying assumptions in learning models. For learning models well matched to a problem, a good correlation between generalisation ability and the Bayesian evidence is obtained.

Thesis by

David J.C. MacKay

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

(Submitted December 10, 1991)

A tale of four theses

David MacKay (1991)

- Bayesian methods [posterior=likelihood*prior/evidence]

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)}$$

- Bayesian networks for inference
  data modelling through parameters w

$$P(w|D,M) = \frac{P(D|w,M)P(w|M)}{P(D|M)}$$

- technically: Gaussian weight distributions?

Radford Neal (1995)

- deep Bayesian networks [regression, classification]
- beyond Gaussian approximation
- hybrid Monte Carlo sampling
- technically: avoid overtraining for large BNNs

$\rightarrow$ Deep BNNs for inference

BAYESIAN LEARNING FOR NEURAL NETWORKS

by

Radford M. Neal

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy,
Graduate Department of Computer Science,
in the University of Toronto

A tale of four theses

Yarin Gal (2016)

- deep learning and uncertainties
- active learning/reinforcement learning
- technically: variational inference
- technically: stochastic regularization

$\rightarrow$ BNNs for uncertainty

UNIVERSITY OF CAMBRIDGE

Uncertainty in Deep Learning

Yarin Gal

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Gonville and Caius College — September 2016

Other situations that can lead to uncertainty include

- noisy data (our observed labels might be noisy, for example as a result of measurement imprecision, leading to *aleatoric uncertainty*),
- *uncertainty in model parameters* that best explain the observed data (a large number of possible models might be able to explain a given dataset, in which case we might be uncertain which model parameters to choose to predict with),
- and *structure uncertainty* (what model structure should we use? how do we specify our model to extrapolate / interpolate well?).

The latter two uncertainties can be grouped under *model uncertainty* (also referred to as *epistemic uncertainty*). Aleatoric uncertainty and epistemic uncertainty can then be used to induce *predictive uncertainty*, the confidence we have in a prediction.

A tale of four theses

Yarin Gal (2016)

- deep learning and uncertainties
- active learning/reinforcement learning
- technically: variational inference
- technically: stochastic regularization

$\rightarrow$ BNNs for uncertainty

> But fitting the posterior over the weights of a Bayesian NN with a unimodal approximating distribution does not mean the predictive distribution would be unimodal! imagine for simplicity that the intermediate feature output from the first layer is a unimodal distribution (a uniform for example) and let's say, for the sake of argument, that the layers following that are modelled with delta distributions (or Gaussians with very small variances). Given enough follow-up layers we can capture any function to arbitrary precision—including the inverse cumulative distribution function (CDF) of any multimodal distribution. Passing our uniform output from the first layer through the rest of the layers—in effect transforming the uniform with this inverse CDF—would give a multimodal predictive distribution.

UNIVERSITY OF CAMBRIDGE

Uncertainty in Deep Learning

Yarin Gal

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Gonville and Caius College

September 2016

A tale of four theses

Yarin Gal (2016)

- deep learning and uncertainties
- active learning/reinforcement learning
- technically: variational inference
- technically: stochastic regularization

→ BNNs for uncertainty

Manuel Haußmann (2021)

- many proper derivations
- active learning, reinforcement learning
- stochastic differential equations
- technically: BNN variational inference

INAUGURAL – DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich-Mathematischen Gesamtfakultät

der

RUPRECHT-KARLS-UNIVERSITÄT
HEIDELBERG

vorgelegt von

Manuel Haußmann, M.Sc.

geboren in Stuttgart, Deutschland

Jet regression

Jet properties with uncertainties

- train many networks
  different architectures/hyperparameters
  different trainings
  different initalizations
  different data sets
- histogram network output $f(x)$, use $f(x) \pm \Delta f(x)$
- remember NN function $f_\omega(x)$ described by weights ω

$\to$ Bayesian network $\quad \Delta f_\omega(x)$ from $\Delta \omega_j$

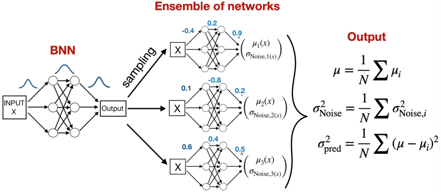
Energy measurement for jet j

- expectation value from probability distribution

$$\langle E \rangle = \int dE \; E \; p(E)$$

- Bayesian network
  sample weight distributions $p(\omega|T)$

$$p(E) = \int d\omega \; p(E|\omega) \; p(\omega|T)$$



Likelihood loss

Replacing the MSE

- start from variational approximation [think $q(\omega)$ as Gaussian with mean and width]

$$p(E) = \int d\omega \; p(E|\omega) \; p(\omega|T) \approx \int d\omega \; p(E|\omega) \; q(\omega)$$

