Everything that Works Works Because it's Bayesian: Why Deep Nets Generalize?

May 25, 2017 ·

Everything that Works Works Because it's Bayesian: Why Deep Nets Generalize?

The Bayesian community should really start going to ICLR. They really should have started going years ago. Some people actually have.

For too long we Bayesians have, quite arrogantly, dismissed deep neural networks as unprincipled, dumb black boxes that lack elegance. We said that highly over-parametrised models fitted via maximum likelihood can't possibly work, they will overfit, won't generalise, etc. We touted our Bayesian nonparametric models instead: Chinese restaurants, Indian buffets, Gaussian processes. And, when things started looking really dire for us Bayesians, we even formed an alliance with kernel people, who used to be our mortal enemies just years before because they like convex optimisation. Surely, nonparametric models like kernel machines are a principled way to build models with effectively infinite number of parameters. Any model with infinite parameters should

be strictly better than any large, but finite parametric model, right? Well, we have been proven wrong.

But maybe not, actually. We Bayesians also have a not-so-secret super-weapon: we can take algorithms that work well, reinterpret them as approximations to some form of Bayesian inference, and voila, we can claim credit for the success of an entire field of machine learning as a special case of Bayesian machine learning. We are the BORG of machine learning: eventually assimilate all other successful areas of machine learning and make them perfect. Resistance is futile.

We did this before: L1 regularisation is just MAP estimation with sparsity inducing priors, support vector machines are just the wrong way to train Gaussian processes. David Duvenaud and I even snatched herding from Max Welling and Alex Smola when we established **herding is just Bayesian quadrature done**

slightly wrong.

But so far, we just couldn't find a way to claim credit for all of deep learning. Some of us tried to come through the back-door with Bayesian neural networks. It helps somewhat that Yann LeCun himself has written **a paper on the topic**. Yarin managed to claim **dropout is just variational inference done wrong**. But so far, Bayesian neural networks are just a complementary to existing successes. We could not so far claim that deep networks trained with stochastic gradient descent are Bayesian. Well, fellow Bayesian, our wait may be over.

Why do Deep Nets Generalise?

HINT: because they are really just an approximation to Bayesian machine learning.

One of the hottest topics at ICLR this year was generalisation in deep neural networks, and it seems to continue with a number of NIPS submissions. If you're not aware of this trend, read up on it, it's pretty interesting. It turns out, neural networks by themselves are indeed the stupid, dumb, highly overparametrised black-boxes we always said they were after all. People did an experiment **(Zhang et al, 2017)**: you shuffle all labels in your training dataset - removing all information between inputs and labels - and the network will happily overfit to this complete garbage and achieve zero training error. If you have completely

random labels, surely the only way to learn them is by memorizing each example and basically learning a sort of look-up-table. If you learn by memorization, you didn't really learn anything about generalisation in your data, your generalisation error will be huge. Related observations were made by **Krueger et al, (2017)**. See Also **Ben Recht's ICLR slides**

So what is it - if not the network itself - that makes deep network training generalise? It appears to be that **stochastic gradient descent** may be responsible. **(Keskar et al, 2017)** show that deep nets genealise better with smaller batch-size when no other form of regularisation is used. And it may be because SGD biases learning towards flat minima, rather than sharp minima. This is something that (surprise-surprise) **Hochreiter and Schmidhuber (1997)** have worked on before. Most recently, **(Wilson et al, 2017)** show that these good generalisation properties afforded by SGD diminish somewhat when using popular adaptive SGD methods such as Adam or rmsprop. Finally, there is contradictory work by **Dinh et al, (2017)** who claim sharp minima can generalize well, too. See also **(Zhang et al, 2017)** and **Jorge Nocedal's ICLR slides** on the same topic.

In summary: The reason deep networks work so well (and generalize at all) is not just because they are some brilliant model, but because of the specific details of how we optimize them. Stochastic gradient descent does more than just converge to a local optimum, it is biased to favour local optima with certain

desirable properties, resulting in better generalization. SGD itself, and the question of flat vs sharp minima, should therefore be of interest to Bayesians still trying to wrap their heads around the success of dumb deep networks.

> We only need a small, vague claim that SGD does something Bayesian, and then we're winning.

Update: apparently, Bayesians have already started the process:

- Mandt, Hoffman and Blei (2017) **Stochastic Gradient Descent as Approximate Bayesian Inference**

Flat minima and Minimum description length

So SGD tends to find flat minima, minima where the Hessian - and consequently the inverse Fisher information matrix - has small eigenvalues. Why would flat minima be interesting from a Bayesian pespective?

