Combining ID & AD: simultaneous particle identification and anomaly detection at collider experiments using Bayesian neural networks

Wiehl 2022

Johannes Erdmann, Burim Ramosaj, **Daniel Wall**

- Bayesian vs. Detern
- Our example use d
	- Multiclass classific **interest in the EM calorimeter** images
- Classification Perfo
- Anomaly Detection
- Conclusions

Outl

Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. *3 Ê|***X** *q◊*

• Basic idea: learn the whole distribution over the NN weights w given training data {X,Y} $\frac{1}{100}$ idea: learn the whole distribution over the NN weights w given training data $\{X,Y\}$ Kult barameters *west a function f are likely to have generated* V from V ? Note that this integral is only defined when *q◊*(*Ê*) is absolutely continuous w.r.t. *p*(*Ê|***X***,* **Y**) (fion over the NN weights w given training data $\{X, Y\}$ The likely to have generated Y from X ? (i.e. for the whole distribution over the NN weights w given training data $\{X, Y\}$ the minimum of this optimisation objective (often a local minimum). m_{c} and m_{c} divergence allows us to a predictive distribution as to approximate the prediction as m_{c} Note that this integral is only defined when *q◊*(*Ê*) is absolutely continuous w.r.t. *p*(*Ê|***X***,* **Y**) (ution over the NN weights w given training data {X, Y} t the minimum of the minimum of the minimum t **Ar**
er
, to *q*ú *◊* **y**ú $\overline{\mathbf{a}}$ **i**c |
|
|rai

• Sampling from optimal $q\theta^*(w) \to$ distribution of predictions instead of a point estimate Minimising the KL divergence allows us to approximate the predictive distribution as $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx$ ⁄
∕∕∕ $p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\omega})q_\theta^*(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}$ • Sampling from optimal $q\theta^*(w)$ → distribution of predictions $(\lambda^* | A^* \mathbf{V} \mathbf{V}) = (\lambda^* | A^* \mathbf{V})$ $p(y | X, A, I) \approx$ which defines the objective we will refer to $\mathbf r$ \mathbf{I}) \sim \int $P(\mathbf{y} | \mathbf{A}, \boldsymbol{\omega}) q_{\theta}(\boldsymbol{\omega})$ u $\boldsymbol{\omega}$ $(x * x x x)$ $(x * x * y x)$ $\int f(\theta) |f(x)|^2 dx$
 $\int f'(\theta) |f(x)|^2 dx$ which defines the objective we will refer to $\mathbf r$ $\left(\mathbf{r}, \mathbf{r} \right) \sim \int P(\mathbf{y} | \mathbf{x}, \boldsymbol{\omega}) q_{\theta}(\boldsymbol{\omega}) \mathrm{d} \boldsymbol{\omega}$ **2.1.1 Variational inference** $|X$
on $\int q$
rtai) *||* \blacksquare $\overline{Y)}$
bu \approx $\int\limits_{\bf{hD}}$ *A*, \widetilde{q}_{t} *,* JU ¥ *Ê* $\begin{aligned} \n\text{in} \ \mathbf{z} \ \mathbf{z} \ \mathbf{d} \ \mathbf{z} \end{aligned}$ *,*

• Often intractable \rightarrow approximate with simpler function $q_{\theta}(w)$ minimizing KL divergence • Often intractable \rightarrow approximate with simpler function $q_{\theta}(w)$ minimizing KL divergence $\overline{}$ *Ê |***x** ú*,* $\overline{}$ *q ◊*

