An approach to explainable deep learning using fuzzy inference

Conference Paper · May 2017

DOI: 10.1117/12.2268001

David Bonanno
United States Naval Research Laboratory

Leslie Smith
United States Naval Research Laboratory

Paul Andrew Elmore
Johns Hopkins University / Applied Physics Laboratory

Frederick E. Petry
United States Naval Research Laboratory

An approach to explainable deep learning using fuzzy inference

David Bonanno*[a], Kristen Nock[a], Leslie Smith [a], Paul Elmore[b], Fred Petry[b]
[a]U.S. Naval Research Laboratory, 4555 Overlook Avenue SW, Washington, D.C., USA 20375
[b]U.S. Naval Research Laboratory, 1005 Balch Boulevard, Stennis Space Center, MS, USA 39529

ABSTRACT

Deep Learning has proven to be an effective method for making highly accurate predictions from complex data sources. Convolutional neural networks continue to dominate image classification problems and recursive neural networks have proven their utility in caption generation and language translations. While these approaches are powerful, they do not offer explanation for how the output is generated. Without understanding how deep learning arrives at a solution there is no guarantee that these networks will transition from controlled laboratory environments to fieldable systems. This paper presents an approach for incorporating such rule based methodology into neural networks by embedding fuzzy inference systems into deep learning networks.

Keywords: Deep Learning, Explainable AI, Fuzzy Inference, Sensor Fusion

1. INTRODUCTION

As autonomous systems continue to develop so too will their responsibilities. Currently, autonomous systems are operating to collect a tremendous wealth of information gleaned from multiple sensing modalities. The vastness of this data makes it prohibitively difficult to exploit; often times teams of people are required to manually process, interpret, and exploit the information that has been collected. As this method is often costly, both in terms of time and monetary constraints, it is popular to turn towards machine learning tools to enhance the exploitation process.

Many machine learning tools can be effectively wielded by analysts to reduce the computational burden on an analyst. One such machine learning tool, deep learning (DL), looks to encode complex mathematical representations using stacked auto encoders. One popular algorithm within DL is convolutional neural networks (CNNs), which have proven their utility in object classification [8] and detection [13] within imagery. CNNs can be used to process large amounts of data and sort out what and where an object is located without requiring analysts to manually sort through the imagery. While this is a useful tool, there exists a breakdown in communication between the operator and the CNN. The CNN is able to accurately generate a classification label but does not necessarily report on features that were present allowing a classification to be inferred. For example, a CNN may be able to correctly identify an object as being a 'cat' but not have any representation of 'whiskers' or 'fur.' Similarly, the analyst would not be able to communicate the importance of a specific feature or trait to the CNN which limits the amount and nature of feedback from an analyst to a CNN.

This problem fundamentally limits the utility of such tools. Without understanding how a CNN arrives at a solution, it is impossible to understand how adaptable the system is. This poses a complex challenge for many autonomous systems, as many of these machine learning tools are developed with controlled imagery and trained on labeled data. However, when an autonomous system is deployed the imagery may fundamentally change and there are no guarantees that the machine learning tools will operate effectively given these changes.

By introducing an additional machine learning process, fuzzy inference, an explainable rule-based structure can be realized in DL alleviating these problems. Fuzzy inference processes allow an analyst to generate rule-based structures. By creating these rules it is possible for an analyst to bias features generated from DL providing feedback to the system.

Next-Generation Analyst V, edited by Timothy P. Hanratty, James Llinas, Proc. of SPIE Vol. 10207, 102070D · © 2017 SPIE · CCC code: 0277-786X/17/$18 · doi: 10.1117/12.2268001

Additionally, through these rule-based structures, an analyst can be easily understand how a decision has been made by the system.

2. MOTIVATION

It is increasingly necessary for explanations to be communicated effectively both to and from autonomous systems. These explanations can offer insight into why an action has been chosen, or why a specified object has been given a specific classification label. The latter example can be realized by including rule-based fuzzy inference systems with expressive DL tools. The DL tools can learn to generate information dense feature labels which are then further interpreted by the fuzzy inference systems offering both a label and an explanation. Because the fuzzy inference systems depend on human defined structured rules, the system would be both easily to understand and easy to modify.

2.1 Fuzzy Inference

Fuzzy-inference systems, such as ANFIS [5] allow for complex non-linear problems to be approximated using if-then statements. These systems have a wide variety of applications and can encode both objective measurements and subjective information. Such systems have been used for a wide array of applications, including multimodal classification [12], medical imaging [4], and market prediction [2].

