

← Go to **ICML 2023 Conference** homepage (/group?id=ICML.cc/2023/Conference)

On the Functional Similarity of Robust and Non-Robust Neural Representations



András Balogh (/profile?id=~Andr%C3%A1s_Balogh1), Márk Jelasity (/profile?id=~M%C3%A1rk_Jelasity1)

Published: 24 Apr 2023, 23:55, Last Modified: 24 Apr 2023, 23:55

ICML 2023 Poster

Conference, Area Chairs, Authors, Reviewers, Senior Area Chairs

Revisions (/revisions?id=sFqfXphJh5)

Edit ▾

Keywords: model stitching, functional representational similarity, adversarial robustness, robust representations

TLDR: We study the similarity of robust and non-robust neural representations with model stitching, and find that they are similar regarding accuracy, but the similarity w.r.t. robust accuracy depends on the location of the stitching layer.

Abstract:

Model stitching---where the internal representations of two neural networks are aligned linearly---helped demonstrate that the representations of different neural networks for the same task are surprisingly similar in a functional sense. At the same time, the representations of adversarially robust networks are considered to be different from non-robust representations. For example, robust image classifiers are invertible, while non-robust networks are not. Here, we investigate the functional similarity of robust and non-robust representations for image classification with the help of model stitching. We find that robust and non-robust networks indeed have different representations. However, these representations are compatible regarding accuracy. From the point of view of robust accuracy, compatibility decreases quickly after the first few layers but the representations become compatible again in the last layers, in the sense that the properties of the front model can be recovered. Moreover, this is true even in the case of cross-task stitching. Our results suggest that stitching in the initial, preprocessing layers and the final, abstract layers test different kinds of compatibilities. In particular, the final layers are easy to match, because their representations depend mostly on the same abstract task specification, in our case, the classification of the input into n classes.

Supplementary Material: zip (/attachment?id=sFqfXphJh5&name=supplementary_material)

Financial Aid: abalogh@inf.u-szeged.hu

Paper Checklist Guidelines: I certify that all co-authors of this work have read and commit to adhering to the Paper Checklist Guidelines, Call for Papers and Publication Ethics.

Submission Number: 4440

Filter by resolution ▾

Filter by author ▾

Search keywords...

Sort: Newest First

Everyone Program Chairs Authors Area Chairs Reviewers Submitted

Senior Area Chairs Reviewer B1YL Reviewer kf1a Reviewer rEdH

10 / 10 replies shown



Paper Decision

Decision

✍ Program Chairs (👤 kyunghyun.cho@nyu.edu (/profile?id=kyunghyun.cho@nyu.edu), emma.p.brunskill@gmail.com (/profile?id=emma.p.brunskill@gmail.com), krausea@ethz.ch (/profile?id=krausea@ethz.ch), ebrun@cs.stanford.edu (/profile?id=ebrun@cs.stanford.edu), +7 more (/group/info?id=ICML.cc/2023/Conference/Program_Chairs))

📅 24 Apr 2023, 22:11 (modified: 24 Apr 2023, 23:38) 👁 Program Chairs, Authors

Decision: Accept (Poster)

Comment:

This submission performs a variety of experiments using model stitching to investigate the compatibility of representations from robust and non-robust networks, when stitched to optimize either clean or robust accuracy. It shows that maintaining robustness with stitching two robust accuracy requires optimizing robust accuracy, but stitching robust and non-robust networks leads to loss of robustness. Two reviewers assess the paper positively, while one raises concerns regarding the relationship between the results and those of Cianfarani et al. (2022). I share some of these concerns—it is difficult to interpret situations where representations are compatible but the activations are not similar—but I nonetheless feel that the investigation here is a worthwhile contribution to the literature analyzing the differences between robust and non-robust neural networks. I thus recommend acceptance.

Official Review of Submission4440 by Reviewer B1YL

Official Review ✍ Reviewer B1YL 📅 11 Mar 2023, 19:04 (modified: 14 Mar 2023, 14:39)

👁 Program Chairs, Area Chairs, Authors, Reviewer B1YL, Reviewers Submitted, Senior Area Chairs

Summary:

This paper studies how well two different convolutional neural network models applied for computer vision tasks can be stitched together to achieve similar "plain" accuracy and similar robust accuracy. Stitching is performed by inserting a learned linear mapping from the internal representation of one network to that of another and then replacing the rest of the network with the one being stitched to. The authors perform experiments testing what the resulting networks achieve in terms of accuracy, robust accuracy (on a test set of adversarial examples), and model inversion. The authors also perform some analysis to explain the observed behaviors.

