B-Cos Alignment for CNNs and Vision Transformers with additional extensions

Kian Hajireza
hajireza@kth.se

Martín Bravo
mebd@kth.se

Julian Alcibar Zubillaga
julianaz@kth.se

Project

- We chose the "*B-Cos Alignment for Inherently Interpretable CNNs and Vision Transformers*" which is part of the "*Interpreting and Understanding Deep Networks*" topic.
- The grade that we are aiming for is A or B.
- The project title is B-Cos Alignment for CNNs and Vision Transformers with additional extensions

Project Plan

Experiments

The main milestones of this project are:

- Implementing the paper.
- Perform additional studies that are not in the paper in regards to machine learning research best practices.
- If time permits, we aim to apply the B-Cos implementation for regression tasks.

Regarding implementing the paper task the sub milestones are to:

- For the tests with CIFAR-10 we plan to implement the B-Cos extension for a ResNet-56.
- For the tests with ImageNet we plan to implement the B-Cos extension for different types of convolutional neural networks such as VGGs, ResNet, DenseNets and ResNext and also Vision Transformers. To explore the explanation for the model decision of the vision transformers, we will use the grid-pointing game
- Evaluate the interpretability of the B-Cos extension using localization scores and compare to the benchmark which in this case are other post-hoc explanation methods. Due to limited time we thought of studying only LIME, GradCAM, DeepLIFT, Grad and IntGrad.
- For evaluating the performance of the implementations we will be constructing a benchmark in conjunction with ablation studies to test the effect of the B-Cos extension. Note that the benchmark will be without bias terms in the convolution and normalization layers to isolate the effect of the B-Cos extension. The performance will be evaluated through the test accuracies the models achieve.

Note that the list above should only result in a grade E.
Submilestones in regards to best machine learning research practices are listed below:

- It is not clear based on the paper and its supplemental information what split they used for their training, validation, and test set. Therefore we will be implementing cross-validation for improved reproducibility.
- Additionally, it is not clear what the probability distribution of their model accuracies looks like. Therefore we will be looking to complement the study with more information such as standard deviation or perhaps even violin plots.
-
- The paper did not investigate any statistical significance regarding the accuracies of their proposed method compared to their benchmarks. Therefore we will be using the student's t-test to determine the statistical significance of the results.
- In the article, the authors tested the accuracy of their model for CIFAR-10 for different B values. A test that we are interested in doing is adding the parameter B as a trainable parameter. We believe that this would have to be formulated as a minimax optimization problem, where the maximum value of the parameter B is minimized.
- If time allows, an additional test that we are interested in doing is to implement an implementation of the B-Cos in the context of neural ODEs for a classification task and use a vanilla neural ODE as a benchmark.

By completing the top 3 items in the list above we believe it should motivate a grade C. Completing all of the items listed above would motivate a grade A-C.

Regarding the regression task, the sub-milestones are to:

- Implement B-Cos for a neural network that is suitable for regression.
- Test the B-Cos implementation for neural ODEs.
- Perform regression using the B-Cos implementation for different physical systems such as Lotka-Volterra, chemical reaction system, and Lorenz system.
- Implement a benchmark in ablation studies to determine the effect of the B-Cos implementation.

Datasets

We will use CIFAR-10 and ImageNet as the datasets for the classification tasks. For the potential regression task, we aim to generate the data synthetically for different physical systems such as Lotka-Volterra, chemical reaction systems or Lorenz system.

Benchmarking with Standard Architectures

We will use vanilla CNNs and Vision Transformers using VGG, ResNet-56, and DenseNet architectures on both datasets as the benchmark.

Computational Requirements

The computational requirements for executing our plan will involve training deep learning models on datasets like CIFAR-10 and ImageNet and working with synthetic data for scientific applications such as Neural ODEs. Given these needs, we will utilize Google Colab Pro to access GPUs and optimize training time.

Given all the tasks we will have to do, our requirements for this project are the following:

1. **GPU:** Given that we will use deep learning models, these are quite intensive. Therefore, we will be required to use a high-performance GPU, given the availability of Google Colab Pro, we will need a Tesla T4 or P100.

2. **RAM:** Handling Large Datasets like ImageNet will require performing memory-intensive operations. We will need at least 16Gb.

3. **Storage:** Store models and datasets will require at least 100Gb.

Deep Learning Software Packages

Our deep learning framework will primarily be based on PyTorch and related libraries, including `pytorch-lightning` ($\geq$ 1.8.0,$<$ 2.0.0) for training management, `torch` ($\geq$ 1.13), `torchvision` for dataset handling, `torchmetrics` ($\geq$ 0.11.0) for evaluation, `einops` for tensor operations, `torchdiffeq` for solving Neural ODEs.

How will we measure the success of the projects?

The success of our project will be measured by the ability to effectively extend the B-Cos alignment methodology to classification and regression tasks, assessing its impact on **model performance**, **interpretability**, and **generalization**. Key metrics will include improvements in **accuracy, precision, recall, and loss** and the ability to produce meaningful **feature attributions and explanations**. We will perform extensive ablation studies, statistical significance testing, and cross-validation to ensure robust results.

Skills/Knowledge Goals

- **Julian:** I have worked with transformers and CNNs, however, I have never worked with Neural ODEs, so in this project, I seek to learn a little about these types of architectures and how they differ from what I already know and to be able to apply the concepts I learned about interpretability and understanding deep neural networks.

- **Kian:** I aim to learn more about reproducibility when it comes to deep learning. I don't have a lot of experience with CNNs and transformers, so learning more about those types of networks is also something I aim to do. Lastly, I am furthering my cooperation skills and working with version control systems like Git.

- **Martín:** I have worked as a Machine Learning Engineer in the industry, successfully implementing various machine learning models, and enhancing my problem-solving and coding skills. Through this project, I aim to deepen my understanding of B-Cos Alignment and its applications in both classification and regression tasks.

Why do we deserve this grade?

As we can see, our experiment involves several steps that will allow us to implement the B-Cos method effectively. By extending it to regression tasks and applying it to complex models such as neural ODEs we will contribute valuable insights to the field. These extensions will not only expand the boundaries of the scope of the original paper but will also address key gaps such as the lack of statistical significance testing and cross-validation. The added complexity of incorporating hyperparameter optimization, benchmarking against standard architectures, and performing ablation studies will deepen our understanding of the capabilities and limitations of the method. This comprehensive approach ensures that our project will significantly advance research, justifying an A or B grade.