Structured rule based systems have the advantage of being able to be biased by subjective information. This provides an opportunity for an analyst to provide expert information to the system, improving classification results or changing the behavior of the system. This feedback bias can additionally be used to speed up learning for the autonomous systems while maintaining stability [11].

2.2 Deep Learning

Unlike the fuzzy-inference counterparts, deep learning methods including CNNs and Recurrent Neural Networks (RNNs) do not require hand crafted features. Instead, they rely on predefined structures and labeled data and are able to learn features on their own. DL methods have been successful in complex sensor problems including fine grained image classification [8], speech recognition [3] and action recognition [1].

A remaining challenge for these DL methods involves extracting and interpreting the features that are learned by the networks. For many applications a classification label is not enough information; in these instances it is important to additionally be able to understand the high level features that have been generated by the networks. Preliminary work has shown that the networks do create, at some level, neurons which can be used to represent higher level abstractions including human and cat detectors, even when the networks are not trained to do so explicitly [9].

Features generated in this fashion are important for two reasons. First, these features are generated in an unsupervised system freeing the system from the burdens of having hand crafted labels. The problem of hand labeling data has often stalled development of DL systems as generating tens of thousands of training examples is tedious and difficult work. Second, having the ability to find specific neurons in a DL system which correspond to a specific requested feature allow an analyst to hook into the network and pull out feature arrays. Instead of only having a class label for 'car,' as is traditionally done in DL, it would be possible to find specific neurons for components of a car ('windshield', 'trunk', etc.). These features can be exploited directly by an analyst or fed into fuzzy-inference systems where they can automatically be grouped together to generate classification labels through rule based mechanisms.

2.3 Hybrid Learning Approaches

Having multiple machine learning algorithms be developed for a problem is by no means an original concept. Deep learning has been coupled with many algorithms including random forest algorithms [10]. Here, high level features are generated through DL and fed into the random forest algorithm allowing information from multiple sensors to be merged in a meaningful way. DL has also been expressed as hierarchical structures [6] which has extended classification problems to even finer grains of classification. DL has even been combined with fuzzy logic for multiple instance problems allowing hand crafted rules to be applied across an image [7].

We look to extend these lessons further by allowing the DL to operate unsupervised and to select the desirable features as defined by fuzzy-inference systems. This will allow robust systems to be developed using rich information from DL even when hand labeled data is unavailable. These features will then be interpreted by fuzzy-inference systems offering explanations for why a classification label has been selected by the system.

