

Bayesian Deep
Learning with
10 % of the
weights

Rob
Romijnders

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Bayesian Deep Learning with 10 % of the weights

Practical approach to Bayesian deep learning

Rob Romijnders

robromijnders.github.io

PyData Amsterdam, 2018

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

- The goal
- Historical perspective
- Bayesian inference
- Parameter posterior
- Uncertainty
- Pruning

3 Experiments and results

- Pruning
- Uncertainties

4 Closing

Problems with neural networks

Neural networks have two problems:

- ① Neural networks give no **uncertainty** in predictions
→ easily fooled by **adversarial examples**
- ② Neural networks have **millions of parameters**

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning
Uncertainties

Closing

Motivation

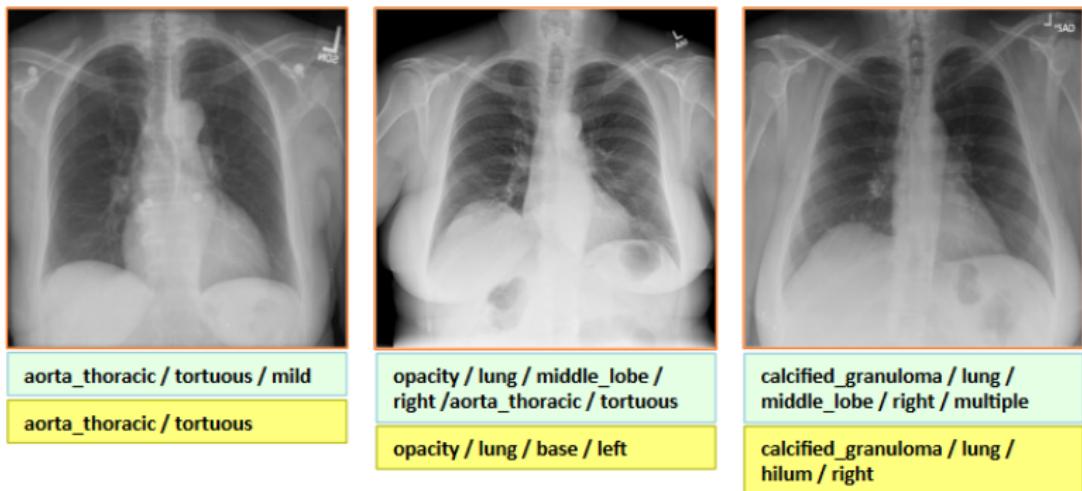


Figure: Uncertainty is important when making diagnoses

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Motivation



Figure: Uncertainty is important when making a critical decision

Motivation



Figure: Uncertainty is important when prediction bitcoin

Adversarial attack



Figure: Uncertainty is necessary to find adversarial examples

Bayesian Deep
Learning with
10 % of the
weights

Rob
Romijnders

Introduction

Motivation

Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Embedded applications



Figure: Pruning reduces the memory and computation usage (Pruning = dropping parameters)

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Real time inference



Figure: Pruning reduces the computation requirements

Bayesian Deep Learning with 10 % of the weights

Rob
Romijnders

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments and results

Pruning

Uncertainties

Closing

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Pseudo code

In summary, this talk covers the following pseudo code

```
model = Model()
```

```
model.train(data)
```

```
if application == 'embedded':  
    model.prune()
```

```
# Actually, the next line is all we care about:  
prediction, uncertainty = model.predict(input)
```

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Historical perspective

This content is not new: it has been around for decennia/centuries

Being Bayesian about neural networks

- Bayes lived in 18th century
- Variational inference for neural networks: Hinton and van Camp (1993)
- Bayesian Neural networks: Neal (1995)

Uncertainties for a model

Shannon published information theory in 1948

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Bayes rule

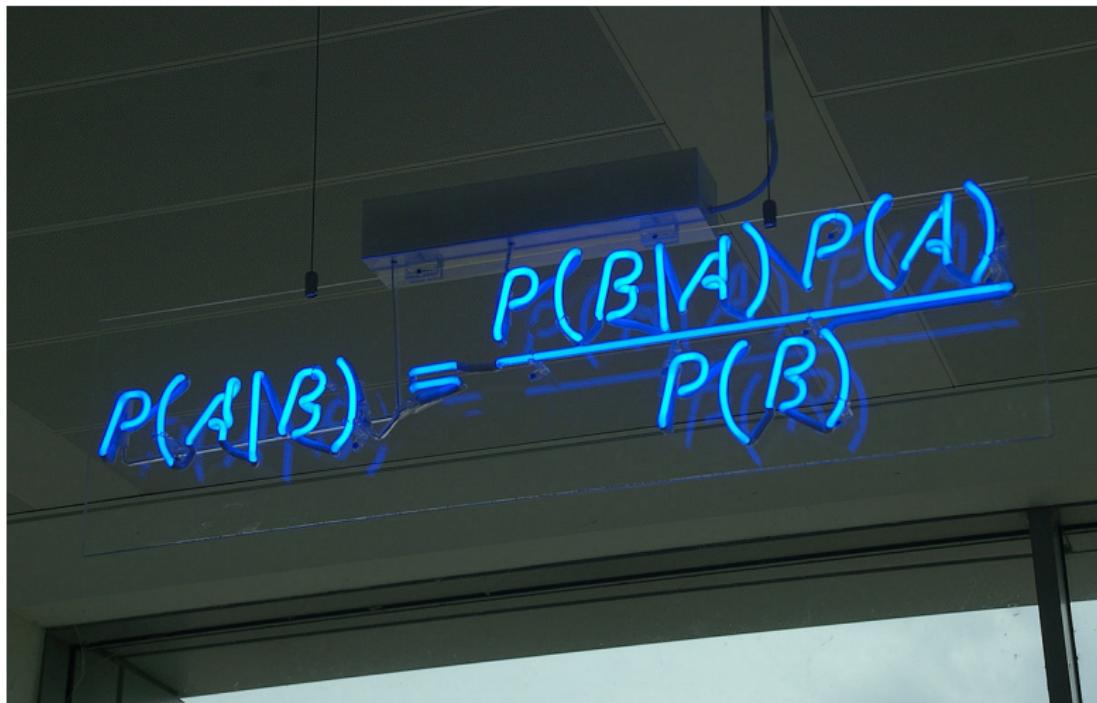


Figure: Every presentation on Bayesian machine learning has this image.
So this presentation too

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Bayes rule

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$

$$\log \text{posterior} = \log \text{likelihood} + \log \text{prior} + \text{constant}$$

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Bayes rule

We have been using Bayes' rule all the time

Weight decay .. Ridge regression .. L2 regularisation

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior
Uncertainty
Pruning

Experiments
and results

Pruning
Uncertainties

Closing

$$\begin{aligned} -\log \text{posterior} &= -\log \text{likelihood} & -\log \text{prior} & + \text{constant} \\ \text{loss} &= \text{classification loss} & + \lambda \sum_i w_i^2 & + \text{constant} \end{aligned}$$

