Learning Concepts using Deep Neural Networks

Sunil Kumar Vengalil,[1] Neelam Sinha[2] and G Srinivasaraghavan,[3]

[1]*International Institute of Information Technology Bangalore*

[2]*International Institute of Information Technology Bangalore*

[3]*International Institute of Information Technology Bangalore*

Abstract

Keywords: Machine Learning, Artificial Intelligence, Computer Science
DOI: 10.2018/JMLFS000001

1. INTRODUCTION

With the advent of Deep learning in the past decade, Machine Learning and Deep Learning models have changed their role from a "merely research idea" to an essential component in almost all domains ranging from Banking Finance and Insurance Sector(BFIS), retail, manufacturing to medical[1] and healthcare. However, one of the major pain points of deep learning models, when used in industry, is the lack of explainability, i.e., they are unable to provide the exact reason for prediction. One the one hand, researchers build complex ensemble and deep architectures in order to increase the classification accuracy, which results in increased accuracy but the more complex the model is, the more difficult to provide explanations for the predictions. This trade off between accuracy and explainability of Machine Learning and Deep Learning models is well known [12]. In many use cases, like medical, customer churn prediction in Banking Finance and Insurance Sector(BFIS), fraud and spam detection etc, it is as equally important to get an explainable prediction as getting correct predictions. An explainable prediction can be easily converted to an action item. Sometimes one might even want to compromise with a less accurate model, if it can provide better explanations for the predictions. This behaviour of Machine Learning and Deep Learning models is in sharp contrast with how humans learn.

Learning in human possess the following special characteristics which the ML and DL models starve to achieve.

1. When humans learns something new, over a period of time they gradually 1)internalize the concepts 2) abstracts the concepts and 3) reuse the concepts in related tasks- all these accomplished usually from a very few samples [14].
2. The learned concepts are more generic as opposed the predictions, they make using the concept, in a particular scenario. Each scenario might be different, but the concept used to make an inference will be same. In other words, they use the same concept to make predictions for different input scenarios i.e the mapping between concept and scenario is one to many.
3. Humans can always provide an explanation for what they predict using the concepts they learned (Irrespective of the fact that the prediction might be wrong).

In this article, we explore how such behaviours can be incorporated into deep learning models. Explainability in DL models is something which has been actively researched and many approaches have been suggested in the literature in the last decade[11]. See section 2.1 for an extensive literature survey of explainability on deep learning models. However, our objective in this work is not just to inject explainability into a trained deep learning model. We rather focus on how the underlying model can learn **generalized concepts** which are used to make predictions. There have been previous studies [14] on developing models that learn concepts and use the learned concepts in a human-like manner using Bayesian Probabilistic Language [15].

In this study, we report results of experiments performed on image datasets of varying complexity starting from MNIST, CIFAR-10, CIFAR-100 and Marmot (a dataset to detect tabular regions in images of documents) [13] datasets. We propose an approach for learning tasks like classification and segmentation while the model also learns the concepts associated with the dataset/task.

Our approach is to modify the training procedure of deep learning models so that the model not only learns to perform a specific task, say for example image classification, but also learns a set of **generic modularized concepts** and **composition rules or operators** with the following objectives:

- The model's final prediction can be explained using the concepts and operators learned making the model more interpretable (i.e model can provide explanations/reasons for its prediction in a "human understandable" fashion).
- New concepts can be generated by applying operators on the learned concepts.
- A task like image classification or segmentation, which is solved using supervised methods today and hence necessitates huge number of (typically in millions for images) annotated data, can be solved in semi-supervised fashion requiring very less human annotation effort.
- We anticipate to accelerate the training process as the concepts are learned prior to (or during) learning the final classification task.

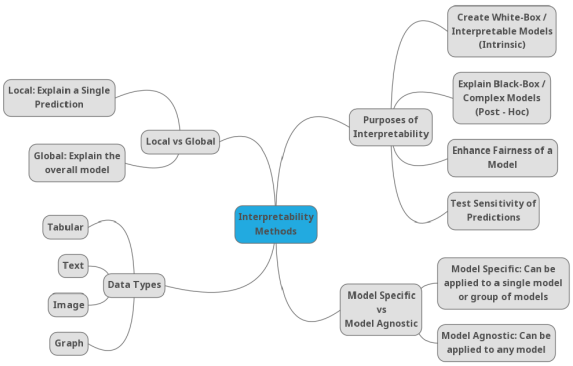

FIGURE 1: Taxonomies of Explainable AI techniques. Image taken from [11]

- The learned concepts can be re-used for other related tasks. This is in contrast with current transfer learning approaches where the model parameters that are being shared across tasks need not be **modularized**.