Strengths And Weaknesses:

Overall this is an interesting paper with an intuitive setup and some interesting observations. It is particularly interesting that inserting a single non-robust transformation when stitching a network to itself is enough to remove its robustness property. The experiments are well setup and well motivated. The deeper dive on when the representations "break" and make the network susceptible to adversarial examples is also very interesting.

Weaknesses:

- This paper might benefit from a broader scope in its experiments. It could include more and newer models and architectures (such as Visual Transformers, self-supervised, weakly supervised, and text supervised models) and it could expand to bigger datasets (such as ImageNet and/or LAION). However, I accept that there is a computational barrier to carrying out such experiments and this is probably not to be held against the paper.
- The paper could also consider other interesting functional properties beyond accuracy on adversarial examples and accuracy on the standard test set. It could look at OOD generalization and other datasets meant to test for robustness in different ways (such as ImageNet-C, A, O, R, V2, etc)

Questions:

I do not have any questions; overall, I believe this paper is novel and of interest to the community as-is.

Limitations:

Yes.

Ethics Flag: No

Soundness: 4 excellent

Presentation: 4 excellent

Contribution: 3 good

Rating: 7: Accept: Technically solid paper, with high impact on at least one sub-area, or moderate-to-high impact on more than one areas, with good-to-excellent evaluation, resources, reproducibility, and no unaddressed ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Rebuttal by Authors

Rebuttal

Authors (András Balogh (/profile?id=~András_Balogh1), Márk Jelasity (/profile?id=~Márk_Jelasity1))

17 Mar 2023, 20:05 (modified: 20 Mar 2023, 15:40)

Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Revisions (/revisions?id=8B9kzk1nHu)

Rebuttal:

Although no questions were asked, we wanted to thank the reviewer for the suggestions on other functional properties to look at. We have been thinking about investigating other properties ourselves, and we agree that these ones would be rather interesting to study. It would be interesting to identify what properties “move together” and which ones are independent.



Official Review of Submission4440 by Reviewer kf1a

Official Review Reviewer kf1a 28 Feb 2023, 13:33 (modified: 25 Mar 2023, 16:57)

Program Chairs, Area Chairs, Authors, Reviewer kf1a, Reviewers Submitted, Senior Area Chairs

Revisions (/revisions?id=Pq9LR5Gxs)

Summary:

This paper uses the model stitching framework to quantify the functional similarity of robust and non-robust networks. They show that robust and non-robust representations are functionally compatible w.r.t. natural accuracy, but this does not preserve robust accuracy. Similarly, they show that robust and non-robust representations are not functionally compatible w.r.t. robust accuracy, except when stitching the last few layers. Subsequent experiments on cross-task stitching and model inversion explain why stitching at the last few layers preserves functional compatibility.

Strengths And Weaknesses:

1. The study's reliance on the model stitching framework carries forward several limitations. For instance, the method only works when both networks share the same architecture, leading to the question of which layers to stitch across architectures. Additionally, the choice of stitching layer is based mostly on heuristics, leaving open questions regarding why 1x1 convolutions are the right approach to stitch layers, given that they cannot implement rotations or orthogonal linear transformations.
2. Discrepancy in findings: The findings of this study differ from those of Cianfarani et al. (2022), which analyzed layer-wise representations of larger robust and non-robust networks and found significant differences in their internal representations. While the authors of the present study suggest that these discrepancies may be due to issues with representational similarity metrics, the lack of agreement between the two studies raises concerns about the reliability of the findings.
3. The paper's main takeaways and practical implications are not well-defined. The experiments are limited to CIFAR-10, and it is unclear whether the findings can be generalized to larger networks and datasets. The analysis on why last layers are stitchable is not particularly novel or insightful given past work on understanding layerwise representations (as discussed in the paper). Additional work is necessary to establish the significance and generalizability of the findings.