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Stochastic gradient descent

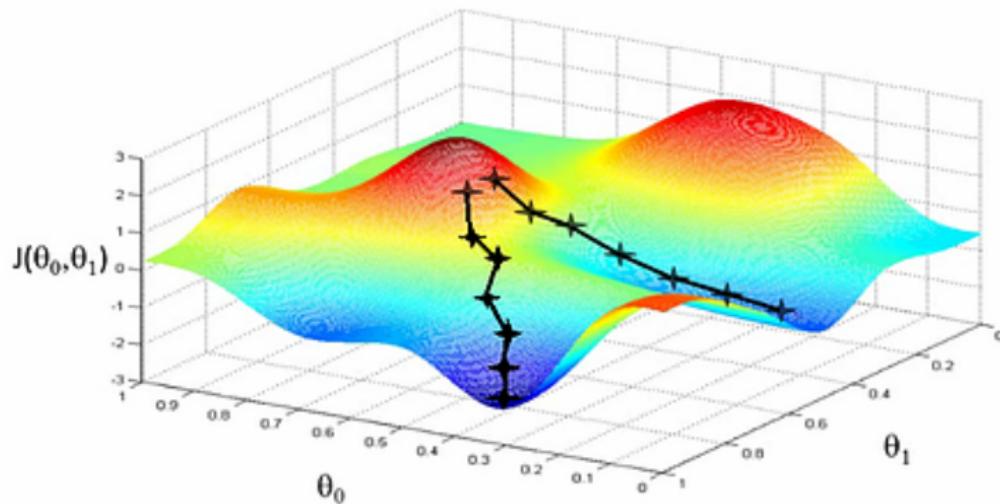


Figure: But we inferred only one parameter vector

Bayesian Deep Learning with 10 % of the weights

Rob
Romijnders

Introduction

Motivation

Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

Experiments and results

Pruning

Uncertainties

Closing

Bayesian deep learning

From this...

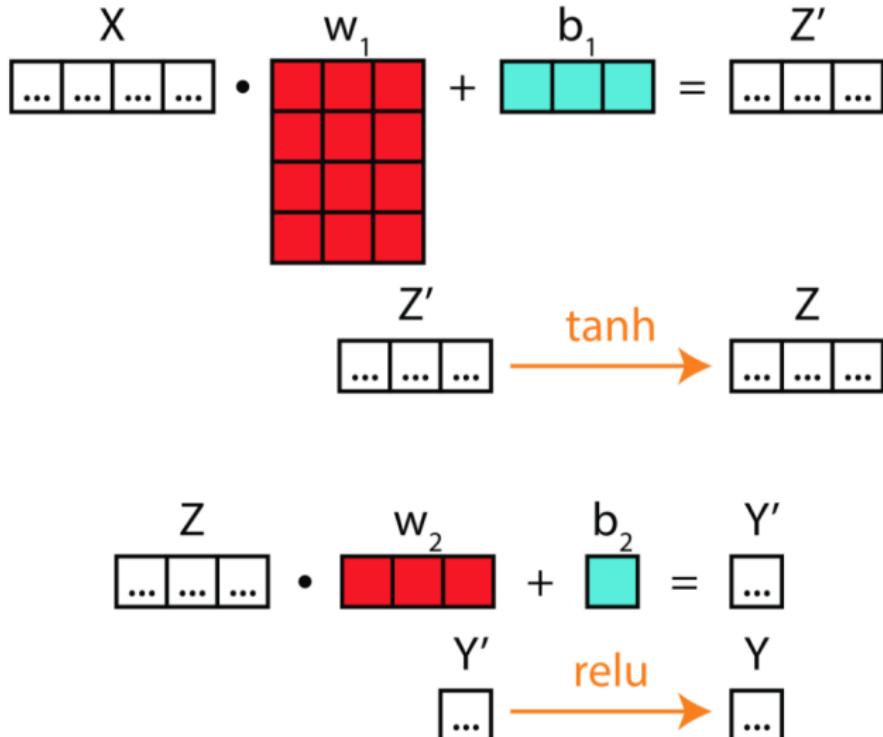


Figure: Used with kind permission of [Eric Ma](#)

...to this

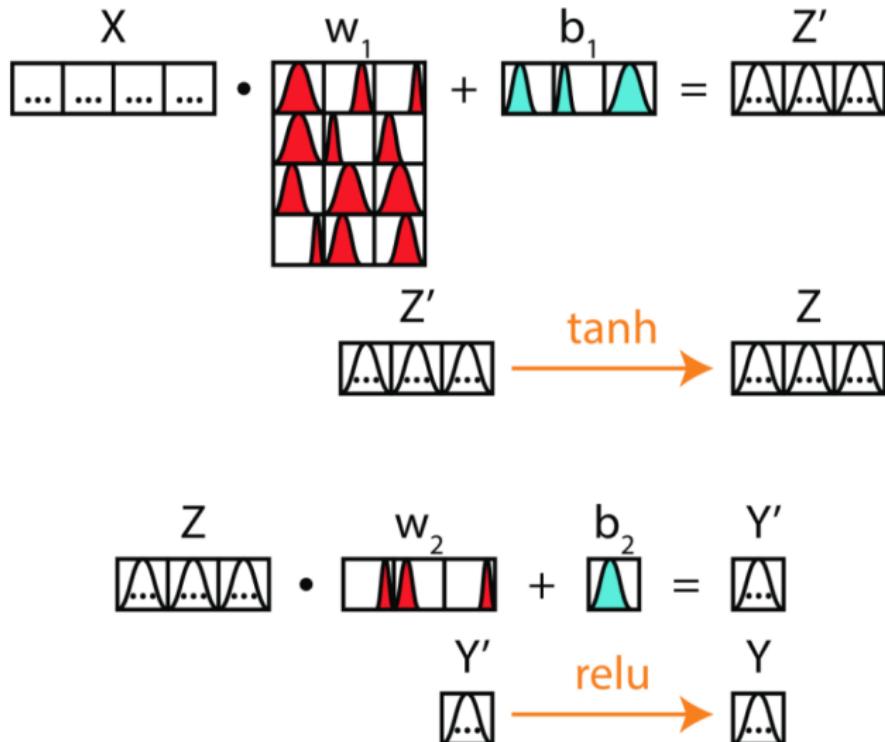


Figure: Used with kind permission of [Eric Ma](#)