- similarity through minimal KL-divergence [Bayes' theorem to remove unknown posterior]

$$\begin{aligned}
\text{KL}[q(\omega), p(\omega|T)] &= \int d\omega \; q(\omega) \; \log \frac{q(\omega)}{p(\omega|T)} \\
&= \int d\omega \; q(\omega) \; \log \frac{q(\omega) p(T)}{p(T|\omega) p(\omega)} \\
&= \text{KL}[q(\omega), p(\omega)] - \int d\omega \; q(\omega) \; \log p(T|\omega) + \log p(T) \int d\omega \; q(\omega) \\
&= \text{KL}[q(\omega), p(\omega)] - \int d\omega \; q(\omega) \; \log p(T|\omega) + \log p(T)
\end{aligned}$$

- well-defined evidence lower bound (ELBO)

$$\begin{aligned}
\log p(T) &= \text{KL}[q(\omega), p(\omega|T)] - \text{KL}[q(\omega), p(\omega)] + \int d\omega \; q(\omega) \; \log p(T|\omega) \\
&\geq \int d\omega \; q(\omega) \; \log p(T|\omega) - \text{KL}[q(\omega), p(\omega)]
\end{aligned}$$

$\to$ loss with likelihood $p(T|\omega)$ and prior $p(\omega)$

$$L = -\int d\omega \; q(\omega) \; \log p(T|\omega) + \text{KL}[q(\omega), p(\omega)]$$

Link to standard networks

Regularization and dropout

- Gaussian prior

$$\text{KL}[q_{\mu,\sigma}(\omega), p_{\mu,\sigma}(\omega)] = \frac{\sigma_q^2 - \sigma_p^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \log \frac{\sigma_p}{\sigma_q}$$

- deterministic network $q(\omega) \to \delta(\omega - \omega_0)$

$$L \approx - \log p(T|\omega_0) + \frac{(\mu_p - \omega_0)^2}{2\sigma_p^2} + \text{const}$$

  standard network with fixed L2-regularization

$\to$ deterministic counterpart

- Monte-Carlo dropout

  meant to reduce overfitting
  remove random weights during training
  loss with Bernoulli distribution [weight $x\omega_0 = 0, \omega_0$]

$$L = - \int dx \, \left[\rho^x (1-\rho)^{1-x}\right]_{x=0,1} \log p(T|x\omega_0) \approx -\rho \, \log p(T|\omega_0)$$

$\to$ trivial version of variational training

Weight sampling

Weight space

- expectation value using trained network $q(\omega)$

$$\begin{aligned} \langle E \rangle &= \int dE d\omega \; E \; p(E|\omega) \; q(\omega) \\ &\equiv \int d\omega \; q(\omega) \overline{E}(\omega) \qquad \text{with} \qquad \overline{E}(\omega) = \int dE \; E \; p(E|\omega) \end{aligned}$$

- output variance

$$\begin{aligned} \sigma_{\text{tot}}^2 &= \int dE d\omega \; \left(E - \langle E \rangle \right)^2 \; p(E|\omega) \; q(\omega) \\ &= \int d\omega \; q(\omega) \left[\overline{E^2}(\omega) - 2 \langle E \rangle \overline{E}(\omega) + \langle E \rangle^2 \right] \\ &= \int d\omega \; q(\omega) \left[\overline{E^2}(\omega) - \overline{E}(\omega)^2 + \left(\overline{E}(\omega) - \langle E \rangle \right)^2 \right] \equiv \sigma_{\text{stoch}}^2 + \sigma_{\text{pred}}^2 \end{aligned}$$

Two uncertainties

- contribution vanishing for $q(\omega) \to \delta(\omega - \omega_0)$

$$\sigma_{\text{pred}}^2 = \int d\omega \; q(\omega) \left[\overline{E}(\omega) - \langle E \rangle \right]^2$$

- contribution in weight space

$$\sigma_{\text{stoch}}^2 \equiv \sigma_{\text{model}}^2 = \int d\omega \; q(\omega) \left[\overline{E^2}(\omega) - \overline{E}(\omega)^2 \right] = \int d\omega \; q(\omega) \; \sigma_{\text{stoch}}(\omega)^2$$