Questions:

Please see strengths And weaknesses

Limitations:

Please see strengths And weaknesses

Ethics Flag: No

Soundness: 2 fair

Presentation: 2 fair

Contribution: 2 fair

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Rebuttal by Authors

Rebuttal

Authors (András Balogh (/profile?id=~András_Balogh1), Márk Jelasity (/profile?id=~Márk_Jelasity1))

17 Mar 2023, 19:49 (modified: 20 Mar 2023, 15:40)

Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Revisions (/revisions?id=4pRcjMseL)

Rebuttal:

The reviewer touches on some very interesting, fundamental questions regarding our work. We are happy to have this chance to share our views on these questions. We will give our rebuttal in two parts.

Stitching as a tool: We completely agree that our work inherits any limitations that the stitching methodology might have. Nevertheless, at this point, this is the most suitable methodology we are aware of to study functional similarity; indeed, it helped us to shed light on some previously unexplored properties of robust representations

Choice of stitching layer: Indeed, the choice of stitching layer is heuristic. This has been discussed in depth in related work: to simplify the argument considerably, we need a transformation that is rich enough to be useful (otherwise meaningful similarity cannot be captured), but not too rich (otherwise the stitching layer could learn the task itself). In particular, the 1x1 convolution for convolutional architectures was proposed independently by Bansal et al and Csiszár et al at NeurIPS 2021, where both papers argued that this choice is in the “sweet spot” in the above sense.

Stitching different architectures: While this is a highly non-trivial question, we respectfully disagree with the strong formulation that stitching “only works when both networks share the same architecture”; to the best of our knowledge, this aspect has not yet been explored in depth (apart from stitching convolutional architectures with different width), but stitching different architectures is certainly a possibility. Also, in our preliminary experiments, we played with stitching along non-identical layers (ie, skipping some layers in the front or end network) and that also worked quite well, in fact, in some cases even slightly better than stitching along matching layers (but we omitted these results to focus our resources on other issues).



Rebuttal by Authors

Rebuttal

Authors (András Balogh (/profile?id=~András_Balogh1), Márk Jelasity (/profile?id=~Márk_Jelasity1))

17 Mar 2023, 19:54 (modified: 20 Mar 2023, 15:40)

Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Revisions (/revisions?id=DY3M5FXPqI)

Rebuttal:

This is part II of our rebuttal response.

Cianfarani et al paper: Unfortunately, we will need some further clarifications from the reviewer on how the results of Cianfarani et al might contradict our results, because we do not find any obvious discrepancies. We found this an interesting and important reference, but there, centered kernel alignment (CKA) is used, and the robust and non-robust representations are not compared directly, but indirectly by looking at the internal layer-wise similarity within a given network, and the response of a given network to adversarial inputs. Therefore, the overlap between this paper and ours is rather minimal. The only point we could find where there is some overlap is the observation that adversarial inputs perturb the deeper layers more while the representations of a few initial layers are almost unchanged even in non-robust networks. This particular conclusion is consistent with our own results, where we can see that when stitching in the initial layers, robustness can be preserved to a certain extent even if the front network is non-robust.

Reliability of our findings: It is a rather strong accusation to question the reliability of the results of our paper. It would help if we were given some evidence to be able to react to any specific issues. In general, we tested our implementation extensively, and we provided the source code as well so that the community can

verify that the methods we describe indeed give the results that we describe.

Scope of our experiments: Note that in the supplementary part (appendix) we do present experiments with larger models and other (although not larger) datasets. These experiments are consistent with those presented in the paper. Nevertheless, we completely agree that, as we also describe in the limitations section of the paper, it would be very useful to perform many more experiments with additional architectures and datasets. Our resources allowed us to perform the set of experiments described in the paper.