Parameters of a Gaussian

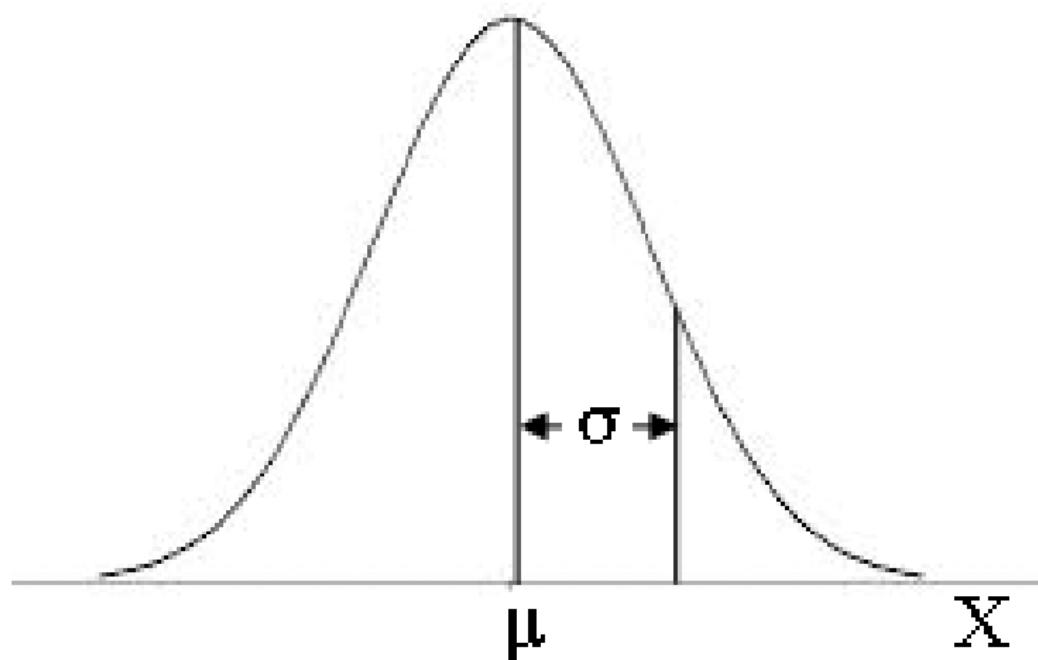


Figure: For a Gaussian, we need parameters μ and σ

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Posterior probability

The parameter posterior will:

- Enable more samples for prediction → uncertainty over prediction
- Tell us which parameters have high zero-probability → pruning

Loss functions

Loss functions

old loss

$$= \text{classification loss} + \sum_i \underbrace{\lambda w_i^2}_{\text{L2 penalty}}$$

new loss

$$= \text{classification loss} + \sum_i \underbrace{\frac{1}{2} \lambda \mu_i^2}_{\text{L2 penalty}} - \underbrace{\log \sigma_i + \frac{1}{2} \lambda \sigma_i^2}_{\text{penalty on } \sigma}$$

Intuition

loss

$$= \underbrace{\text{classification loss} + \sum_i \frac{1}{2} \lambda \mu_i^2}_{\text{loss on location of weights}} - \log \sigma_i + \underbrace{\frac{1}{2} \lambda \sigma_i^2}_{\text{loss on } \sigma}$$

Summary

- **What do we care about?**
Uncertainties and pruning
- **How we do that?**
Bayesian inference
- **How we do that?**
Approximate the parameter posterior
- **How we do that?**
Find a μ and σ per parameter
- **How we do that?**
Minimize the loss function on the previous slide

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Use entropy as uncertainty metric

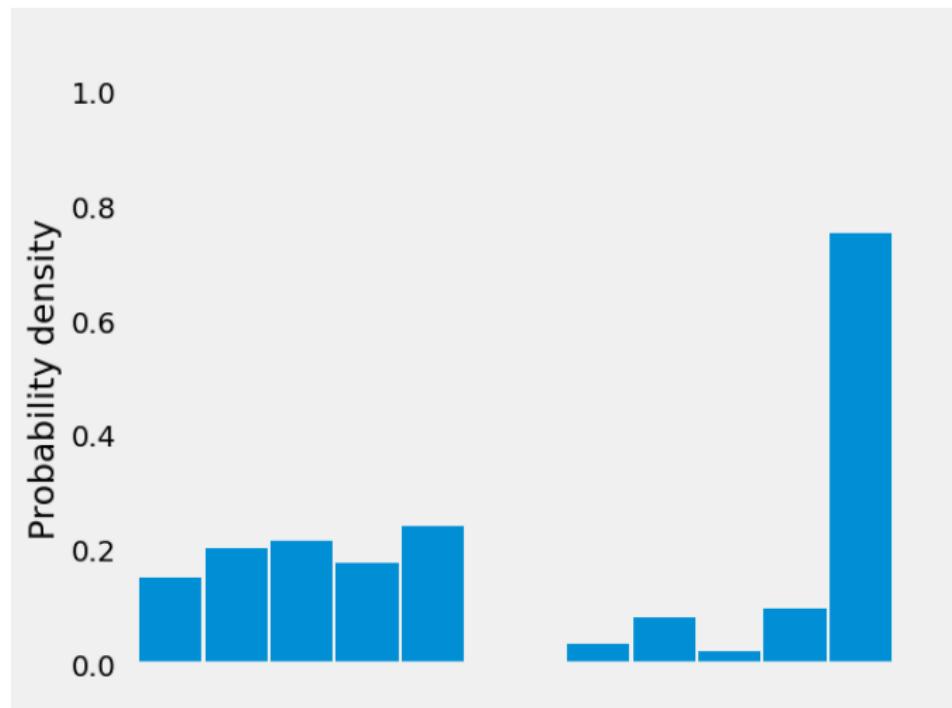


Figure: Which prediction has least uncertainty?

Use entropy as uncertainty metric

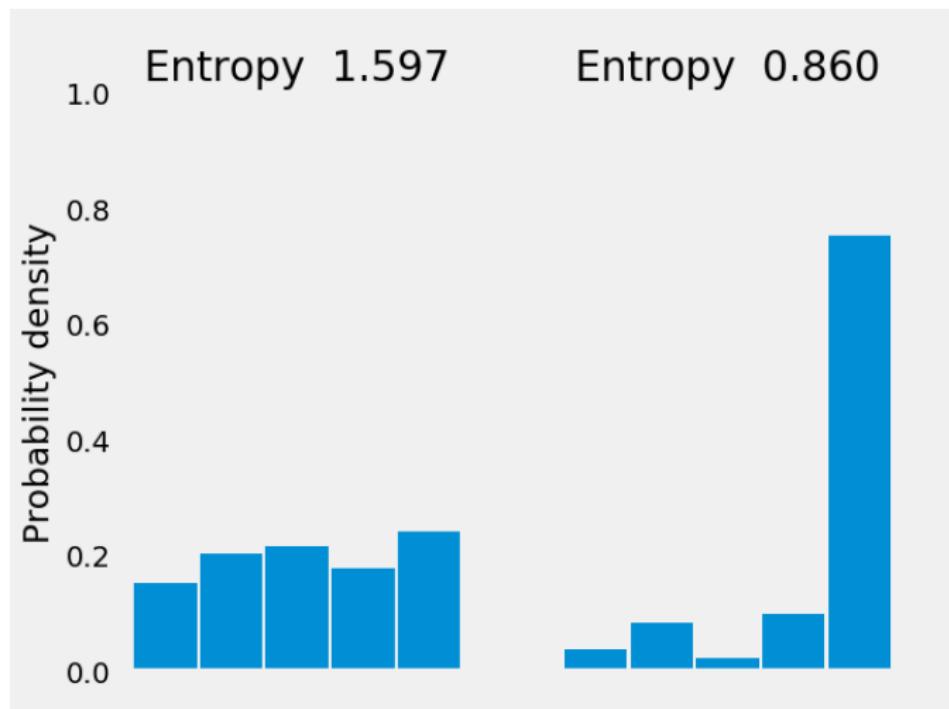


Figure: Which prediction has least uncertainty?

