Deep Bayesian Inference: a natural coupling to deep neural networks

Wenfeng Feng
Henan Polytechnic University
fengwf@hpu.edu.cn

Xin Zhang
Henan Polytechnic University
zhangxin@hpu.edu.cn

Abstract

Deep where the observed evidence come from corresponding layer in deep neural network.

1 Introduction

Learning the conditional categorical distribution over a finite set of target labels or categories given the input observations plays an essential role in machine learning. For example, in supervised learning, particular classification, we fit a conditional categorical distribution over a set of target classes given inputs. In self-supervised learning, such as causal (or masked) language modeling, the joint probability distribution of the word sequence is factorized as the product of the conditional categorical distributions over the vocabulary of the next word (or the masked words) given the previous words (or the other unmasked words). In reinforcement learning, a policy is represented as a conditional categorical distribution over a finite set of actions given input states or observations. add citations

The conditional categorical distribution over a finite set of possible target labels given the input observations is commonly represented as:

$$p(\boldsymbol{y}|\boldsymbol{x}) = \boldsymbol{S}(\boldsymbol{f}(\boldsymbol{x})). \tag{1}$$

In the above equation, $\boldsymbol{x}$ is the input observations, y is a random variable that can take its values from a set of n possible target labels, i.e. $\boldsymbol{y} \in \{y_1, \cdots, y_n\}$. The function $\boldsymbol{f}$ transform (convert) the input observations $\boldsymbol{x}$ to vectors in a n-dimensional real vector space $\mathbb{R}^n$. These non-normalized vectors generated by function $\boldsymbol{f}$ are often called logits, which is then passed to Softmax function $\boldsymbol{S}$ that normalize the logits to a categorical probability distribution.

In modern machine learning, the function $\boldsymbol{f}$ is usually implemented by a deep neural network with multiple layers between inputs and outputs. Many deep neural network architectures evolved. CNN, Transformer, Mlp-mixer, etc.

A deep neural network makes it easier to learn multiple levels and aspects of input data features and concepts, layer by layer. A deep neural network are able to create deep representations, at every layer, the network learns a new, more abstract, higher-level, or different aspect representation of the input data features and concepts. https://www.researchgate.net/post/Why_Are_Deep_Networks_Better_Than_Shallow_Ones

On the other hand, Bayesian inference is an iterative process utilizing Bayes' Theorem to deduce(infer) a probability distribution based on new observed data features coming in iteratively.

Since at each layer in a deep neural network, we get a new level or aspect representation of the input data, they can serve as the new observed data in each Bayesian iteration.

In this sense, a Bayesian inference composed of multiple Bayesian iteration, where new observed data is provided by the corresponding layer in a deep neural network. Thus, we call such architecture as Deep Bayesian Inference.

Preprint. Under review.

Rather than, all the different levels and aspects of input data feature can only be aggregated to the last layer of the neural network, and then are converted to categorical probabilities at the same time by the Softmax function, they are converted to probability distribution layer by layer.

Therefore, we utilize Bayesian inference to convert the data features leaned by the neural network to likelihood, layer by layer. Such a deep Bayesian inference architecture naturally coupled to the deep neural network, and stabilize,Advantage of DBI

We coupled DBI to three popular neural network architectures, Transformer, MlpMixer, ConvMixer, respectively.

The primary results of experiments show that: 1. Adding DBI to Transformer decreases the performance of classification. This may because Transformer capture global attention, and Softmax function in Transformer. 2. Adding DBI to MlpMixer could partially release overfitting problem. 3. Adding DBI to ConvMixer increase its performance obviously. This may be because convolution capture local attention layer by layer.

a layer learns something and then passes a representation to the next layer, which learns something else

The Bayesian inference provides a new approach to explanations of deep neural networks.

2 Deep Bayesian Inference

$$p(y=c|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y=c)p(y=c)}{\sum_{c=1}^{C} p(\boldsymbol{x}|y=c)p(y=c)} \tag{2}$$

When dealing with high-dimensional data, it is often useful to reduce the dimensionality by projecting it to a lower dimensional subspace which captures the "essence" of the data. (P46)

2.1 Bayesian Inference

Bayes theorem is a principle way to calculate a conditional probability.

The prior distribution is the available knowledge about the parameters in a statistical model, which is typically determined before any data are collected and observed.

The observed evidence is expressed in terms of the likelihood function of the data given the parameters, is the conditional probability distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ of the data $\boldsymbol{x}$ given parameters $\boldsymbol{\theta}$. In Bayesian inference, the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ comes into the posterior as a function of $\boldsymbol{\theta}$ for observed data $\boldsymbol{x}$. (van de Schoot et al., 2021)

At each Bayesian iteration, the likelihood function is provided by the output of the corresponding layer of the deep neural network.

