View Reviews

Paper ID
2518

Paper Title
Uncertainty propagation in neural networks for sparse coding

Reviewer #1

Questions

1. Please provide an "overall score" for this submission.
4: An okay submission, but not good enough; a reject. I vote for rejecting this submission, although I would not be upset if it were accepted.

2. Please provide a "confidence score" for your assessment of this submission.
2: You are willing to defend your assessment, but it is quite likely that you did not understand central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.
This paper proposes a Bayesian version of the LISTA algorithm [11], which performs better on small data and provides uncertainty estimates which can be used for active learning.

I found this paper quite confusing for two reasons: firstly the motivation/evaluation, and secondly the presentation of the model/inference. My confusion is sufficient that I maintain a non-negligible probability that I have substantially missed the point (this is reflected in my confidence score), but here are my thoughts nonetheless.

1) Motivation / evaluation

The LISTA approach is for solving the l1 regularized linear regression problem efficiently, not about building neural network models. The presented approach is described in the the first three paragraphs of the introduction as a BNN, but in fact the model itself is a linear (and non-bayesian) one. It confuses me to talk of uncertainties, as the uncertainties are in an inference procedure, not in the model itself. It may be useful to consider uncertainties in deterministic settings (e.g. Bayesian optimization and more generally the whole field of probabilstic numerics), but if this is setting then I am confused by 6.3, which appears to be using the uncertainty in the downstream task of active learning. I don't understand how the uncertainties are being used, as the model itself is deterministic. Perhaps the model is that the weights in the linear model are the output of the Bayesian LISTA model, but in this case I find it a strange task as the model is already specified in (1). The original LISTA algorithm offered improvements over ISTA, even though it was an approximation to it ('It seems that a small amount of data-specific mutual inhibition is all that is needed to explain away unnecessary components of the code vector' [11]),

but this quality is not discussed in the presented work.

There appear to be two sorts of sparsity: the l1 regularisation in the original model, which manifests itself in the soft thresholding non-linearity, and the spike and slab priors in the Bayesian LISTA network. It confuses me how these two sparsities are related. In the non-bayesian version the components of beta are already sparse, so I don't understand the motivation to add additional sparsity priors over the beta_l.

To summarize the above into three questions:
* what is the motivation for being Bayesian about the LISTA algorithm?
* how does in the uncertainty in the Bayesian LISTA procedure relate to the original linear model?
* how does the sparse promoting prior interact with the sparsity promoting non-linearity?

2 Model / inference

The separation of model and inference is not clear to me. In the 'forward propagation' step there are number of approximations, but then there is also an approximate posterior in section 5. I am confused as to whether the model has been simplified to yield tractable inference (albeit with a approximate posterior) or whether the approximations in section 4 are part of the approximate inference.

Aside from the above issues, there are a few imprecise statements that I feel should be addressed:

10 'highly efficient'. In what sense? How is this claim justified?

21 'effectively' in what sense?

96 'closely approximated' in what sense? The supplementary material contains two histograms, which show a convincing match of a single 1D marginal, but I would have thought this hardly sufficient for such claims. This claim could perhaps be rephrased to reflect the empirical nature of the evaluation.

132 Is the assumption of mutual independence over the elements of S and beta_{l-1}? This assumption is not discussed, and it seems to me quite a significant assumption.

Some small points:

37 typo missing an 'and'

in (8) of the supplementary material, what is 's'?

I find the notation of B for the final betas quite confusing, as e.g. beta is used in section 6.

Comments following the guidelines as requested:

Quality: Strengths: The general tone, presentation and referencing of the paper is good. There is a detailed supplementary material and experiments that do more than compare log-likelihoods. Weaknesses: the motivation of the paper is unclear to me, and it unclear what the experiments should be assessing.

Clarity: Strengths: the paper is well written and the language is generally clear. Weaknesses: some of the language is imprecise, and I find several aspects confusing, as explained above.

Originality: Strengths: the approximation of Gaussian with spike-slab is novel to my knowledge, and using bayesian approaches for this deterministic problem is novel. Weaknesses: it is not shown why being Bayesian is necessary in the context, and how the uncertainty in the algorithm relates to the model.

Significance: it may be that presented approach is generally useful as a BNN with the soft thresholding non-linearity. Potentially, the efficacy of the of approximations over this non-linearity could inspire new BNN architectures Weaknesses: It not shown that the construction is useful in general, the improvements over LISTA are only demonstrated on a limited set of synthetic/mnist experiments, and not across all measures.

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?
3: Very confident

Reviewer #2

Questions

1. Please provide an "overall score" for this submission.
5: Marginally below the acceptance threshold. I tend to vote for rejecting this submission, but accepting it would not be that bad.

2. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals. You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.
The paper suggests a Bayesian version of Learned ISTA, which learns a fixed number of fully connected weights/soft thresholding layers to approximate sparse encodings of input vectors with respect to a fixed dictionary. Intermediary decomposition coefficients are approximated by a spike and slab distribution to account for sparsity. Some form of probabilistic propagation is used to learn the parameters, using assumed density filtering and expectation propagation. Numerical experiments demonstrate that Bayesian LISTA can learn with a small number of samples, as well

as outperform LISTA on some metrics.

The topic of managing uncertainty in sparse coding and/or multi-layer model is of interest. Although the proposed method appears interesting, the contribution is quite incremental, with most of the method relying on existing tools for probabilistic back-propagation.

In addition, the numerical experiments fail to be fully convincing in my opinion. In particular:

- The numerical experiments compare the methods in terms of NMSE and F-measure (where each element of the support of the true output is treated as a positive example). Bayesian LISTA outperforms LISTA on NMSE (esp. with more layers/iterations), while LISTA yields better F-measures. Unfortunately, the paper has no insight as to why that is. For instance, does the proposed method somehow weakens the ability of LISTA to avoid selecting similar codewords, leading to less sparse encodings? A discussion and/or experiments comparing supports or sparsity of the output encodings of the method would shed some light on what the proposed method achieves.

- Both the synthetic and MNIST experiments use a somewhat contrived Gaussian design. From what I understand, the MNIST experiment presented in the paper is about generating random Gaussian linear combinations of the MNIST images, and then trying to recover the original coefficients. No sparsity is imposed. This looks like an odd task to consider, and I am confused as to what the goal was here.

The paper is clear and well-written.

Remarks

equation 5: should `y` be `y^{(n)}`?

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?
2: Somewhat confident

Reviewer #3

Questions

1. Please provide an "overall score" for this submission.
4: An okay submission, but not good enough; a reject. I vote for rejecting this submission, although I would not be upset if it were accepted.

2. Please provide a "confidence score" for your assessment of this submission.
3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

3. Please provide detailed comments that explain your "overall score" and "confidence score" for this submission. You should summarize the main ideas of the submission and relate these ideas to previous work at NIPS and in other archival conferences and journals.

You should then summarize the strengths and weaknesses of the submission, focusing on each of the following four criteria: quality, clarity, originality, and significance.

Summary

This paper proposes a bayesian framework to estimate uncertainty in neural network for sparse coding. It computes the propagation of the uncertainty in the network through the forward pass and derives a probabilistic backpropagation algorithm to infer the different parameters. The performances are then evaluated on different compress sensing problems and the uncertainty estimation is used in an active learning algorithm to improve the code recovery performance.

Overall assessment

The global idea of this paper is interesting as it permits to extend uncertainty quantification for activation signal to the activation estimated with neural sparse coding. The uncertainty model derived seems correct but I did not checked the proof of the lemmas in supplementary. However, the probabilistic backpropagation is not detailed enough. The update rules for the backpropagation algorithm should be given for each parameter, at least in the supplementary. The minimization of the KL divergence should also be discussed to give a clear view of the computational burden it creates. Also, the scalability of the proposed technique compared to classical backpropagation is not discussed at all in this part.

The experimental validation of the proposed method is weak. My main concern is that the loss used to train the network is not detailed. It seems from (1) that the network is trained with the LASSO loss. In this case the choice of alpha is a critical part in the performance of the algorithm and determine the value of the parameter lambda (alpha / E) which is set to a fix value 0.1 in all experiments. As this parameter determine the sparsity of the sparse code resulting from the network and on the coefficient amplitude bias, it has a strong impact both measure F and NMSE. The introduction of this parameter lambda makes it very hard to evaluate the results.

Moreover, for the simulated experiments, the scale of the problem seems very limited as only 1000 samples are used to train the network and 100 samples used in the test set, which do not appear statistically significant. Even with this small setting, the results with the bayesian LISTA are not clearly better than the one of the other methods. The F-measure score of both LISTA and Bayesian LISTA are worst than the one from ISTA in term of F-measure but these 2 algorithms are initialized with the same performances.
The experimental setup for the second experiment on MNIST, is also very limited. Only 200 samples are used from the 60,000 available, suggesting that the proposed technique scale poorly. The results are also displayed relatively to the number of training iteration and not with varying number L of layers/iterations as in the previous experiment, which do not really make sense in this setting.

Typos and nitpicks

- l61: Remove (.) in "E(.)"
- l66: E is the energy function and the largest eigenvalue...

- l66: S should not depend on W as it makes is harder to understand that both are optimized imdependently.
- Algorithm1: In LISTA, lambda should be set to 'alpha / E'.
- Eq (16): the distribution T is not defined.

4. How confident are you that this submission could be reproduced by others, assuming equal access to data and resources?

1: Not confident