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

From this...

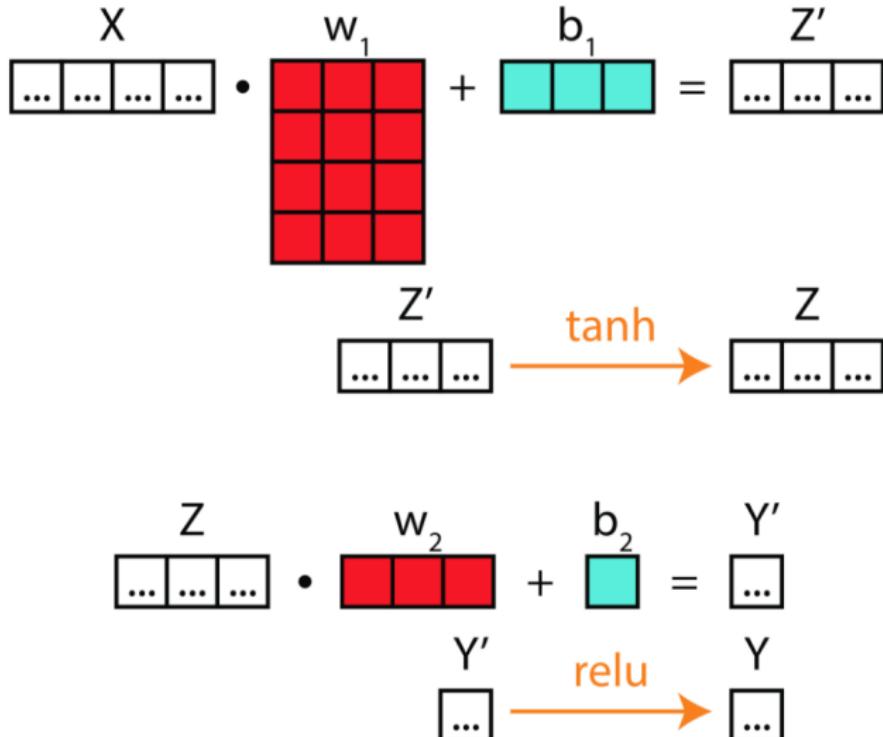


Figure: Used with kind permission of [Eric Ma](#)

...to this

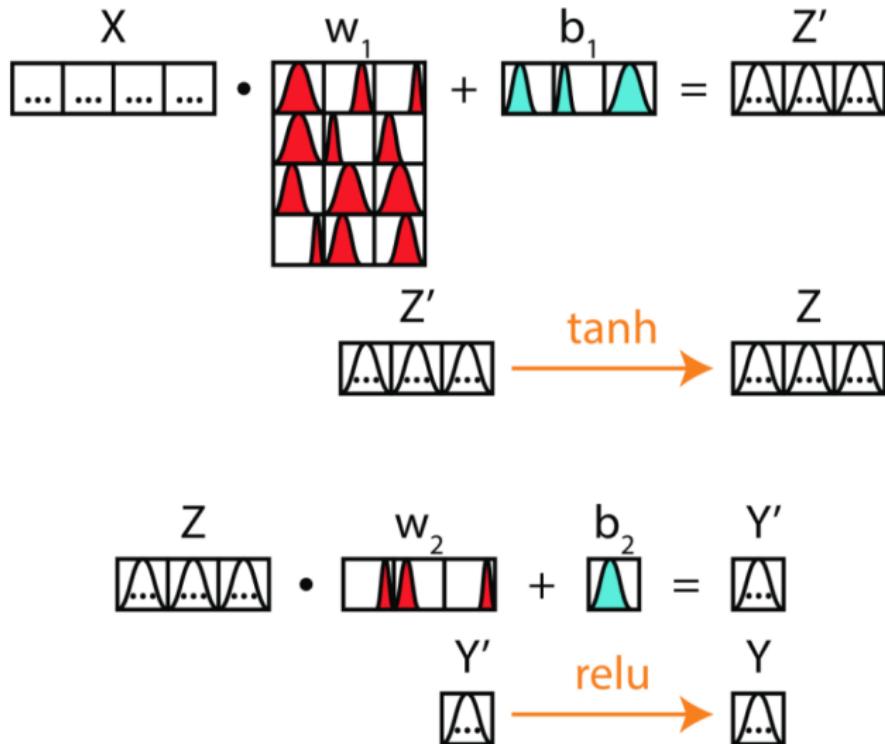


Figure: Used with kind permission of [Eric Ma](#)

From this...

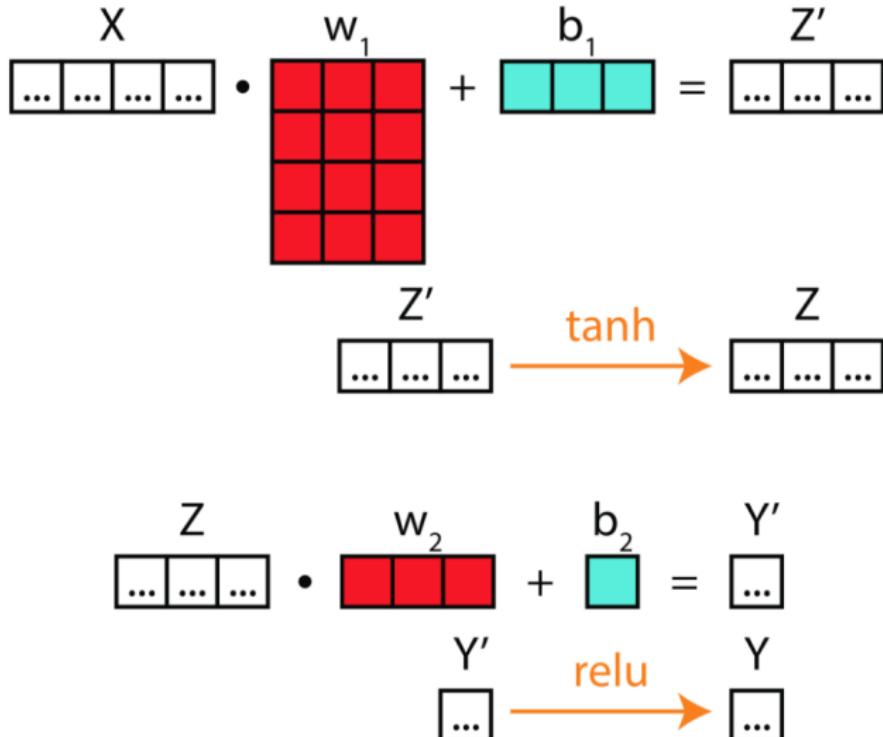


Figure: Used with kind permission of [Eric Ma](#)

From this...