Both prior distribution and likelihood function are combined via Bayes' Theorem in the form of the posterior distribution. The posterior distribution reflects the updated knowledge about the parameters, balancing the prior knowledge with the observed evidence. The posterior distribution is a kind of weighted average (mean) of the prior distribution, where the likelihood function of the parameters are the weights.

The obtained posterior results can then be used as the prior of the next iteration of Bayesian inference.

Bayesian inference is an iterative process utilizing Bayes' Theorem to deduce(infer) a probability distribution based on new observed data coming in iteratively (Harper, 2009). Bayesian inference allows you to update your beliefs iteratively as new information(data) comes in. It works as follows: you have a prior belief about the distribution of your target, then, after you receive some new data, you can update your beliefs by calculating the posterior distribution by Bayes rule. Afterwards, we get even more data come in. So our posterior becomes the new prior. We can update the new prior with the likelihood derived from the new data and again we get a new posterior. This cycle can continue so long as new data comes in, so we can continuously updating our beliefs.

`https://towardsdatascience.com/probability-concepts-explained-bayesian-inference-for-parameter-e`

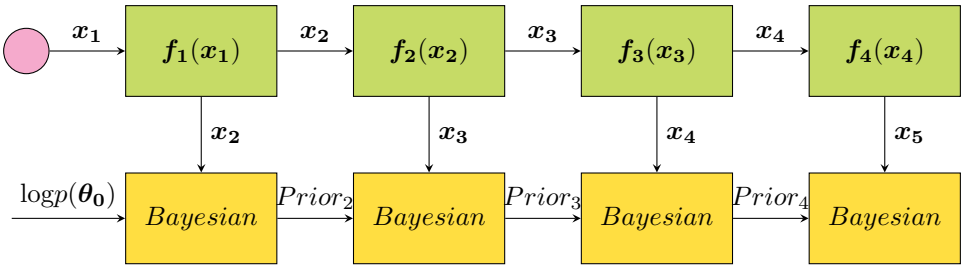

Figure 1: Sample figure caption.

2.2 Language Model

A goal of language modeling is to learn the joint probability function of sequences of words in a language (Bengio et al., 2000).

`https://www.inference.vc/implicit-bayesian-inference-in-sequence-models/`: We can think of these one-step-ahead predictive distributions as implicitly performing Bayesian inference.

3 Deep Bayesian Inference Architecture

DBI is a naturally extension of

Advantage Similar to auxiliary loss, DBI reduce the vanishing gradient problem for earlier layers, stabilize the training and is used as regularization. `https://stats.stackexchange.com/questions/304699/what-is-auxiliary-loss-as-mentioned-in-pspnet-paper`

Deep generative models are used to estimate the likelihood of each observation and to create new samples from the underlying distribution. Suppose the underlying distribution is a categorical distribution, whose parameters are $p_i, i \in [1, N]$. Then the likelihood is defined as $P(p_i|c)$, where c is the context observations. The likelihood works as weights in a Bayesian iteration to update the prior belief. (Ref: An introduction to deep generative modeling)

3.1 Related Works

Bayesian Neural Network A Bayesian neural network (BNN) is commonly defined as a stochastic artificial neural network trained using Bayesian inference (Jospin et al., 2022). While the prior distribution in BNN is taken over the possible model parameterization (is specified for each weight and bias), the prior distribution of Deep Bayesian Inference is taken over the possible parameterization of data distribution.

Normalizing Flow

4 Experiments

Experiment environment and setting We use Kaggle's free accessed GPUs to do our experiments. Kaggle provides a NVIDIA TESLA P100 GPU for each user which has 15.9GB memory, with a limit on each user's GPU use of 30 hours/week. Due to the limitation of the hardware, we mainly do experiments on Tiny ImageNet, rather than ImageNet.

4.1 Image Classification

Vision Transformers and CNN Dosovitskiy et al. (2020)

Tiny ImageNet The Tiny ImageNet dataset is a modified subset of the original ImageNet dataset. Here, there are 200 different classes instead of 1000 classes of ImageNet dataset, with 100,000 training examples and 10,000 validation examples. The resolution of the images is just 64x64 pixels.

5 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 7 regarding figures, tables, acknowledgments, and references.

6 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

6.1 Headings: second level

Second-level headings should be in 10-point type.

6.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

7 Citations, figures, tables, references

These instructions apply to everyone.

7.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `neurips_2022` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{neurips_2022}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

7.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

7.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

7.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

8 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

[1] Sample of the first footnote.

[2] As in this example.

Table 1: Sample table title

	Part	
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

9 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

  Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

9.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments and Disclosure of Funding

Thanks Kaggle for providing free GPUs. Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

References

Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 13.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Marc Harper. 2009. The replicator equation as an inference dynamic. *arXiv preprint arXiv:0911.1763*.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. 2022. Hands-on bayesian neural networks—a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48.

Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, et al. 2021. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 5.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? **[TODO]**
(b) Did you mention the license of the assets? **[TODO]**
(c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**

5. If you used crowdsourcing or conducted research with human subjects...
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**