The gist of our approach is augmenting the training set with primitive concepts associated with data and introducing losses in hidden layers during training so that the predictions, and the hidden layer feature maps that lead to the final predictions, are more explainable and re-usable. Many existing algorithms [2] [3] [4] for building explainability into deep learning models are applied after the model is fully trained. However, our approach differs significantly as we are modifying the loss function and training data for incorporating explainability. In addition to solving the primary task of explainability, the approach also benefits from the fact that the model can be trained with a smaller number of epochs. This is because all the hidden layers, where the losses are introduced, start learning the features in parallel, whereas in existing architectures training losses are computed only at the output layer and it takes a larger number of epochs for the initial layers to learn useful features. Further, since the loss functions can be introduced at any layer directly, the approach will not suffer from vanishing gradient.

Even though our approach can be applied to any deep learning models, we pick a widely used generative model, Variational Autoencoder [18] in order to illustrate our approach. Our work is motivated by the concept of learning based on visual concepts introduced by Lake et.al. in 2015 [14]. We generate visual concepts, similar to those used in [14] and use the images of generated concepts for training a variational autoencoder, whereas in [14] the concepts are used to build a probabilistic prediction framework.

The rest of the document is organized as follows. 2 summarizes existing work on explainability, concept learning and semi-supervised learning. The proposed approach is detailed in 3. Section 4 detailed out our experimental results and inferences. Finally we summarize our result and inferences in section 5

2. RELATED WORK

2.1. Explainability in Deep Learning Models

An excellent review of existing Explainable AI techniques can be found in [11].

Adding explainability to deep neural networks has been one of the key focus areas of the deep learning research community in the last decade. There are many different taxonomies for Explainable AI(XAI) which is shown in figure 1.

One such classification is based on model-agnostic (which can be applied to any model) versus model-dependent where the approach depends on the specific prediction algorithms used by the model. The early works on this includes the use of explanation vectors introduced by David Baehrens et.al. in their paper [5]. Local Interpretable Model-Agnostic Explanations (LIME) [2] is one of the most widely used tools to explain the predictions of any machine learning model. LIME works by approximating the model behaviour locally by a simple interpretable surrogate model like linear model or decision tree. The approximation is restricted to a small local change in the input. As this technique looks only at the input and output of the model, this technique can be applied to any machine learning models. However, the main drawback of the approach is that the surrogate model is only a local approximation and fails to capture the global behaviour. Rabold et. al. in their recent work [6] builds on top of LIME and suggests an approach where classifiers' decisions can be explained in terms of logic rules. Their approach is unique in the sense that the predictions can be explained using relationships of objects within the image, whereas in most other approaches the decision is based on presence or absence of certain features.

Another famous technique called Layer-wise Relevance Propagation was introduced by Bach et.al. [7] where the classifier decision is back propagated and a relevance score is computed at each of the layers backwards until the input units are assigned a relevance score. The work by Wojciech Samek et.al. [8] provides a quantitative evaluation of Layer-wise Relevance Propagation and suggests it as a better model as opposed to sensitivity based methods.

Scott et.al. introduced a unified approach for explainability which combines multiple methods and brings in a new class of additive feature importance measure (called SHapley) [3]. Their method, known as SHapley Additive ExPlanations (SHAP) considers all the interactions between features and provides an average measure of importance for each feature.

Another class of widely used algorithms for Convolutional Neural Networks targets at finding and highlighting the regions in the input image that is responsible for the prediction. These techniques, originally introduced in paper [4], generate a heatmap

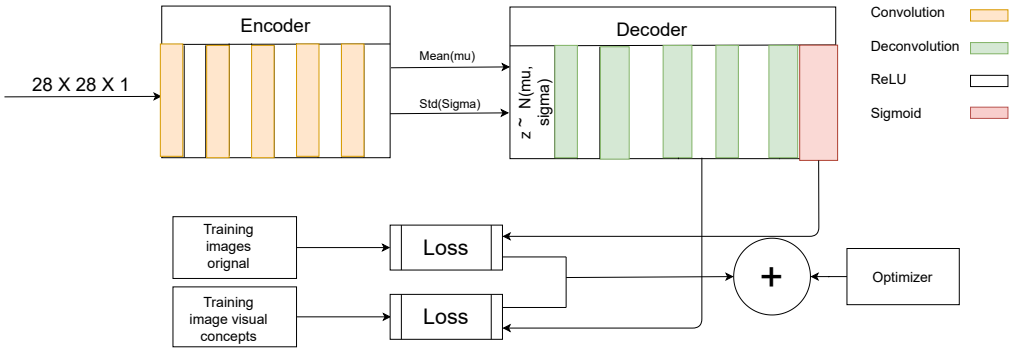

FIGURE 2: Architecture of the proposed model for MNIST dataset. An additional loss component were added at layer 3 of decoder in Variational Autoencoder.