\n- \n**Bayesian Neural Networks**\n
	\n- Basic idea: learn the whole distribution over the NN weights w given training data {X,Y}.
	\n- Which parameters w of a function f are likely to have generated Y from X?
	\n- \n
	$$
	p(w|X, Y) = \frac{p(X, Y|w)p(w)}{p(X, Y)} = \frac{p(Y|X, w)p(X|w)p(w)}{p(Y|X)p(X)} = \frac{p(Y|X, w)p(w)}{\int p(Y|X, w)p(w)dw}
	$$
	\n with for example\n
	$$
	p(y = a|x, \omega) = \frac{\exp(f_a^{\omega}(x))}{\sum_{d'} \exp(f_a^{\omega}(x))}
	$$
	\n
	\n- \n Often intractable → approximate with simpler function qe(w) minimizing KL divergence\n
	$$
	KL(q_\theta(\omega) || p(\omega|X, Y)) = \int q_\theta(\omega) \log \frac{q_\theta(\omega)}{p(\omega|X, Y)} d\omega = -\int q_\theta(\omega) \log p(Y|X, \omega) d\omega + KL(q_\theta(\omega)||p(\omega)) + \text{const}
	$$
	\n
	\n- \n Sampling from optimal qe*(w) → distribution of predictions instead of a point estimate\n
	$$
	p(y^* | x^*, X, Y) \approx \int p(y^* | x^*, \omega) q_\theta^*(\omega) d\omega
	$$
	\n
	\n- \n**Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and a rXiv:0712.4042.**\n
	\n

\n- **Bayesian Neural Networks .** Basic idea: learn the whole distribution over the NN weights w given training data {X,Y} **.** Which parameters w of a function f are likely to have generated Y from X ?
\n- $$
p(w|X,Y) = \frac{p(X,Y|w)p(w)}{p(X,Y)} = \frac{p(Y|X,w)p(X|w)p(w)}{p(Y|X)p(X)} = \frac{p(Y|X,w)p(w)}{f(p(Y|X,w)p(w)dw}
$$
\n with for example $p(y=d|\mathbf{x},\omega) = \frac{\exp(f_q^w(\mathbf{x}))}{\sum_{d'}\exp(f_d^w(\mathbf{x}))}$ \n
\n- Often intractable → approximate with simpler function $q_{\theta}(w)$ minimizing KL divergence $KL(q_{\theta}(\omega) \parallel p(\omega|\mathbf{X}, \mathbf{Y})) = \int q_{\theta}(\omega) \log \frac{q_{\theta}(\omega)}{p(\omega|\mathbf{X}, \mathbf{Y})} d\omega = -\int q_{\theta}(\omega) \log p(\mathbf{Y}|\mathbf{X}, \omega) d\omega + KL(q_{\theta}(\omega) || p(\omega)) + const$ \n
\n- Sampling from optimal $q_{\theta}^s(w) \rightarrow$ distribution of predictions instead of a point estimate $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \int p(\mathbf{y}^*|\mathbf{x}^*, \omega) q_{\theta}^*(\omega) d\omega$ \n
\n- Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and a rXiv:0712.4042.

KL divergence minimisation is also equivalent to maximising the *evidence lower bound* n Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. $\hskip 10mm 3$ rning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. \overline{a} *Ê*) log

- **2.1.1 Variational inference** *◊*, whose structure is **JUL** \bullet
	- *Which parameters w of a function f are likely to have generated Y from X ?* $\overline{}$ \overline{h} *Ê|***X**

◊

Bayesian Neural Networks easy to evaluate our approximation to be as close as possible as possible as possible as possible as possible a
2.1 Bayesian modelling as possible as Note that this integral is only defined when *q◊*(*Ê*) is absolutely continuous w.r.t. *p*(*Ê|***X***,* **Y**) \mathbf{r}