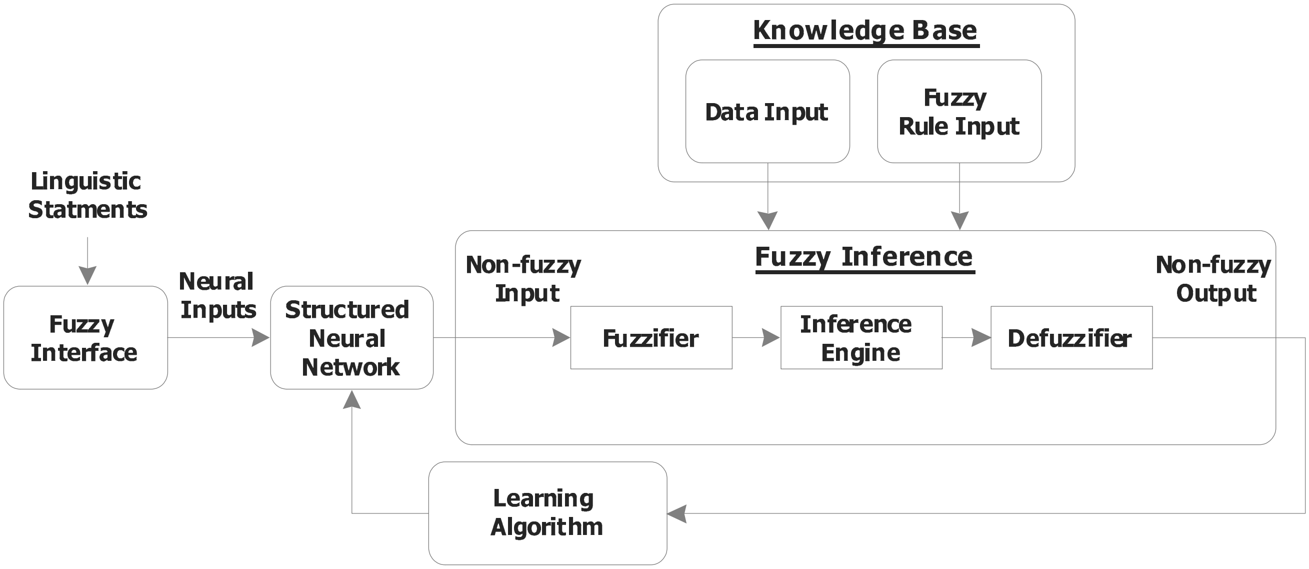

Figure 1. User provided linguistic statements are translated into symbolic representations that can be used to initialize the neural network. These can manifest in parameters like the number of neurons per layer and number of layers in the overall neural network architecture. From there, the neural network drives the fuzzy inference block. Other inputs include the fuzzy rules which can be adjusted to incorporate expert knowledge on the input data. A benefit to this system is that these fuzzy rules are used to explain the behavior of the fuzzy system. Note however that expert knowledge is not required, as the network will be set up to automatically extract fuzzy rules from the numerical data.

3. TECHNICAL APPROACH

A hypothetical system (as seen in Figure 1) can be created using two components. The first is deep learning feature generation which can be used to create representative features from sensor data directly. The deep learning system would initially be trained on unlabeled data; desirable features would be extracted using methods similar to [9].

Once these features are extracted from the deep learning system, they will be integrated into fuzzy-inference systems. These systems can incorporate both the features detected from the deep learning as well as subjective information from an analysts as a method of biasing the system. These two pieces together can be used for classification purposes. The final system would therefore be able to report both classification results and the specific features and rules that were activated for the system to arrive at its conclusion. Additionally, the final system could be further biased by an analyst as a form of feedback.

3.1 Deep learning feature generation

A deep learning architecture can be defined to handle raw data including either imagery or one dimensional signals, depending on the needs of an analyst. The weights can then iteratively be learned in an unsupervised manner without the need for the data to be labeled. This process will generate high level features even without labeled data, as shown in [9].

Once these networks are trained, the activation of each neuron in the network can be tested for its responsiveness to labeled stimuli, both positive and negative. A single input can have multiple labeled features present. This validation process can also be used to further fine tune neurons in the network through back propagation. The responsive neurons can be used as classifiers for the desirable features and fed into a fuzzy-inference system.

3.2 Integration with Fuzzy Inference systems

The features extracted from that data via DL will be further processed by ANFIS, allowing the system to model human reasoning and also provide a mechanism for biasing the system with feedback from an analyst [5]. In order to train the system, both the feature vector inputs and the correct labeled output will be required.

Such as system would have the flexibility to function either with rules defined by an expert, or optionally generate a reasonable set of rules on its own. Having an expert define the rules is desirable as it may bootstrap the learning process and allow the user of the system to define the features and relationships that are important for a specific application.

This development extends the deep learning algorithms to include fuzzy logic via the extension principal. The introduction of fuzzy logic enables fuzzy aggregation which accounts for both the granularity of the data collection as well as a means to handle incomplete information from each modality. These mechanisms also offer a means for subjective evaluation of the value of each modality. The fuzzy arithmetic will percolate into the deep learning architecture allowing Dempster-Shafer beliefs to be realized.

User provided linguistic statements are translated into symbolic representations that can be used to initialize the neural network. These can manifest in parameters like the number of neurons per layer and number of layers in the overall neural network architecture. From there, the neural network drives the fuzzy inference block. Other inputs include the fuzzy rules which can be adjusted to incorporate expert knowledge on the input data. A benefit to this system is that these fuzzy rules are used to explain the behavior of the fuzzy system. Note however that expert knowledge is not required, as the network will be set up to automatically extract fuzzy rules from the numerical data.

The learning algorithm block analyzes the input pattern provided by the fuzzy inference block. Adjusted system weights are back propagated through the system, automatically adjusting the behavior of the predictors, and allowing the system to adapt over time.

4. FUTURE PLANS

Developing the proposed system will allow effective communication between autonomous systems and analysts. Future research to make this system realizable has two major areas of development. The first will be on effective techniques for accurate feature generation. The major difficulty in realizing any of these systems is overcoming the lack of labeled data. By developing deep learning systems which are able to create features as described in Section 3.1, deep learning will be able to expand into new data limited environments.