Implementation

Approximations and implementation

- network output in weight and phase space

$$\text{BNN}\ :\ x, \omega \to \begin{pmatrix} \overline{E}(\omega) \\ \sigma_{\text{stoch}}(\omega) \end{pmatrix}$$

- Gaussian weights & likelihood

$$L = \int d\omega\ q_{\mu,\sigma}(\omega)\ \sum_{\text{jets } j} \left[\frac{\left|\overline{E}_j(\omega) - E_j^{\text{truth}}\right|^2}{2\sigma_{\text{stoch},j}(\omega)^2} + \log \sigma_{\text{stoch},j}(\omega) \right]$$
$$+ \frac{\sigma_q^2 - \sigma_p^2 + (\mu_q - \mu_p)^2}{2\sigma_p^2} + \log \frac{\sigma_p}{\sigma_q}$$

- heterostedastic loss, deterministic network

$$L = \sum_{\text{jets } j} \left[\frac{\left|\overline{E}_j(\omega_0) - E_j^{\text{truth}}\right|^2}{2\sigma_{\text{stoch},j}(\omega_0)^2} + \log \sigma_{\text{stoch},j}(\omega_0) \right]$$

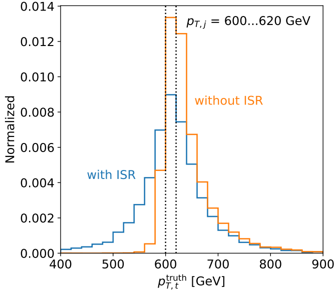
- supervised uncertainties

  training statistics
  stochastic training data
  systematics from data
  label augmentations
  model limitations

Jet measurements with error bars

Measure $p_{T,t}$ of hadronically decaying top [Kasieczka, Luchmann, Otterpohl, TP]

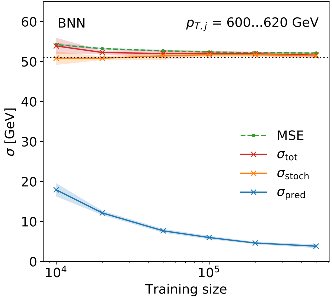
- BNN regression $p_{T,t}$
  p_T of (fat) jet decent estimate for $p_{T,t}^\text{truth}$
- non-Gaussian truth label
  symmetric in ISR-jet 'QCD heat bath'
  without ISR jets need for correction

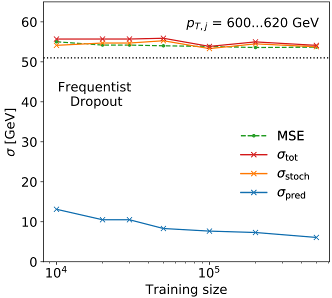


Jet measurements with error bars

Measure $p_{T,t}$ of hadronically decaying top [Kasieczka, Luchmann, Otterpohl, TP]

- BNN regression $p_{T,t}$
  p_T of (fat) jet decent estimate for $p_{T,t}^{\text{truth}}$
- non-Gaussian truth label
  symmetric in ISR-jet 'QCD heat bath'
  without ISR jets need for correction
- training sample size
  separate $\sigma_{\text{stoch}} \gg \sigma_{\text{pred}}$
  statistics not the problem [LHC theme]
  noisy label inherent limitation
  checked with deterministic networks

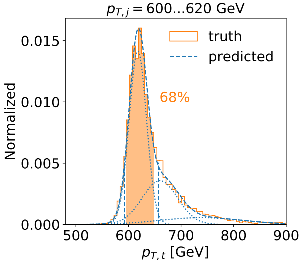




Jet measurements with error bars

Measure $p_{T,t}$ of hadronically decaying top [Kasieczka, Luchmann, Otterpohl, TP]

- BNN regression $p_{T,t}$
  p_T of (fat) jet decent estimate for $p_{T,t}^{\text{truth}}$
- non-Gaussian truth label
  symmetric in ISR-jet 'QCD heat bath'
  without ISR jets need for correction
- training sample size
  separate $\sigma_{\text{stoch}} \gg \sigma_{\text{pred}}$
  statistics not the problem [LHC theme]
  noisy label inherent limitation
  checked with deterministic networks
- non-Gaussian network output
  remember $p_{T,t}^{\text{truth}}$ non-Gaussian
  model $p(T|\omega)$ as Gaussian mixture
  weight distribution $q(\omega)$ still Gaussian