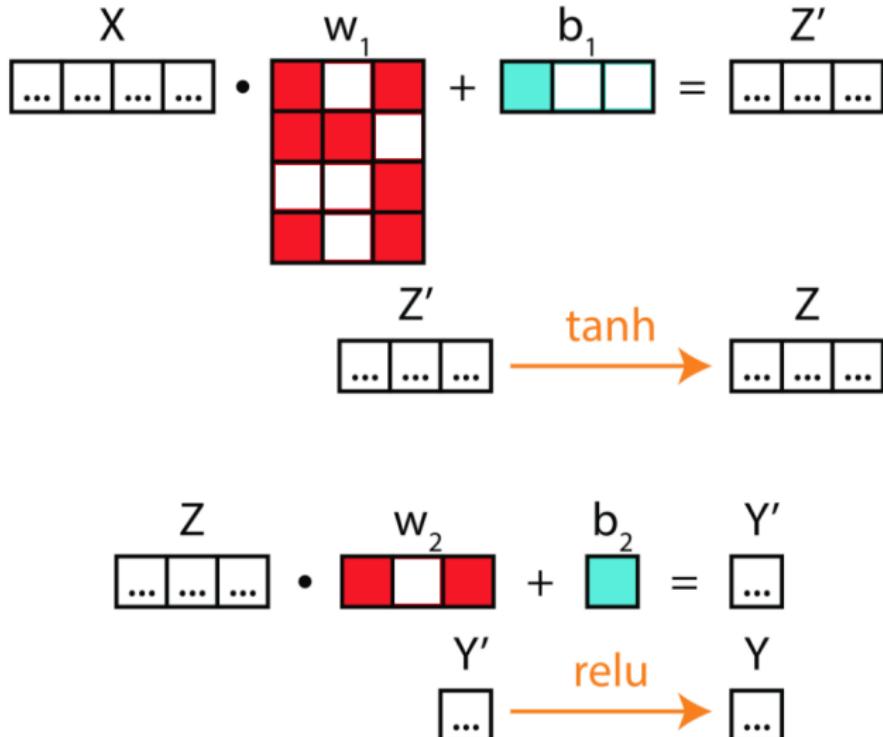


Figure: Used with kind permission of [Eric Ma](#)

Pruning according to posterior

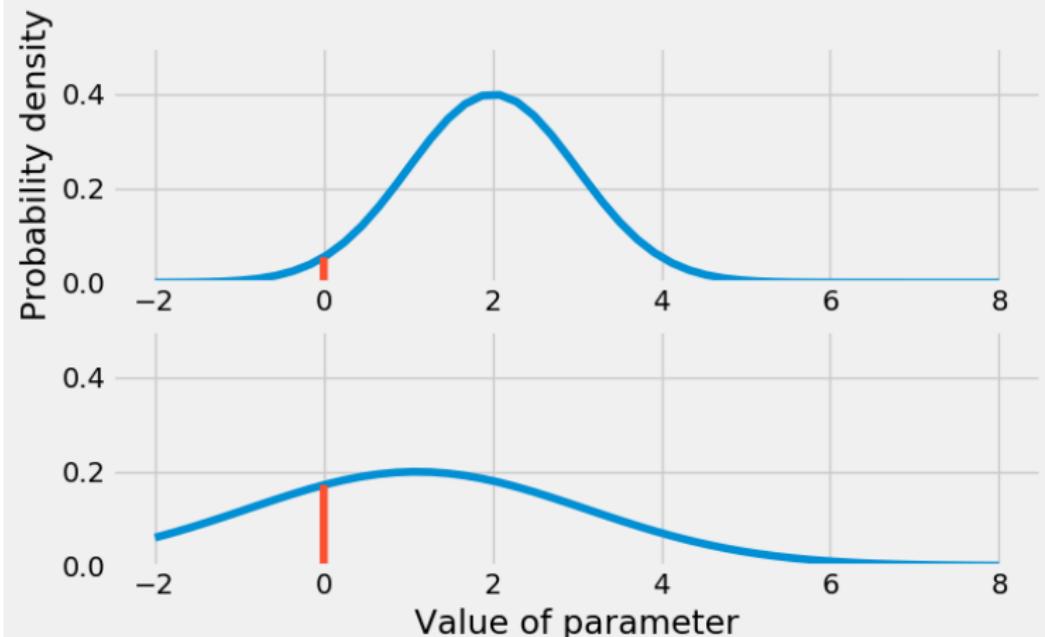


Figure: Which parameter would you rather prune?

Data sets

Fun

No deep learning project is complete without **MNIST**

Serious

Two most common applications of deep learning:

- Image recognition: **CIFAR10** data set
- Time series classification: **UCR - ECG's**
 - Train set only 500 time series → Bayesian's don't overfit