The second future area of research involves integrating fuzzy inference systems into deep learning. The proposed research extends the techniques into the deep learning architectures via the extension principal. Here, fuzzy logic will be brought into the physical layers of the network allowing fuzzy logic to be learned within the architecture. Research into learning from data containing epistemic uncertainty will also be

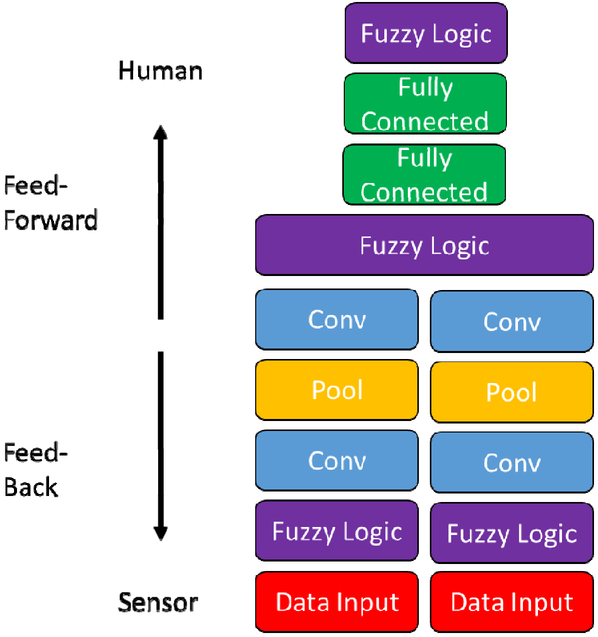

Figure 2. Research will be conducted for the development of sensor fusion techniques between spectral and spatial information obtained from HSI and LIDAR respectively. This approach will use deep learning techniques to allow mid and high level sensor fusion between sensors enhancing the content derived from collected data. This information can be readily used by an analyst or an autonomous system to quickly and reliably make informed decisions

conducted enabling deep learning to be applied to data with missing information. Such a system will minimally require preprocessed data, developed deep learning architectures, and the incorporation of subjective information through fuzzy logic. This work will improve the understanding of the system's strengths and limitations, quantify the uncertainty, and develop methods for potential transitions.

Going forward, these deep learning and fuzzy inference systems will also be developed for multimodal sensor fusion applications. Our systems will improve sensor fusion processing by automatically fusing information using deep learning techniques and fuzzy logic. These techniques will enable amalgamation between multiple sensing modalities, including spectral and spatial information, and will simultaneously increase the reliability of information generated and reduce the amount of time required to extract such information.

Our current approach to the sensor fusion problem (Figure 2) involves incorporating fuzzy inference systems and fuzzy logic throughout the network. This will allow inferences and biases from an analyst to effect the network at all levels; this will allow an analyst to provide feedback on the data, fusion, and overall classification for a sensor fusion system.

5. REFERENCES

[1] Baccouche, Moez, et al. "Sequential deep learning for human action recognition." *International Workshop on Human Behavior Understanding*. Springer Berlin Heidelberg, (2011)

[2] Boyacioglu, Melek Acar, and Derya Avci. "An adaptive network-based fuzzy inference system (ANFIS) for the prediction of stock market return: the case of the Istanbul stock exchange." *Expert Systems with Applications* 37.12 (2010): 7908-7912.

[3] Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, (2013)

[4] Hosseini, Monireh Sheikh, and Maryam Zekri. "Review of medical image classification using the adaptive neuro-fuzzy inference system." *Journal of medical signals and sensors* 2.1 (2012): 49.

[5] Jang, J-SR. "ANFIS: adaptive-network-based fuzzy inference system." *IEEE transactions on systems, man, and cybernetics* 23.3 (1993): 665-685.

[6] Katole, Atul Laxman, et al. "Hierarchical Deep Learning Architecture For 10K Objects Classification." *arXiv preprint arXiv:1509.01951* (2015).

[7] Khalifa, Amine Ben, and Hichem Frigui. "Multiple Instance Fuzzy Inference Neural Networks." *arXiv preprint arXiv:1610.04973* (2016).

[8] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. (2012).

[9] Le, Quoc V. "Building high-level features using large scale unsupervised learning." *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, (2013).

[10] Merentitis, Andreas, and Christian Debes. "Automatic fusion and classification using random forests and features extracted with deep learning." *Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International*. IEEE, (2015).

[11] Nauck, Detlef, Frank Klawonn, and Rudolf Kruse. *Foundations of neuro-fuzzy systems*. John Wiley & Sons, Inc., (1997).

[12] Picón, Artzai, et al. "Fuzzy spectral and spatial feature integration for classification of nonferrous materials in hyperspectral data." *IEEE Transactions on Industrial Informatics* 5.4 (2009): 483-494.

[13] Szegedy, Christian, Alexander Toshev, and Dumitru Erhan. "Deep neural networks for object detection." *Advances in Neural Information Processing Systems*. (2013).