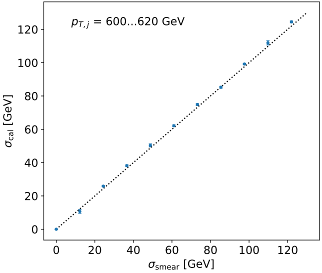
Data augmentation

Calibration means error propagation

- calibration means label measured elsewhere
- training on smeared data?
  training with smeared labels!
- Gaussian noise over label
- added to the stochastic uncertainty

$$\sigma^2_{\text{tot}} = \sigma^2_{\text{stoch}} + \sigma^2_{\text{pred}} \\ = \sigma^2_{\text{stoch},0} + \sigma^2_{\text{cal}} + \sigma^2_{\text{pred}}$$

$\rightarrow$ error extracted correctly



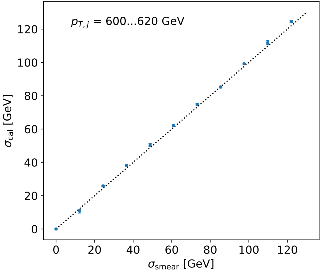
Data augmentation

Calibration means error propagation

- calibration means label measured elsewhere
- training on smeared data?
  training with smeared labels!
- Gaussian noise over label
- added to the stochastic uncertainty

$$\begin{aligned}\sigma_{\text{tot}}^2 &= \sigma_{\text{stoch}}^2 + \sigma_{\text{pred}}^2 \\ &= \sigma_{\text{stoch},0}^2 + \sigma_{\text{cal}}^2 + \sigma_{\text{pred}}^2\end{aligned}$$

$\rightarrow$ error extracted correctly



Jet regression bottom lines

- BNN regressionion working
- statistical uncertainty controlled
- stochastic uncertainty sizeable
- non-Gaussian output working
- training-data augmentation
- calibration straighforward

Precision amplitudes

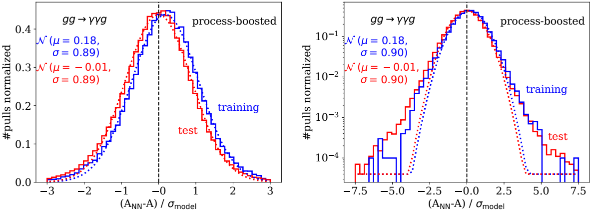
Loop amplitudes $gg \to \gamma\gamma g(g)$ [Badger, Butter, Luchmann, Pitz, TP]

- amplitudes A over phase space points x_j — simple regression
- weight-dependent pull

$$\frac{\overline{A}_j(\omega) - A_j^{\text{truth}}}{\sigma_{\text{model},j}(\omega)}$$

- training data exact in x and A
- improvement $\to$ interpolation by weighting [by pull or σ]

$$L = \int d\omega \; q_{\mu,\sigma}(\omega) \sum_{\text{points } j} n_j \times \left[\frac{\left| \overline{A}_j(\omega) - A_j^{\text{truth}} \right|^2}{2\sigma_{\text{model},j}(\omega)^2} + \log \sigma_{\text{model},j}(\omega) \right] \cdots$$



Precision amplitudes

Loop amplitudes $gg \to \gamma\gamma g(g)$ [Badger, Butter, Luchmann, Pitz, TP]

- amplitudes A over phase space points x_j — simple regression
- weight-dependent pull

$$\frac{\overline{A}_j(\omega) - A_j^{\text{truth}}}{\sigma_{\text{model},j}(\omega)}$$

- training data exact in x and A
- improvement $\to$ interpolation by weighting [by pull or σ]

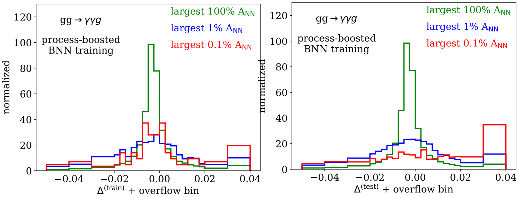
$$L = \int d\omega\ q_{\mu,\sigma}(\omega) \sum_{\text{points } j} n_j \times \left[\frac{\left|\overline{A}_j(\omega) - A_j^{\text{truth}}\right|^2}{2\sigma_{\text{model},j}(\omega)^2} + \log \sigma_{\text{model},j}(\omega) \right] \cdots$$