• Basic idea: learn the whole distribution over the NN weights w given training data {X,Y} ic idea learn the whole distribution over the NN weights w given training data $\{X|Y\}$ Kultuba champetors *Wester (Kunstian fare likely to have generated* V from V ? Note that this integral is only defined when *q◊*(*Ê*) is absolutely continuous w.r.t. *p*(*Ê|***X***,* **Y**) (fion over the NN weights w given training data $\{X, Y\}$ The matrice minimum of the minimum of the minimum \mathbf{X} and \mathbf{X} are local minimum \mathbf{X} and \mathbf{X} are local minimum of the minimum o (i.e. for the whole distribution over the NN weights w given training data {X,Y} the minimum of this optimisation objective (often a local minimum). Note that this integral is only defined when *q◊*(*Ê*) is absolutely continuous w.r.t. *p*(*Ê|***X***,* **Y**) $\{X, Y\}$ is a form weights w given training data $\{X, Y\}$ t the minimum of the minimum of the minimum t and t a local minimum t a local minimum t and t \geq t \geq *Art*
er
, to ú *◊* **y**ú $\overline{\mathbf{a}}$ **:** |
|
|rai

◊ \overline{a} proximate with si

 $q_{\theta}(\omega)$ is the every measurable set $q_{\theta}(\omega)$ in the denote by α $\bm{\omega})\,\parallel p(\bm{\omega}|\mathbf{A},\,\mathbf{Y})) = \int q_\theta(\bm{\omega})\log\frac{1}{p(\bm{\omega}|\mathbf{X},\mathbf{Y})}\mathrm{d}\bm{\omega}\!=\!-\!\int q_\theta(\bm{\omega})\log p(\mathbf{Y}|\mathbf{A},\bm{\omega})\mathrm{d}\bm{\omega}+\mathbf{KL}(q_\theta)$ $\text{KL}(q_{\theta}(\boldsymbol{\omega}) \mid p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})) = \int q_{\theta}(\boldsymbol{\omega}) \log \frac{q_{\theta}(\boldsymbol{\omega})}{p(\boldsymbol{\omega})|\mathbf{X}},$ $p(\boldsymbol{\omega}|\mathbf{X}, \mathbf{Y})$ $d\boldsymbol{\omega} = -\int q_{\boldsymbol{\theta}}(\boldsymbol{\omega}) \log p(\boldsymbol{\omega})$ $\mathbf{w}(\mathbf{X},\mathbf{Y}))=\int q_\theta(\boldsymbol{\omega})\log\frac{q_\theta(\boldsymbol{\omega})}{p(\boldsymbol{\omega}|\mathbf{X},\mathbf{Y})}\mathrm{d}\boldsymbol{\omega}\!=\!-\!\int q_\theta(\boldsymbol{\omega})\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega}+\mathrm{KL}(q_\theta(\boldsymbol{\omega})||p(\boldsymbol{\omega}))\;.$ 2.1 Bayesian modelling **19** *Ê|***X Y**)) = $\overline{r})$
ig, i ω = *p* $\overline{\Omega}$ $\begin{aligned} \n\text{in} \quad \omega \\ \n\text{and} \quad \text{and} \quad \text{and}$ \overline{a}

 $p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \int$ • Sampling from optimal $q\theta$ ^{*^{*}*(*W*) → distribution of predictions} $(\lambda^* | A^* \mathbf{V} \mathbf{V}) = (\lambda^* | A^* \mathbf{V})$ $(x * x x x)$ $(x * x * y x)$ **2.1.1 Variational inference** an approximating *variational* distribution $\begin{aligned} \mathbf{p} \mathbf{p} \ \mathbf{p}$ to the posterior distribution obtained from the original model. We thus minimise the $\int q$
rtai $\frac{\mathcal{V}}{\mathcal{V}}$
intuitively a measure of similar
 \mathcal{V}^*) *||* \mathcal{L}

Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. Description from Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. *A Ê|***X** *q◊*