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

MNIST examples



Figure: Examples of MNIST. Train set: 50k samples. Test set: 10k samples

CIFAR examples

airplane



automobile



bird



cat



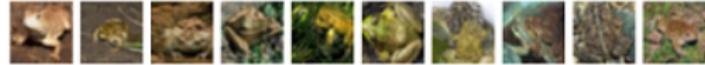
deer



dog



frog



horse



ship



truck



Figure: Examples of CIFAR. Train set: 50k samples. Test set: 10k samples

Introduction

Motivation

Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

ECG examples

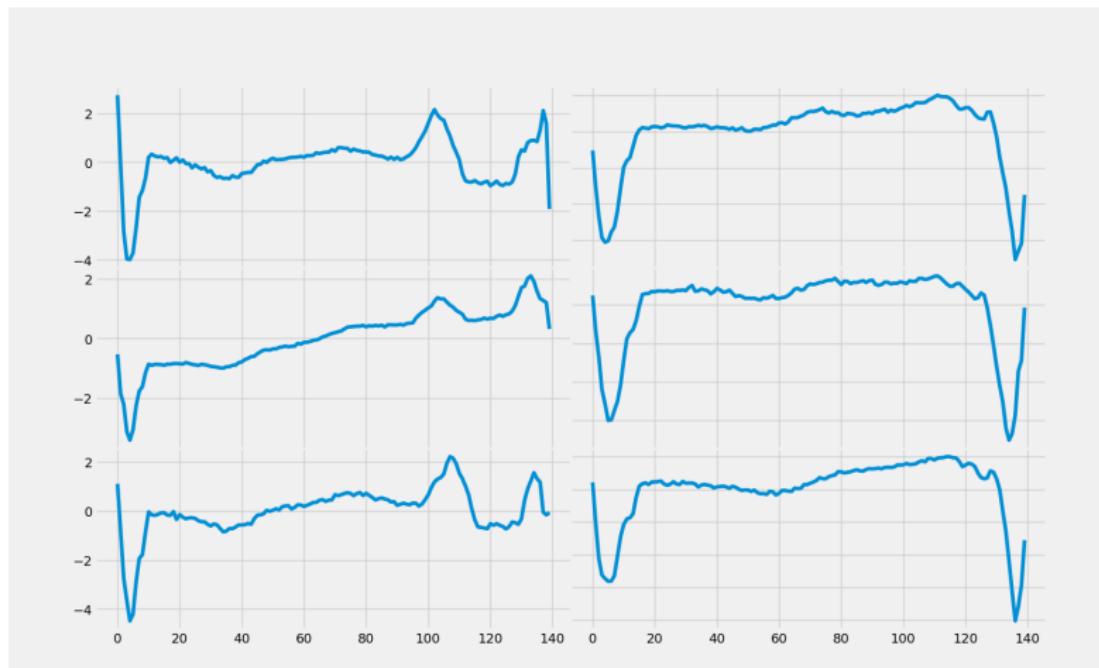


Figure: Examples of ECG. Train set: 500 samples. Test set: 4500 samples

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

1 Introduction

Motivation

2 Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

3 Experiments and results

Pruning

Uncertainties

4 Closing

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

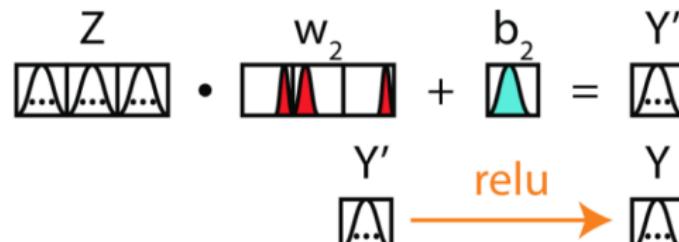
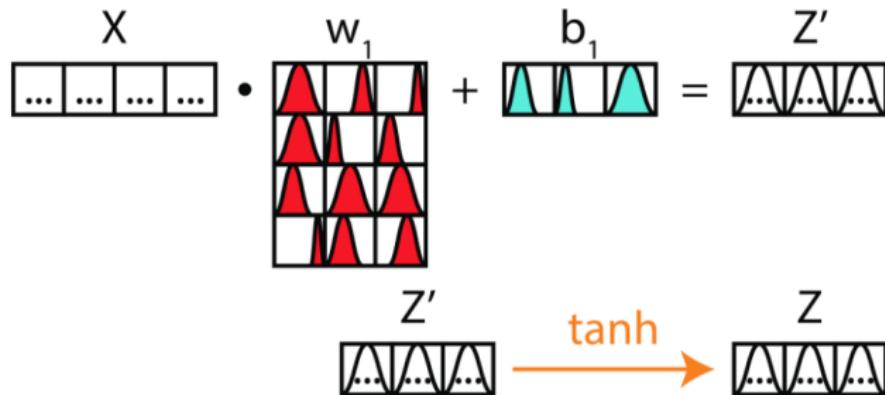
Pruning