Precision regression

- quality of network amplitudes

$$\Delta_j^{(\text{train/test})} = \frac{\langle A \rangle_j - A_j^{\text{train/test}}}{A_j^{\text{train/test}}}$$

$\to$ Beyond fit-like regression



Precision amplitudes

Loop amplitudes $gg \to \gamma\gamma g(g)$ [Badger, Butter, Luchmann, Pitz, TP]

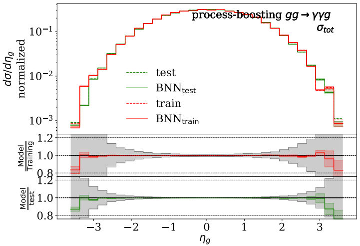
- amplitudes A over phase space points x_j — simple regression
- weight-dependent pull
$$\frac{\overline{A}_j(\omega) - A_j^{\text{truth}}}{\sigma_{\text{model},j}(\omega)}$$
- training data exact in x and A
- improvement $\to$ interpolation by weighting [by pull or σ]

$$L = \int d\omega \; q_{\mu,\sigma}(\omega) \sum_{\text{points } j} n_j \times \left[\frac{\left| \overline{A}_j(\omega) - A_j^{\text{truth}} \right|^2}{2\sigma_{\text{model},j}(\omega)^2} + \log \sigma_{\text{model},j}(\omega) \right] \cdots$$

Precision regression

- quality of network amplitudes
$$\Delta_j^{(\text{train/test})} = \frac{\langle A \rangle_j - A_j^{\text{train/test}}}{A_j^{\text{train/test}}}$$

$\to$ Beyond fit-like regression



Classification problem

SciPost Physics Submission

The Machine Learning Landscape of Top Taggers

G. Kasieczka (ed)[1], T. Plehn (ed)[2], A. Butter[2], K. Cranmer[3], D. Debnath[4], B. M. Dillon[5], M. Fairbairn[6], D. A. Faroughy[5], W. Fedorko[7], C. Gay[7], L. Gouskos[8], J. F. Kamenik[5,9], P. T. Komiske[10], S. Leiss[1], A. Lister[7], S. Macaluso[3,4], E. M. Metodiev[10], L. Moore[11], B. Nachman,[12,13], K. Nordström[14,15], J. Pearkes[7], H. Qu[8], Y. Rath[16], M. Rieger[16], D. Shih[4], J. M. Thompson[2], and S. Varma[6]

1 Institut für Experimentalphysik, Universität Hamburg, Germany
2 Institut für Theoretische Physik, Universität Heidelberg, Germany
3 Center for Cosmology and Particle Physics and Center for Data Science, NYU, USA
4 NHECT, Dept. of Physics and Astronomy, Rutgers, The State University of NJ, USA
5 Jozef Stefan Institute, Ljubljana, Slovenia
6 Theoretical Particle Physics and Cosmology, King's College London, United Kingdom
7 Department of Physics and Astronomy, The University of British Columbia, Canada
8 Department of Physics, University of California, Santa Barbara, USA
9 Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia
10 Center for Theoretical Physics, MIT, Cambridge, USA
11 CP3, Universitéxx Catholique de Louvain, Louvain-la-Neuve, Belgium
12 Physics Division, Lawrence Berkeley National Laboratory, Berkeley, USA
13 Simons Inst. for the Theory of Computing, University of California, Berkeley, USA
14 National Institute for Subatomic Physics (NIKHEF), Amsterdam, Netherlands
15 LPTHE, CNRS & Sorbonne Université, Paris, France
16 III. Physics Institute A, RWTH Aachen University, Germany

gregor.kasieczka@uni-hamburg.de
plehn@uni-heidelberg.de

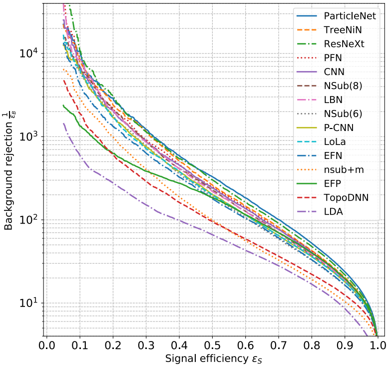
July 24, 2019

Abstract

Based on the established task of identifying boosted, hadronically decaying top quarks, we compare a wide range of modern machine learning approaches. Unlike most established methods they rely on low-level input, for instance calorimeter output. While their network architectures are vastly different, their performance is comparatively similar. In general, we find that these new approaches are extremely powerful and great fun.