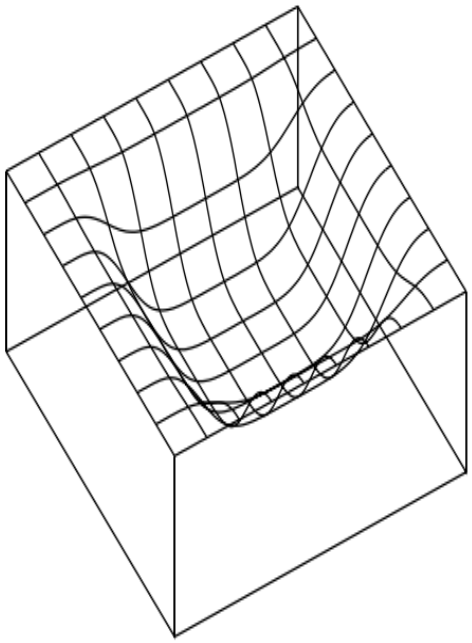


Figure 1: *Example of a "flat" minimum.*

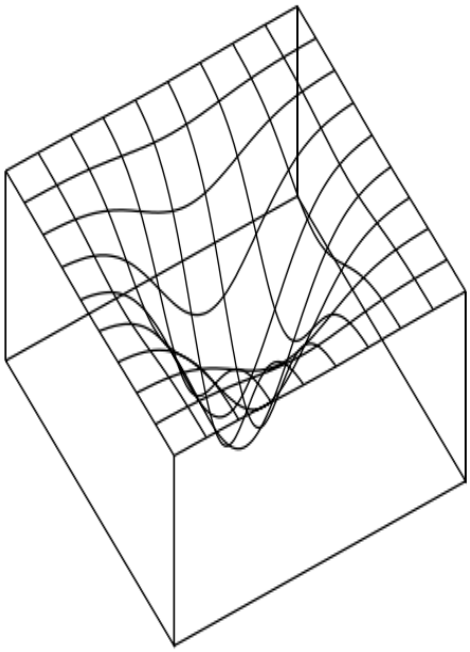


Figure 2: *Example of a "sharp" minimum.*

It turns out, **(Hochreiter and Schmidhuber, 1997)** motivated their work on seeking flat minima from a Bayesian, minimum description length perspective. Even before them, **(Hinton and van Camp, 1993)** presented the same argument in the context of Bayesian neural networks. The intuitive argument goals as follows:

If you are in a flat minimum, there is a relatively large region of parameter space where many parameters are almost equivalent inasmuch as they result in almost equally low error. Therefore, given an error tolerance level, one can describe the parameters at the flat minimum with limited precision, using fewer bits while

keeping the error within tolerance. In a sharp minimum, you have to describe the location of your minimum very precisely, otherwise your error may increase by a lot.

Update: I removed the discussion of Jeffreys priors from here as people pointed out I was wrong.

Conclusions

Everything that works works because it's Bayesian. Maybe deep learning only works because it uses SGD, and maybe SGD is a rudimentary way of implementing Bayesian occam's razor. Whether you're a Bayesian or not, there's an interesting theory developing around generalisation in deep networks, and I think everyone - including Bayesians - should be aware of it. It seems like stochastic gradient descent and its tendency to seek out flat minima has a lot to do with why deep nets don't fail as miserably as Bayesians predicted them to. Seeking flat minima makes sense from a minimum description length perspective.

I should point out - if it's not clear - that the title, intro, and to some degree this conclusion were meant as a provocative joke. I'm pretty sure I'm going to get a lot of love for this post nevertheless. Please just don't take me too seriously.

Tim G •2 years ago

Hi Ferenc,
just to add one more interesting reference from the famous Tali Tishby:
https://arxiv.org/abs/1703....
In a nutshell, they observe that SGD essentially consists of two phases - error reduction and compression. They also find some arguments why deep networks work better than shallow ones.
While not strictly showing that SGD is approximate variational Bayes, I would still count this paper towards the "collective"...

Zak Jost > Tim G •2 years ago

I was going to call out the same work. I think the information plane is a very nice visualization of this process, and the various types of mutual information a great framework for understanding the phenomena.