\n- Basic idea: learn the whole distribution over the NN weight
\n- Which parameters *w* of a function *f* are likely to have generated
\n- $$
p(w|X,Y) = \frac{p(X,Y|w)p(w)}{p(X,Y)} = \frac{p(Y|X,w)p(X|w)p(w)}{p(Y|X)p(X)}
$$
\n with for example
$$
p(y = d|x, \omega) = \frac{\exp(f_d^{\omega}(x))}{\sum_{d'} \exp(f_d^{\omega}(x))}
$$
\n
\n- Often intractable → approximate with simpler function $q_{\theta}(x)$
\n- $$
KL(q_{\theta}(\omega) \parallel p(\omega|X,Y)) = \int q_{\theta}(\omega) \log \frac{q_{\theta}(\omega)}{p(\omega|X,Y)} d\omega = -\int q_{\theta}(\omega) \log p(\omega)
$$
\n
\n- Sampling from optimal $q_{\theta}^*(w) \rightarrow$ distribution of predictions
\n

with for example to the posterior distribution obtained from the original model. We thus model $\exp(f_d^{\omega}(\mathbf{x}))$ $\sum_{d'} \exp(f_{d'}^{\omega}(\mathbf{x}))$ $y =$ $\overline{}$ *◊* $\overline{}$ d' \vee

> • Sampling from optimal $q_{\theta}^{*}(w) \rightarrow$ distribution of predictions instead of a point estimate Minimising the KL divergence allows us to approximate the predictive distribution as *A*, $\overline{\mathsf{U}}$ ¥
∗ *,Ê*

• Often intractable \rightarrow approximate with simpler function $q_{\theta}(w)$ minimizing KL divergence • Often intractable \rightarrow approximate with simpler function $q_{\theta}(w)$ minimizing KL divergence $\overline{}$ *Ê |***x** ú *,* **CONTRACTOR** *q ◊*

$$
p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) \approx \int p(\mathbf{y}^*|\mathbf{x}^*, \boldsymbol{\omega}) q_{\theta}^*(\boldsymbol{\omega}) \mathrm{d} \boldsymbol{\omega}
$$

n Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042.

KL divergence minimisation is also equivalent to maximising the *evidence lower bound* n Yarin Gal, Uncertainty in Deep Learning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. \diamond rning, PhD thesis, Cambridge, 2016 and arXiv:0712.4042. \overline{a} *Ê*) log

- **2.1.1 Variational inference** *◊*, whose structure is IU L \bullet \blacksquare
	- *Which parameters w of a function f are likely to have generated Y from X ?* \sim \sim \overline{h} *Ê|***X**

q

 $p(Y|X,w)p(X|w)p(w)$ $p(Y|X)p(X)$ = $p(Y|X,w)p(w)$ $\int p(Y|X,w)p(w)dw$ intuitively a measure of similar interventions of similar terms of similar terms of similar terms of similar t
The contract of similar terms of similar t $P(I|A)P$
 $\lim_{\alpha \to 0} P(I|A)P$ \overline{F} $\overline{\int p(Y|X,w)p(w)dw}$ The true posterior *p*(*Ê|***X***,* **Y**) cannot usually be evaluated analytically. Instead we define $M_{\rm H}$, the KL divergence allows us to approximate the prediction as μ and μ as μ *p*(**y**^ú *|***x**ú *,* **X***,* **Y**) ¥ $\frac{1}{\sqrt{2}}$ $(X, w)p(w)dw$ $\exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))$ m_{c} and m_{c} divergence allows us to a predictive distribution as to approximate the prediction as m_{c} $\overline{}$ \overline{X} $p(X)$ = $\overline{\int p(Y|X,w)p(w)dw}$ **EXAIIIPIE** $p(y = a | \mathbf{x}, \boldsymbol{\omega}) = \frac{\sum_{d'} \exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))}{\sum_{d'} \exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))}$ $M_{\rm H}$, the KL divergence allows us to approximate the prediction as μ and μ *p*(⁄ $(X,w)p(w)dw$ $\exp(f_{d'}^{\boldsymbol{\omega}}(\mathbf{x}))$ is absolutely continuous w.r.t. $\frac{1}{\sqrt{2}}$
 $\frac{1}{\sqrt{2}}$ *q ◊* ($\frac{\mathsf{he}}{\mathsf{x}}$
of v^*
is e^* \mathbf{R}
 \mathbf{R} d *Ê* =: W free
 $\frac{1}{2}$
 $\frac{1}{2}$
 $\frac{1}{2}$
 $\frac{1}{2}$
 $\frac{1}{2}$ *◊* $\overline{\mathcal{U}}$ *Ê*) , **(Contact Contact)** *Ê*)) $\overline{\mathbf{Y}}$ $\begin{array}{c} \mathbf{a} \cdot \\ \mathbf{b} \end{array}$
arged will refer