'Hello world' of LHC-ML

Content



Classification problem

Top tagging with uncertainties [Bollweg, Hausßmann, Kasiecka, Luchmann, TP, Thompson]

- (60±??)% top vs gluon probability
- Bayesian classification network

$$p(c) = \int d\omega \; p(c|\omega) \; p(\omega|T)$$
$$\approx \int d\omega \; p(c|\omega) \; q(\omega)$$

- advantage: parton content not stochastic
  complication: output in closed interval [0, 1]

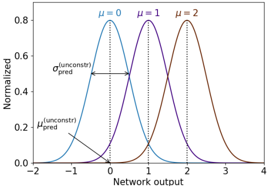
$$\text{Sigmoid}(x) = \frac{e^x}{1+e^x} \quad \Leftrightarrow \quad \text{Sigmoid}^{-1}(x) = \log \frac{x}{1-x}$$

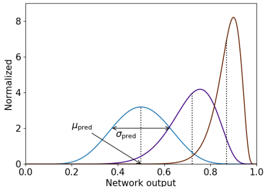
- Gaussian to classification output

$$\mu_\text{pred} = \int_{-\infty}^{\infty} d\omega \; \text{Sigmoid}(\omega) \; G_{\mu,\sigma}(\omega)$$
$$= \int_0^1 dx \; \frac{x}{x(1-x)} \; G_{\mu,\sigma}\left(\log \frac{x}{1-x}\right) \in [0,1]$$

→ correlation σ_pred vs μ_pred

$$\sigma_\text{pred} \approx \mu_\text{pred} \left(1-\mu_\text{pred}\right) \; \sigma_\text{pred}^\text{Gauss}$$

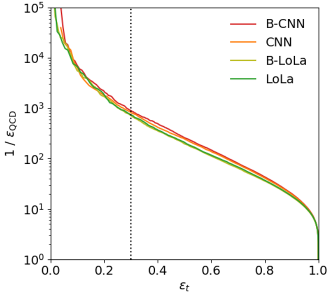




Jet classification with error bars

BNN Top tagging

- data: QCD and top jets [$p_T = 550 \ldots 600$ GeV]
  jet image [DeepTop/CNN]
  ordered constituents [LoLa]
- performance BNN vs deterministic



Jet classification with error bars

BNN Top tagging

- data: QCD and top jets [$p_T = 550 \dots 600$ GeV]
  jet image [DeepTop/CNN]
  ordered constituents [LoLa]
- performance BNN vs deterministic
- prior independence [LHC means frequentist]

σ_{prior}	10^{-2}	10^{-1}	1	10	100	1000
AUC error	0.5 —	0.9561 ± 0.0002	0.9658 ± 0.0002	0.9668 ± 0.0002	0.9669 ± 0.0002	0.9670 ± 0.0002

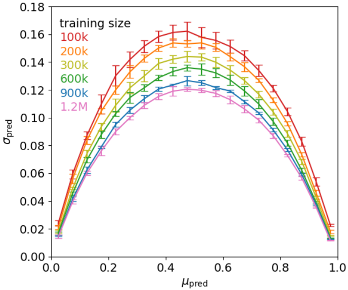
Jet classification with error bars

BNN Top tagging

- data: QCD and top jets [$p_T = 550 \ldots 600$ GeV]
  jet image [DeepTop/CNN]
  ordered constituents [LoLa]
- performance BNN vs deterministic
- prior independence [LHC means frequentist]

σ_{prior}	10^{-2}	10^{-1}	1	10	100	1000
AUC error	0.5 —	0.9561 ± 0.0002	0.9658 ± 0.0002	0.9668 ± 0.0002	0.9669 ± 0.0002	0.9670 ± 0.0002

- $\mu - \sigma$ parabola correlation



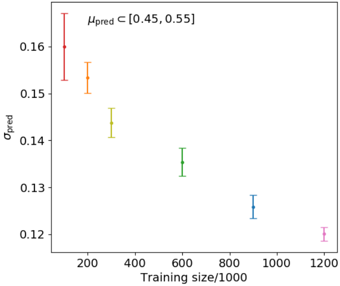
Jet classification with error bars

BNN Top tagging

- data: QCD and top jets [$p_T = 550 \dots 600$ GeV]
  jet image [DeepTop/CNN]
  ordered constituents [LoLa]
- performance BNN vs deterministic
- prior independence [LHC means frequentist]