Riccardo Zecchina •2 years ago

Perhaps the following results can help the discussion. For simple architectures and random patterns we find that there exist rare regions of optima which have a very high local entropy (dense cluster of high likelihood configurations, i.e. large flat minima). These regions have

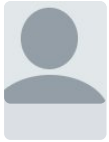
escaped the traditional statistical physics analysis and could be revealed only by a novel technique. Working algorithms end up un such regions in spite of the existence of exponentially many local minima and isolated optimal minima (of course these algorithms should not satisfy detailed balance).

https://arxiv.org/abs/1509....
https://arxiv.org/abs/1605....
https://arxiv.org/abs/1511....
https://arxiv.org/abs/1602....

For analytical purposes the results are for discrete weights and simple architectures (still highly non-convex).
They can be generalized.

dimt •2 years ago

Great article, very well written.
One thing I don't get. Surely a deep net would even fit random labels, with no generalization of course. On the other hand a proper Bayesian approach shouldn't over-fit in this case, right? So maybe some Bayesian properties, but still many differences?



T. •2 years ago• edited

Human brain is like this. You're given a task of memorizing a sequence of cards, you will first find common patterns to memorize them, and it will generalize. When you fail to find any patterns, you can STILL memorize them by memorization, and it probably won't generalize. So human brain is powerful and lazy, resembles observations of DNNs in (Zhang et al, 17).

Егор Зворыкин •2 years ago

(Keskar et al, 2017) show that deep nets geneRalise better with smaller batch-size.

everlasting •2 years ago• edited

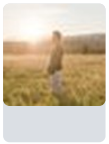
Forgive me for a dumb question, but can you elaborate on this?
> under the Jeffreys prior you a priori favour models where changing the parameters slightly doesn't change the model's predictions much
In the 1d case for example, for θ with higher fisher information, $E[(\frac{dL}{d\theta})^2]$ is larger, so is the expected variation of likelihood given a fixed variation on θ, $E[(L(\theta+d\theta)-L(\theta))^2]$. It doesn't seem consistent with the claim that the prediction doesn't change much...

Ferenc Huszar Mod > everlasting •2 years ago

you're right, I think I got it the wrong way around. This is very interesting.

Also, Jeffreys (in the 1-d case) is a reference prior, so it's minimally informative, which means that the mutual information between the parameter and the data is maximised, which also means the parameter should be highly sensitive to the data, which again implies a sharp minimum.

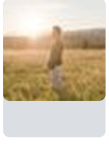
I'll correct this.



Shawn Pan •a year ago

When you mention the flat minimum, it reminds Mackay's philosophy on model selection. It turns out to be the identical.

[1] MacKay, D. J. C. Bayesian Interpolation. Neural Comput. 4, 415–447 (1992).



Shawn Pan •a year ago

Very nice post! I am new to Bayesian Deep learning. I have read the Yarin's paper on dropout as Bayesian inference. I don't get why you say it is VI done *wrong*? From my viewpoint, it is just a special VI with pre-described certain prior and posterior.

Samuel Smith •2 years ago

You can replicate the results of Zhang et al. in a linear model, and show that they are quantitatively explained by evaluating the Bayesian evidence ratio between the model and a null model assigning equal probability to every label:
https://arxiv.org/pdf/1710....

In short:
Random labels: evidence ratio favours the null model
Informative labels: evidence ratio favours the learned model

I do not understand how the ML community has completely forgotten David Mackay's papers from the 90's!
https://authors.library.cal...

Rosenbad > Samuel Smith •2 years ago

That's quite sad indeed...

H Summers •2 years ago

Perhaps you have a modest proposal for a solution?

Sean O'Connor •2 years ago

Why don't you evolve and SGD the same neural network setup and find out?

ddm •2 years ago

Nice post.
I wouldn't say that "there is contradictory work by Dinh et al, (2017)", in fact I understood that they find issues with the definitions of "flat" and "sharp" minima for deep networks, and not with the intuition behind them. Quoting their abstract https://arxiv.org/abs/1703.... "This paper argues that

most notions of flatness are problematic for deep models and can not be directly applied to explain generalization. "

Carles Gelada •2 years ago

But even if large batches generalize worse, they still have good generalization. I see two possibilities of why that might be the case. Maybe the step size also acts as a Bayesian prior or maybe neural networks just don't have truly sharp minima (even the local minima found by batch methods is mostly flat)

yaringal •2 years ago• edited

I think it goes beyond the joy of claiming "my model subsumes yours" (a running joke in bayesian modelling).
Any theoretical framework used to explain deep learning, as long as it can