Uncertainties

Closing

Remember the model

...to this



Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Pruning MNIST

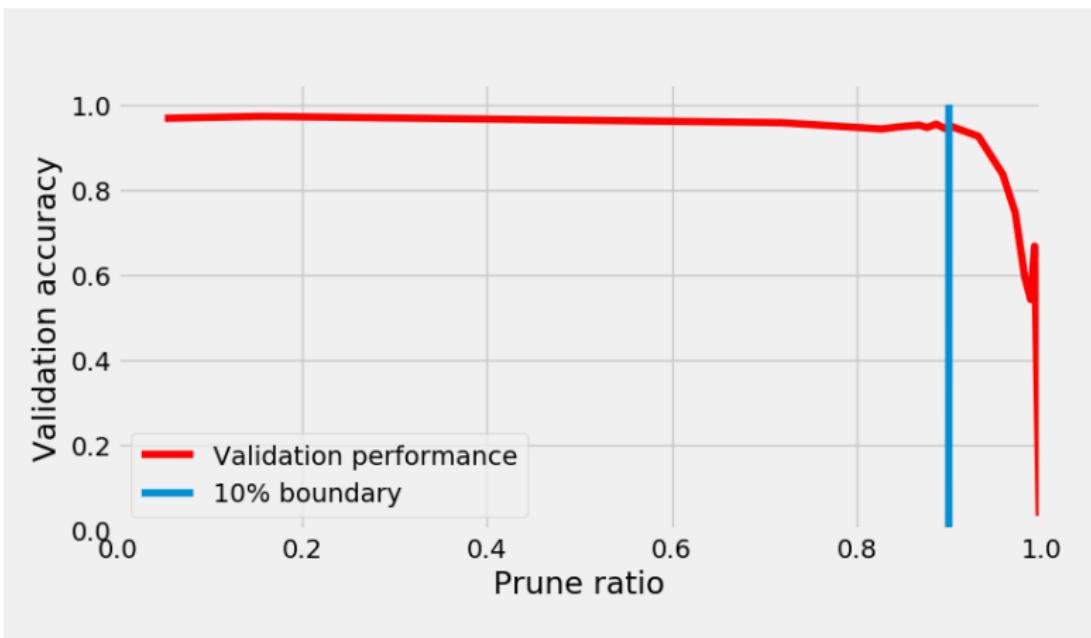


Figure: Pruning curve for MNIST

Pruning CIFAR

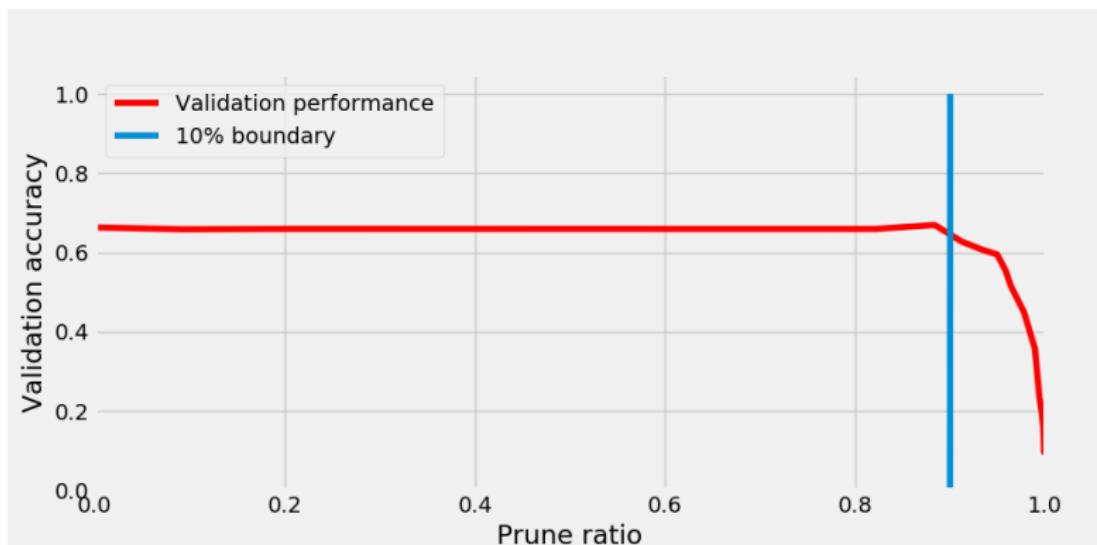


Figure: Pruning curve for CIFAR

Pruning ECG

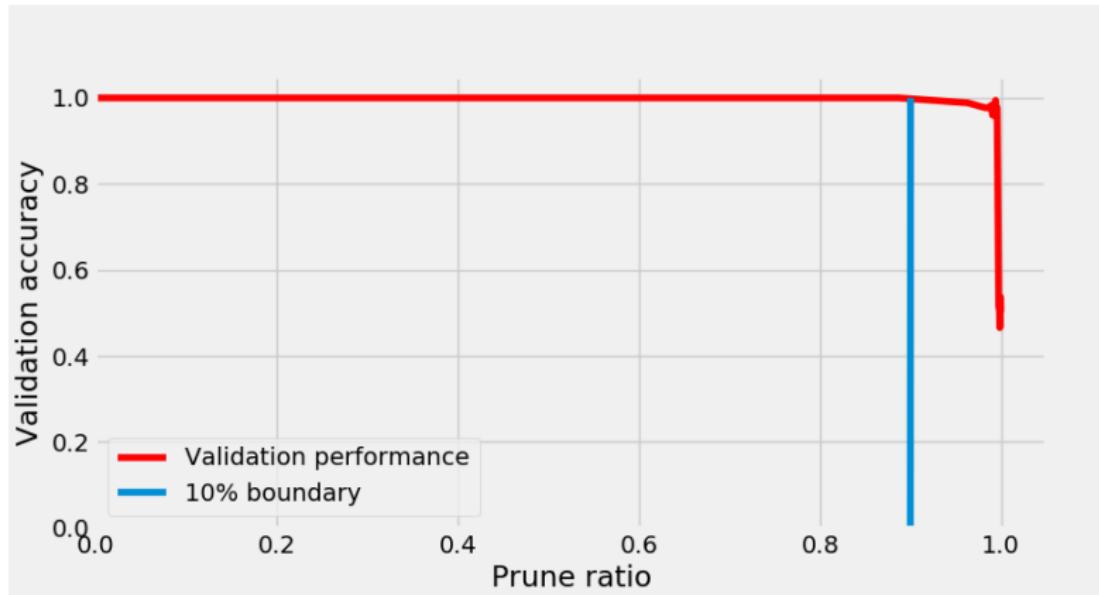


Figure: Pruning curve for ECG

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Outline

① Introduction

Motivation

② Method

The goal

Historical perspective

Bayesian inference

Parameter posterior

Uncertainty

Pruning

③ Experiments and results

Pruning

Uncertainties

④ Closing

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty
Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Experiment uncertainty

How to mutilate images to raise uncertainty?

- Add noise
- Warping
- Rotation

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Noise

Figure: animation only shows with media plugin (use adobe reader)

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Rotation

Figure: animation only shows with media plugin (use adobe reader)

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Warping

Figure: animation only shows with media plugin (use adobe reader)

Take aways

- Get uncertainty for critical predictions
- Robust against adversarial attacks
- Prune networks for small memory and small compute

Questions?

robromijnders.github.io

Material

github.com/RobRomijnders/weight_uncertainty

- All code
- Further reading
- More explanation

Bayesian Deep Learning with 10 % of the weights

Rob
Romijnders

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments and results

Pruning

Uncertainties

Closing

Additional slides

Introduction

Motivation

Method

The goal

Historical perspective
Bayesian inference

Parameter posterior

Uncertainty
Pruning

Experiments and results

Pruning

Uncertainties

Closing

Learning the sigma's

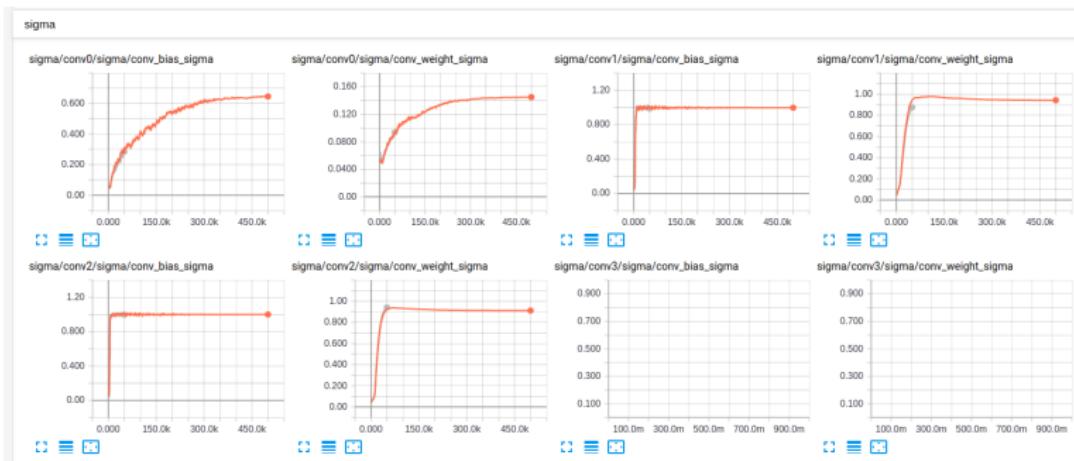
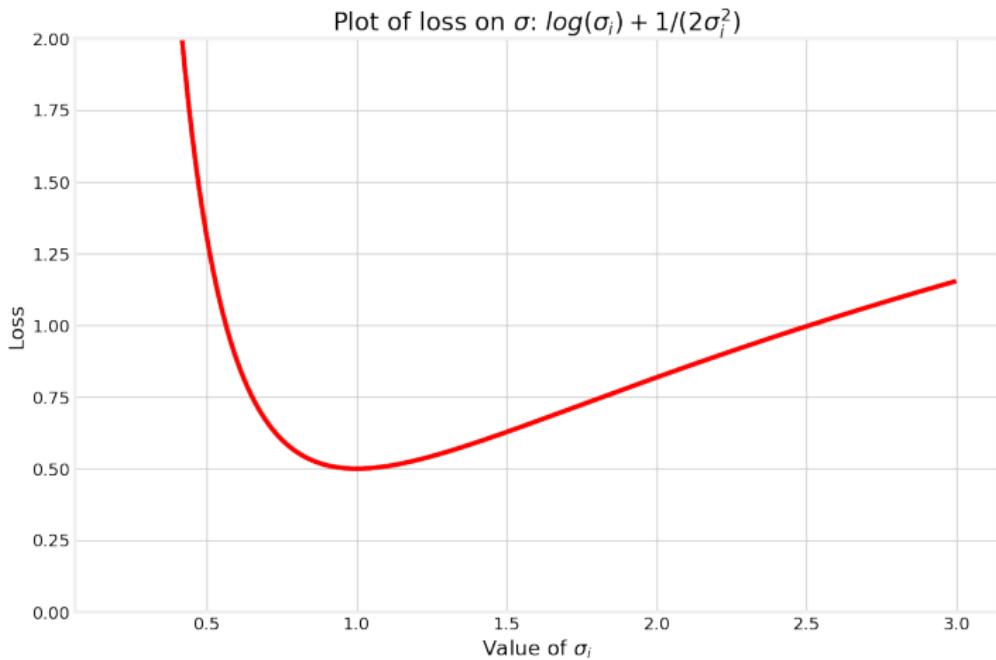