σ_{prior}	10^{-2}	10^{-1}	1	10	100	1000
AUC	0.5	0.9561	0.9658	0.9668	0.9669	0.9670
error	—	± 0.0002	± 0.0002	± 0.0002	± 0.0002	± 0.0002

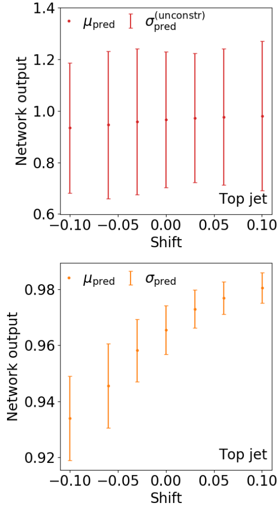
- $\mu - \sigma$ parabola correlation
- training statistics



Data augmentation

Shifted energy scale

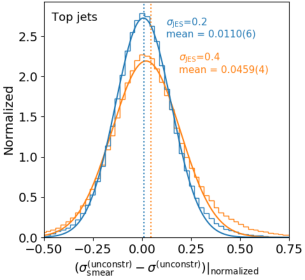
- test on augmented data [specific systematics]
  shift leading pixed by $-10\% \ldots +10\%$
  effect on σ_pred only after sigmoid
  adversarial attack [hierarchical subjets = top]



Data augmentation

Shifted energy scale

- test on augmented data [specific systematics]
  shift leading pixed by $-10\% \dots +10\%$
  effect on σ_pred only after sigmoid
  adversarial attack [hierarchical subjets = top]
- test on noisy data
  20-40% noise on constituents
  minor effect before sigmoid



Data augmentation

Shifted energy scale

- test on augmented data [specific systematics]

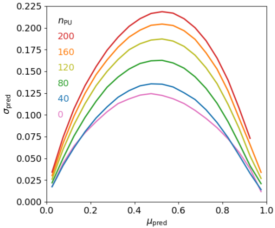
  shift leading pixed by $-10\% \ldots +10\%$
  effect on σ_pred only after sigmoid
  adversarial attack [hierarchical subjets = top]

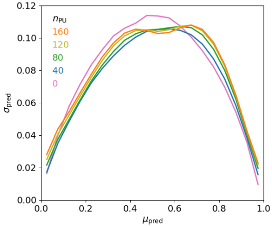
- test on noisy data

  20-40% noise on constituents
  minor effect before sigmoid

- test with noise events [pile-up]

  increased error for constituent architecture
  instability for image architecture





Data augmentation

Shifted energy scale

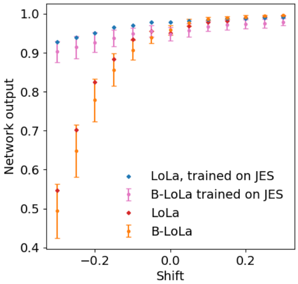
- test on augmented data [specific systematics]

  shift leading pixed by $-10\% \dots +10\%$
  effect on σ_pred only after sigmoid
  adversarial attack [hierarchical subjets = top]
- test on noisy data

  20-40% noise on constituents
  minor effect before sigmoid
- test with noise events [pile-up]

  increased error for constituent architecture
  instability for image architecture
- train on augmented data

  10% noise on constituents
  augmented training softening adversarial attack



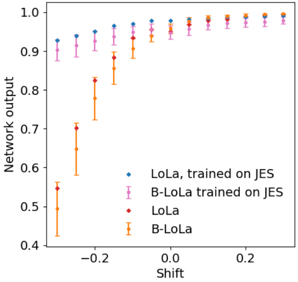
Data augmentation

Shifted energy scale

- test on augmented data [specific systematics]
  shift leading pixed by -10% ... $+10\%$
  effect on σ_{pred} only after sigmoid
  adversarial attack [hierarchical subjets = top]
- test on noisy data
  20-40% noise on constituents
  minor effect before sigmoid
- test with noise events [pile-up]
  increased error for constituent architecture
  instability for image architecture
- train on augmented data
  10% noise on constituents
  augmented training softening adversarial attack

$\rightarrow$ Jet classification bottom lines

BNN classification working
statistical uncertainy controlled
sigmoid output leading pattern
training- and test-data augmentation



Generation problem

Unsupervised Bayesian networks [Bellagente, Haußmann, Luchmann, TP]

- data: event sample [points in 2D space]

  learn phase space density
  normalizing flow mapping to latent space [INN]
  standard distribution in latent space [Gaussian]
  mapping bijective
  sample from latent space