➤ *Replying to Rebuttal by Authors*

Thanks for the response

Official Comment ✍ Reviewer kf1a 📅 25 Mar 2023, 16:53 (modified: 25 Mar 2023, 16:56)

👁 Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

📄 Revisions (/revisions?id=9vSuFIOQ44)

Comment:

- Thank you for clarifying the differences between your work and Cianfarani et al. I had misinterpreted some findings in Cianfarani et al.
 - However, I do not understand this claim: "centered kernel alignment (CKA) is used, and the robust and non-robust representations are not compared directly, but indirectly by looking at the internal layer-wise similarity within a given network, and the response of a given network to adversarial inputs"---why is CKA an "indirect" way to compare representations? For both CKA and stitching, one needs to align the representations somehow, stitching does this via stitching layers whereas CKA takes a similarity-based approach.
 - "Nevertheless, at this point, this is the most suitable methodology we are aware of to study functional similarity"---what does mean "most suitable" mean here? Are there any concrete reasons why stitching is definitively better than representation similarity metrics?
- Re: reliability, I only wanted to *why* the main findings would potentially differ (e.g., is it stemming from CKA vs. stitching? is it the architecture / dataset differences?). I was not making any accusations about the reproducibility of the work..
- I am not convinced about the feasibility / effectiveness of stitching different architectures. This is a major limitation of the stitching framework, and it is certainly possible that there is no "one-to-one" mapping between layers of different architectures.

The rebuttal has addressed my concern about overlap with Cianfarani et al, so I am increasing my score to 4. Replicating these findings on larger datasets (e.g. ImageNet) + discussing practical implications of these findings would make the paper stronger.



➤ *Replying to Thanks for the response*

Further clarifications

Official Comment

✍ Authors (👁 András Balogh (/profile?id=~András_Balogh1), Márk Jelasity (/profile?id=~Márk_Jelasity1))

📅 26 Mar 2023, 08:51

👁 Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Comment:

Based on these new comments, some of the concerns of the reviewer are clearer to us so now we can hopefully provide more satisfactory answers.

Indirect comparison: What we meant is that

1. CKA itself is indirect, because it does not take any function (e.g. accuracy) into account, and also that
2. Cianfarani et al. mainly focus on similarities within a network, and not between networks, hence their comparison of robust and non-robust networks is indirect (they show that these networks have different internal similarities).

But we now understand the apparent contradiction with some of our findings. Fig 2(d) (Cianfarani et al) shows that robust and non-robust representations have a low CKA similarity (on clean inputs) but we show that robust and non-robust representations are functionally compatible regarding clean accuracy. This apparent contradiction, however, is *simply another example of the well-known issue* that CKA similarity cannot predict functional similarity, the very reason we use stitching. This brings us to the next topic:

Stitching vs. CKA: When it comes to judging compatibility regarding clean accuracy, CKA has repeatedly been shown to perform poorly. We give three specific examples very briefly, with references:

1. Ding et al (NeurIPS 2021), Section 3.2: A representation and its own low-rank approximation have a higher CKA similarity than that of the representations of two networks trained with different random initializations, yet the corresponding clean accuracy might drastically differ
2. Mirzadeh et al (ICLR 2021) Section 2: In the context of continual learning, quote "If the number of training epochs per task is small (e.g., in streaming case), the CKA does not change much, even though the accuracy for previous tasks drops drastically".
3. Csiszárík et al (NeurIPS 2021) Fig. 5: One can construct an "evil" loss function that works against CKA similarity so that CKA similarity decreases significantly during training yet performance measures stay constant

Stitching, on the other hand, directly optimizes for a given function (e.g. accuracy) so it has a cleaner semantic interpretation.

One-to-one mapping: Although not directly relevant to our paper, this is still a very interesting issue. In our understanding, the stitching methodology does not strictly require a one-to-one mapping as a stitching layer. Theoretically, an appropriate stitching layer can be defined between any two abstract spaces, even if they are of different dimensionality, and even if they are structured spaces with different structures. We agree that this is an unexplored area regarding efficiency. We are optimistic, though.

Official Review of Submission4440 by Reviewer rEdH

Official Review  Reviewer rEdH  27 Feb 2023, 23:55 (modified: 14 Mar 2023, 14:39)

 Program Chairs, Area Chairs, Authors, Reviewer rEdH, Reviewers Submitted, Senior Area Chairs

Summary:

The authors investigate the effect of affine 1x1 convolutional stitching layers on natural and robust accuracy. The component models considered are trained in a robust or non-robust fashion. The main findings are that natural accuracy is generally preserved while robust accuracy generally becomes destroyed when stitching any combination of models. An improved robust-accuracy-preserving stitching strategy is considered and some conjectures are posed about properties of last-layer representations.