called Class Activation Map that will highlight the most sensitive regions in the input image. Saliency Map, introduced by Symonyan et.al. in 2014 [9] is another variant of class activation map computed by optimizing the input image using the gradient of output prediction with respect to the input image.

Another criteria for classifying XAI techniques is based on the data type used like text, images, structured data (a specific case of which is tabular data) graph etc. TODO add some example cases

Another criteria for classifying XAI techniques is based on the data type used like text, images, structured data (a specific case of which is tabular data) graph etc.

Yet another way of classifying explainability approaches is based on whether the technique is applicable for individual instances or for the whole model. TODO add some examples

2.2. Concept Learning

2.3. Semi-supervised Learning

3. PROPOSED METHOD

3.1. Overview of Approach

We start with a completely unlabelled dataset and solve a task like image classification or segmentation on this dataset by generating concepts and operators required to solve the task at hand. However, the proposed approach is not completely unsupervised. Instead, we propose an iterative algorithm where human feedback is taken interactively and the model is improved based on the feedback. In this respect our approach is comparable to semi-supervised learning and active learning (write differences. active learning the goal is only to minimize manual annotation where as here there is an additional goal of adding explainability and also generalizing the learned concepts)

The proposed approach modifies the training procedure of deep learning models so that the model not only learns to perform the specific task, but also learns a set of **generic modularized concepts** and **composition rules or operators** with the following objectives:

- The models final prediction can be explained using the concepts and operators learned.
- A task like image classification or segmentation, which is solved using supervised methods today, can be solved in semi-supervised fashion requiring very less annotated samples.
- We anticipate to accelerate the training process as the concepts are learned prior to (or while) learning the final classification task.
- The learned concepts can be used for other related task. This is in contrast with current transfer learning approaches where the model parameters that is being shared across tasks need not the modularized

Figure 2 shows the architecture of the proposed model used for the MNIST dataset. For other more complex datasets, the overall structure remains the same except that more number of layers were added in encoder and decoder along with Batch Normalization [19] and regularization. For MNIST, we use a variational autoencoder (VAE) with five layers of convolutional layers in the encoder and five layers of deconvolution layers in the decoder. All layers, including the middle layer that generates the latent images, are fully convolutional. We replaced the middle dense layer of in VAE, with a convolutional layer in order to create a latent image (as opposed to a one dimensional latent vector) that preserves the spatial relationship. We added an additional loss component at the hidden layer 3 of the decoder in order to force this layer to learn the primitive visual concepts of the dataset. The final layer of the decoder has sigmoid activation function and we use binary cross entropy loss as the MNIST images can be treated as binary images.

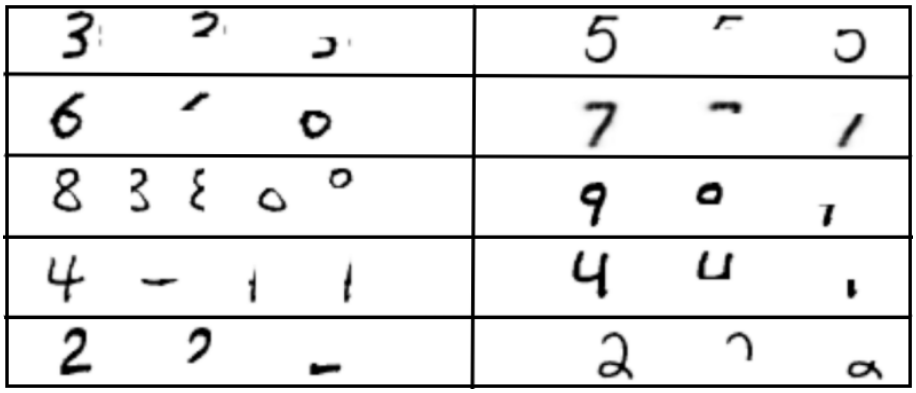

FIGURE 3: Visual concepts generated from each of the digits in MNIST. Since there are two different types of images for digits 4 and 2 separate concepts were generated for each image. The width, height and location of these segments were used to define the mean value of a Normal distribution used for generating more samples.