- Bayesian version

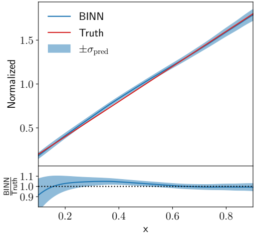
  allow weight distributions
  learn uncertainty map

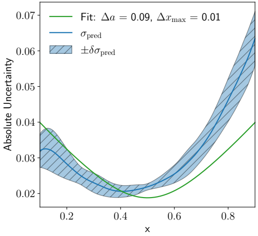
- 2D wedge ramp

$$p(x) = ax + b = ax + \frac{1 - \frac{a}{2}(x_{\max}^2 - x_{\min}^2)}{x_{\max} - x_{\min}}$$

$$(\Delta p)^2 = \left(x - \frac{1}{2}\right)^2 (\Delta a)^2 + \left(1 + \frac{a}{2}\right)^2 (\Delta x_{\max})^2 + \left(1 - \frac{a}{2}\right)^2 (\Delta x_{\min})^2$$

explaining minimum in $\sigma_{\text{pred}}(x)$





Generation problem

Unsupervised Bayesian networks [Bellagente, Haußmann, Luchmann, TP]

- data: event sample [points in 2D space]

  learn phase space density
  normalizing flow mapping to latent space [INN]
  standard distribution in latent space [Gaussian]
  mapping bijective
  sample from latent space

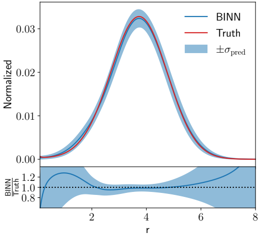
- Bayesian version

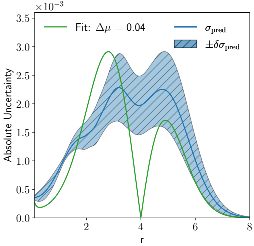
  allow weight distributions
  learn uncertainty map

- 2D wedge ramp
- kicker ramp
- Gaussian ring [$\mu = 4,\ w = 1$]

$$\Delta p = \left| \frac{G(r)}{r} \frac{\mu - r}{w^2} \right|^2 (\Delta\mu)^2 + \left| \frac{(r-\mu)^2}{w^3} - \frac{1}{w} \right|^2 (\Delta w)^2$$

explaining dip in $\sigma_{\text{pred}}(x)$





Generation problem

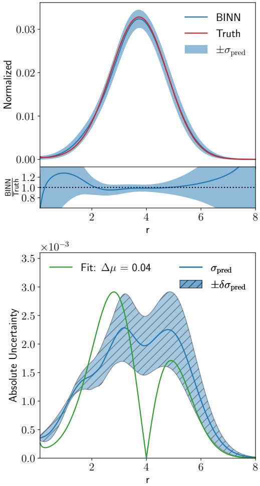
Unsupervised Bayesian networks [Bellagente, Haußmann, Luchmann, TP]

- data: event sample [points in 2D space]
  learn phase space density
  normalizing flow mapping to latent space [INN]
  standard distribution in latent space [Gaussian]
  mapping bijective
  sample from latent space
- Bayesian version
  allow weight distributions
  learn uncertainty map
- 2D wedge ramp
- kicker ramp
- Gaussian ring [$\mu = 4$, $w = 1$]

$$\Delta p = \left| \frac{G(r)}{r} \, \frac{\mu - r}{w^2} \right|^2 (\Delta\mu)^2 + \left| \frac{(r-\mu)^2}{w^3} - \frac{1}{w} \right|^2 (\Delta w)^2$$

explaining dip in $\sigma_{\text{pred}}(x)$

$\to$ INNs just (non-parametric) fits



Bayesian networks

Initially developed for inference they work for...

...regression with error bars

...classification with error bars

...generation with error bars

Modern Machine Learning in Particle Physics

Tilman Plehn, Anja Butter, Barry Dillon, and Claudius Krause

June 21, 2022

Abstract

These lectures notes should lead advanced students with basic knowledge in particle physics and some enthusiasm for machine learning to cutting-edge research in modern machine learning. They accompany a lecture in the 2022 Summer term at Heidelberg University. All examples are chosen from particle physics papers from the last few years, many of them from our Heidelberg group. This is not because our papers are the only interesting applications, but we know them best. For more background information check out Ref. [1] in its online version!