Strengths And Weaknesses:

Strengths:

- The experimentation is very thorough with all robust/nonrobust stitchings explored and displayed in easy-to-read plots.
- The results are fairly surprising and demonstrate a nice empirical insight. In particular, the retention of robust accuracy of models which have robust prefixes/fronts is surprising.

Weaknesses:

- The sheer number of experiments can be a little confusing. There are many combinations of models to consider, and the repetitive plots make it challenging to remember the insights drawn from each. While I can't see any immediate ways to remedy this, the paper does read like a laundry list of experiments.
- Not a lot of novelty in the methodology here: the only novel contribution is the stitching of robust/non-robust models and the slightly nonstandard formulation of the robust loss in Eq 2

Questions:

- For instances where it seems the stitching has failed (i.e. robustly stitching two naturally trained networks as in figure 5), is it possible to overcome the matching failure with a stronger stitching layer? i.e., it's very interesting that even trying the robust stitching procedure is enough
- Along the above lines, what is occurring on the non-monotonic layers in the upper-middle plot of figure 5?

Limitations:

The authors are up-front about the large amounts of GPU time needed to generate the results in this paper, but don't mention any societal impacts (which I imagine would be benign anyway).

Ethics Flag: No

Soundness: 3 good

Presentation: 3 good

Contribution: 3 good

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Confidence: 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.

Code Of Conduct: Yes



Rebuttal by Authors

Rebuttal

Authors (András Balogh (/profile?id=~András_Balogh1), Márk Jelasity (/profile?id=~Márk_Jelasity1))

18 Mar 2023, 22:03 (modified: 20 Mar 2023, 15:40)

Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Revisions (/revisions?id=3PT4U5GJ3Y)

Rebuttal:

We thank the reviewer for the thoughtful questions and comments.

Novelty of the methodology: Indeed, stitching in itself is not novel. We do see some novelty, though. Stitching always requires the specification of a function, according to which similarity is investigated. This has not been stressed before, as clean accuracy has always been the function of interest. However, this flexibility of stitching allows one to investigate the relationship between different functions (for example, whether similarity according to one function implies similarity according to another one, and so on). We use this approach in the context of two functions: clean accuracy and robust accuracy, but this can be applied more generally, as pointed out also by Reviewer B1YL

Stronger stitching layer: We found this an interesting question so we quickly investigated this by using a two-layer stitching in which we apply a ReLU or GeLU nonlinearity after the 1x1 convolution and then we add another 1x1 convolution. The results are the same for the few layers we had time to check. This is an exciting future direction, namely, we could characterize the similarity also by specifying how rich the stitching has to be to have the capacity to replicate the original quality of the stitched networks. Sometimes the information might simply not be there at all, sometimes a stronger stitching layer might be needed. Using only linear stitching layers cannot differentiate between these two cases.

Non-monotonic layers: We have been looking into this question, but unfortunately, we do not yet have a complete answer. We do see this pattern consistently (in our new results with larger architectures, and also Fig 11 in the appendix). One clue is that this happens only if 100% adversarial samples are used. In this case, there is a large difference between blocks 7 and 8 (or in larger networks, there is a local maximum before the last block). If 50% clean samples are used, then the stitching that we get essentially becomes clean stitching, and the adversarial samples are essentially ignored. So the non-monotonic behavior has to do with the stitching fighting with a practically impossible task resulting in different failure modes in different layers (getting no robustness or "cheating" by creating a degenerate network). Note also that adversarial attacks have the strongest effect on the last layers (as was shown by Cianfarani et al in 2022), which might also play a part.

[About OpenReview \(/about\)](/about)

[Hosting a Venue \(/group?id=OpenReview.net/Support\)](/group?id=OpenReview.net/Support)

[All Venues \(/venues\)](/venues)

[Frequently Asked Questions](#)

<https://docs.openreview.net/getting-started/frequently-asked-questions>

[Contact \(/contact\)](/contact)

[Sponsors \(/sponsors\)](/sponsors)[Feedback](#)[Terms of Service \(/legal/terms\)](/legal/terms)[Privacy Policy \(/legal/privacy\)](/legal/privacy)

[OpenReview \(/about\)](/about) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](/sponsors).