Figure: The VI objective increases the sigma's by itself!!

Loss on σ

What does the loss for σ look like?



Make predictions

Sampling

Make multiple predictions with sampled parameters. One can think of this sampling as an ensemble method

```
def make_prediction(input):  
    for param_vec in param_vecs:  
        yield model.get_output(input, param_vec)  
prediction = np.mean(make_prediction(input))
```

Pseudo code

Pseudo code for training our neural network

OLD CODE

```
while not converged:  
    # Get the loss  
    x, y = sample_batch()  
    loss = loss_function(x, y, w)
```

#Update the parameters

```
w_grad = gradient(loss, w)  
w = update(w, w_grad)
```

#####
NEW CODE

```
while not converged:  
    # Get the loss  
    x, y = sample_batch()  
    w = approximation.sample()  
    loss = loss_function(x, y, w)
```

Update the approximation

```
w_grad = gradient(loss, w)  
approximation = update(approximation, w_grad)
```

```
while not converged:
```

```
    # Get the loss
```

```
    x, y = sample_batch()
```

```
    w = approximation.sample()
```

```
    loss = loss_function(x, y, w)
```

```
# Update the approximation
```

```
w_grad = gradient(loss, w)
```

```
approximation = update(approximation, w_grad)
```

Pruning: speed

Bayesian compression for deep learning, Louizos © NIPS2017

Uncertainty: adversarial attack

Adversarial phenomenon in Bayesian deep learning, Rawat, 2017

Gaussian approximation

Approximate with a normal distribution

- Captures local structure of the posterior, which indicates the uncertainty
- Simple for parameter pruning

Anything is better than point estimation !!!

Gaussian approximation

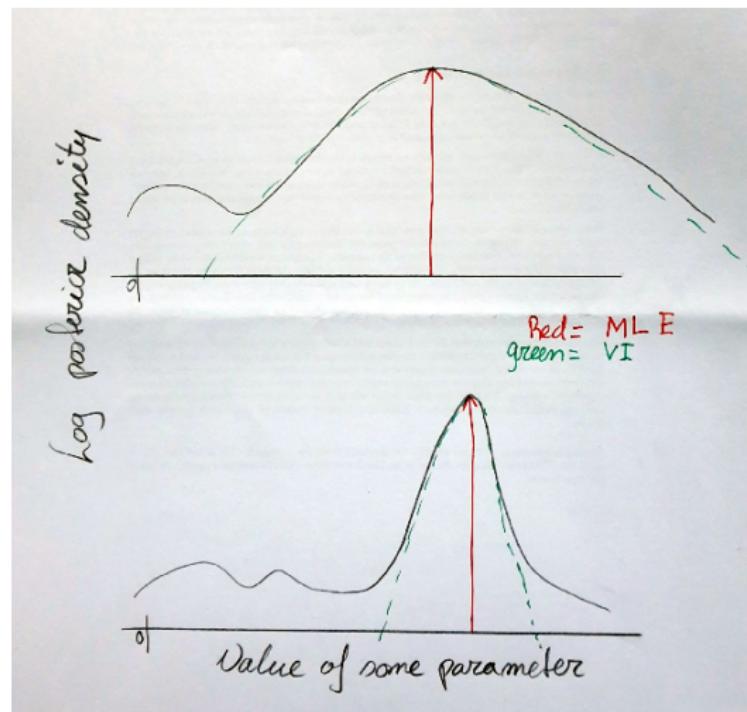


Figure: Any approximation (VI) is better than just a point estimate (MLE)

Bernoulli approximations

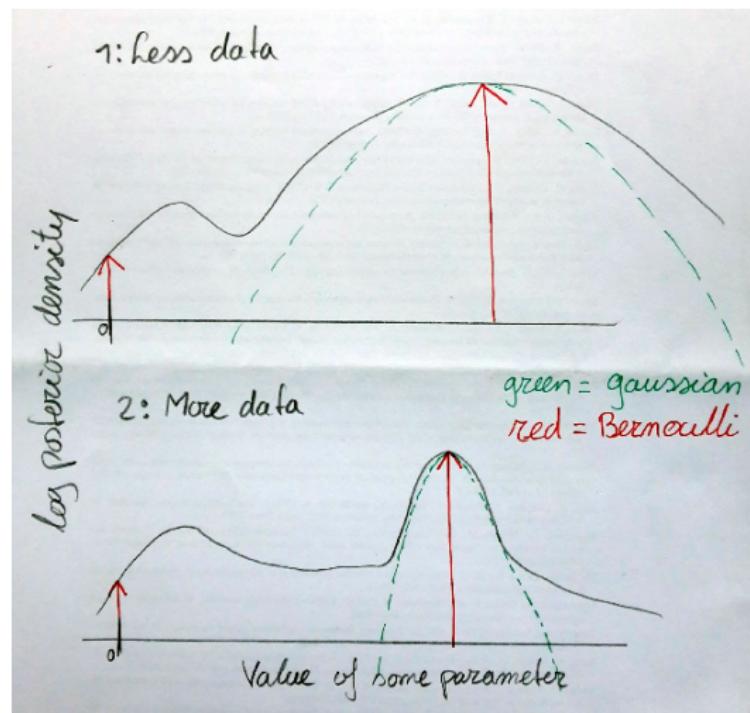


Figure: Bernoulli approximation does not capture local properties

Introduction

Motivation

Method

The goal

Historical
perspective

Bayesian
inference

Parameter
posterior

Uncertainty

Pruning

Experiments
and results

Pruning

Uncertainties

Closing

Adding noise

Figure: animation only shows with media plugin (use adobe reader)