3.2. Datasets

The proposed approach is demonstrated on datasets of increasing complexity starting from MNIST, CIFAR-10, CIFAR-100, Marmot [13] dataset for table detection. For adding the concept loss, the MNIST training set was first augmented by adding 3000 training images for each of the 18 concepts shown in Figure 3. Some of the similar looking concepts in Figure 3, like the horizontal line segments and vertical line segments, were combined into one group to form 18 unique visual concepts. The steps for generating concepts are detailed in the sections below.

3.3. Primary Visual Concepts

Generating Visual Concepts We augment the training set using images of randomly generated segments from MNIST images. For each of the visual concepts shown in Figure 3, 3000 images were generated by randomly sampling the height, width and location of the segment from a normal distribution. For each concept, the mean value of the height, width and location is kept the same as in Figure 3. The number of visual concepts and their corresponding distributions are hyper parameters that will vary from dataset to dataset. To generate the segments for each concept, we sampled heights, widths and location of top left corner (represented by two distributions for x and y coordinates) from the respective concept's distributions and generated slices from randomly chosen training images. Standard deviation for all distributions were kept as 1 pixel in our experiments.

3.4. Composition of concepts

3.5. Network Architecture, Loss function and Training

4. RESULTS AND DISCUSSION

5. CONCLUSION

ACKNOWLEDGEMENTS

The work of Q.H.C. is supported in part by the U. S. Dept. of Energy Grant No. DE-AC02-06CH11357 and in part by the Argonne National Lab. and Univ. of Chicago Joint Theory Institute Grant No. 03921-07-137. The work of S.K. and H.O. is supported in part by the Science and Technology Development Fund (STDF) Project ID 437 and the ICTP Project ID 30. The work of E.M. is supported in part by the U. S. Dept. of Energy Grant No. DE-FG03-94ER40837.

References

[1] Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349.6245 (2015): 255-260.
[2] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "" Why should i trust you?" Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016.
[3] Lundberg, Scott M., and Su-In Lee. "A unified approach to interpreting model predictions." Proceedings of the 31st international conference on neural information processing systems. 2017.
[4] Zhou, Bolei, et al. "Learning deep features for discriminative localization." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
[5] Baehrens, David, et al. "How to explain individual classification decisions." The Journal of Machine Learning Research 11 (2010): 1803-1831.
[6] Rabold, Johannes, et al. "Enriching Visual with Verbal Explanations for Relational Concepts–Combining LIME with Aleph." arXiv preprint arXiv:1910.01837 (2019).
[7] Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." PloS one 10.7 (2015): e0130140.
[8] Samek, Wojciech, et al. "Evaluating the visualization of what a deep neural network has learned." IEEE transactions on neural networks and learning systems 28.11 (2016): 2660-2673.

[9] Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv preprint arXiv:1312.6034 (2013).
[10] Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
[11] Linardatos, Pantelis, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. "Explainable AI: A review of machine learning interpretability methods." Entropy 23.1 (2021): 18.
[12] Wu, Leihong, et al. "Trade-off Predictivity and Explainability for Machine-Learning Powered Predictive Toxicology: An in-Depth Investigation with Tox21 Data Sets." Chemical Research in Toxicology 34.2 (2021): 541-549.
[13] Fang, Jing, et al. "Dataset, ground-truth and performance metrics for table detection evaluation." 2012 10th IAPR International Workshop on Document Analysis Systems. IEEE, 2012.
[14] Lake, Brenden M., Ruslan Salakhutdinov, and Joshua B. Tenenbaum. "Human-level concept learning through probabilistic program induction." Science 350.6266 (2015): 1332-1338.
[15] Tenenbaum, Josh. Bayesian Program Learning and Concept Induction. Massachusetts Institute of Technology Cambridge United States, 2019.
[16] Nanayakkara, Shane, et al. "Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study." PLoS medicine 15.11 (2018): e1002709.
[17] Tim Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Volume 267, 2019, Pages 1-38, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2018.07.007. (https://www.sciencedirect.com/science/article/pii/S0004370218305988)
[18] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
[19] Ioffe, Sergey, and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift." International conference on machine learning. PMLR, 2015.