• Small training datasets lead to larger uncertainties in the BNN predictions (example from top-tagging in 1904.10004)

• BNNs predictions for out-of-distribution

Bayesian Neural Networks

Fig. 1.2 **Predictive mean and uncertainties on the Mauna Loa CO**² **concen-Y. Gal, PhD thesis, Cambridge**

1904.10004

- Classification of images in EM calorimeters = photon identification
- Main background:
	- High-energy $π⁰ → yy$
- Toy EM calorimeter à la 1712.10321
	- ATLAS-like: 3 layers of LAr+Pb
	- 1 m from Geant4 particle gun

Our Example Use Case inner most detector layer is a tracking detector is a tracking detector, in which electrically charged particles are traced particles are for example by creating electron-hole pairs in semiconductors or by ionising gas. Typically, the

- Classification of images in EM calorimeters = photon identification
- Main background:
	- High-energy $π⁰ → yy$
- Toy EM calorimeter à la 1712.10321
	- ATLAS-like: 3 layers of LAr+Pb
	- 1 m from Geant4 particle gun

Our Example Use Case inner most detector layer is a tracking detector is a tracking detector, in which electrically charged particles are traced particles are for example by creating electron-hole pairs in semiconductors or by ionising gas. Typically, the

- Single photon (signal)
- 8 physical background classes:
	- Different purely EM decays of mesons with d different masses \rightarrow different opening angles
- + noise background class (noise burst in 2nd layer with 1% cross talk to neighbouring cells)

• Particle gun kinetic energy: 20 GeV

• Photon ID algorithm at LHC would be trained on effective mixture via parton shower programs

Our Example Use Case

- Setup: 2D CNN with 8 filters and 3x3 kernel size for each calorimeter layer
	- + 10 output nodes all using Flipout (1803.04386)
-
- Assume weights to be Gaussian distributed, uncorrelated and with Gaussian priors • Very similar performance to deterministic NN

• For each image, sample from the weight distributions \rightarrow mean & variance per image

• For each image, sample from the weight distributions \rightarrow mean & variance per image

Mean Std. Dev.

• Now: remove one class at a time during training = anomaly (here: $J/\psi \rightarrow e^+e^-$)

Results

Mean Std. Dev.

- In general, larger uncertainty than before
- Stems from three different cases:
	- A) Examples with ~0 variance
	- B) Examples with >1 active output node
	- C) Examples with "jumpy decisions"

- In general, larger uncertainty than before
- Stems from three different cases:
	- A) Examples with ~0 variance
	- B) Examples with >1 active output node
	- C) Examples with "jumpy decisions"

- In general, larger uncertainty than before
- Stems from three different cases:
	- A) Examples with ~0 variance
	-
	- Examples with "jumpy decisions"

- In general, larger uncertainty than before
- Stems from three different cases:
	- A) Examples with ~0 variance
	-
	-

#Events

• Another example: noise burst as anomaly

Removed class: Noiseburst

Removed class: Noiseburst

• Standard deviations when each class is removed one-by-one

• Some anomalies are easier to identify than others

Conclusions

- Bayesian network's uncertainty estimate may help to identify anomalies
- Semisupervised approach (LHC Olympics 2020,
	-
- We don't know, yet, if it can compete with specialized AD algorithms

2101.08320)

• One advantage:

Provides AD capabilities as a sanitiy/DQ cross check

for standard classification tasks (such as ID)

• At some additional training and prediction costs

