Towards a Face Recognition Model Analyzer

Matthew Johnson, John Angel, Deborah L. McGuinness
Rensselaer Polytechnic Institute
johnsm21@rpi.edu, angelj3@rpi.edu, dlm@cs.rpi.edu

***Abstract*—Machine learning allows computers to learn a model for a given task, such as face recognition, with a high degree of accuracy, using data. However, after these models are generated, they are often treated as black boxes by developers and the limitations of a model are often unknown to end-users. To address these issues, this paper introduces the Face Recognition Model Analyzer (FRMA) ontology and a semantically enabled Result-set viewer. Together these resources describe image features relevant to face recognition and allow users to explore how well a face recognition model does at classifying images that contain an image feature. We evaluated the ontology and Result-set viewer by loading in the Labeled Faces in the Wild [1] dataset, enriching the images with image tags [2], and exploring two popular face recognition models, Facenet [3] and DLib [4]. Using the FRMA ontology and the Result-set viewer, we discovered several classic face recognition model limitations, such as trouble classifying images with occlusions. This evaluation shows that these resources can discover model limitations which can make face recognition model reuse easier for future users.**

***Index Terms*—Ontology, Face Recognition, Machine Learning.**

I. Introduction

Machine learning and the amount of money being spent on machine learning research has exploded over the past ten years. New machine learning techniques are being applied to everything from video game creation, to diagnosing diseases, to better communication between machines and humans. However, one of the limitations of machine learning algorithms is that the learned models are often difficult for humans to understand and are often only evaluated for accuracy on a dataset for a specific task. This form of analysis treats the model like a black box, where only the inputs and outputs for a subset of the solution space are tested, without an understanding of the internal limitations. For many developers of machine learning models, this form of analysis can be sufficient, however, for future users, developers looking to reuse a model for a potentially unanticipated situation, a greater understanding of a model's limitations is needed. Another machine learning limitation is the inability to explain how a model arrives at a prediction, which makes it difficult for users to trust pre-trained models [5]. There are use cases where this is perfectly fine, however, this is unacceptable in scenarios where critical decisions must be made. Because of this, it is often the case that a potential user, making a model for such a scenario, will have insufficient information to make a truly informed decision when choosing which model to use for a new application.

The scope of this problem is quite large, however, this paper focuses on face verification models developed for the Labeled Faces in the Wild dataset (LFW) [1]. We believe that the techniques developed have no limitations that would prevent them from being applied to other supervised machine learning problems and datasets.

The goal of face verification is to determine if the person depicted in two provided images is the same or not. The LFW dataset was chosen because it is a challenging dataset with a large range of variance in pose, lighting, and quality and it is a standard dataset in the face recognition community. At the time of this writing, the LFW website [6] had recorded 76 different models in the unrestricted, labeled outside data results section alone, and many of these have achieved an accuracy of greater than 98%. This large selection of high performing models makes it difficult for future users to compare models for a new dataset or problem domain.

To address these issues, we developed the Face Recognition Model Analyzer (FRMA) ontology[1] and a result-set viewer.[2] The FRMA ontology semantically describes face recognition models, the attributes of the images used to train/test a model, and the predictions generated by a model. The result-set viewer uses the FRMA ontology and allows users to load results and intuitively search image attributes, defined within the ontology, for areas of weakness. The user is then able to explore the strengths and weaknesses of various models from an image attribute perspective using SPARQL queries, or through the result-set viewer, enabling them to better evaluate the effectiveness of a pre-trained model for their new dataset or problem domain.

II. Related Work

There are several research efforts working to improve our understanding of machine learning models. In 2018, Google released their What-If Tool [7]. This tool allows users to visualize their datasets using Facets, edit input data to test What if scenarios, and perform a similar results comparison. Their approach is driven by the model and the input data and is similar to our own approach. The major difference is that our method enriches the input data by using our ontology.

Another avenue of research is machine learning explainability, with the typical goal of explaining how the model arrived at its answer. In [8], researchers compared sensitivity analysis and layer-wise relevance propagation to determine the important pixels in image/human action recognition and words in text document classification. Another group has developed a technique to derive inference rules from the positive and negative examples found in training data [9]. Our approaches are similar in that we both develop an ontology around the

[1] https://tw.rpi.edu/web/Courses/Ontologies/2018/FRMA
[2] https://github.com/FRMA-Ontology/resultset-viewer

978-1-7281-6332-1/20/$31.00 ©2020 IEEE
DOI 10.1109/ICSC.2020.00049

IEEE computer society

input to and the output of machine learning models. We differ in that their focus is on how to explain what is learned, while our ontology is exploring the fitness for model reuse.

III. Resource

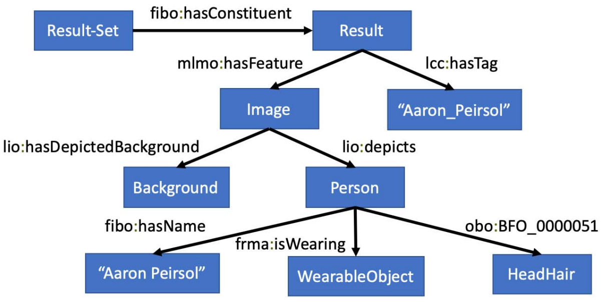
A. Ontology

The goal of the Face Recognition Model Analyzer Ontology (FRMA) is to semantically define the dearth of content within the image rather than merely interpreting the image itself as a flat photograph. This ontology was designed to answer the three following competency questions:

- What type of image attributes does a face recognition model have the most trouble classifying?
- Which of these two face recognition models is better at classifying images of people wearing X?
- Which of these two face recognition models is better at classifying images of people with feature Y?

X represents an object worn by the subject of the image and can include things, such as hats, scarfs, and sunglasses. Y represents a feature of the person in the image and includes static features, such as eye color and dynamic features such as hair style.

The image features within the ontology were driven by the meta-data mined during Kumar's [2] study of Labeled Faces in the Wild. In total, there are 73 different image tags that capture a range of image attributes. From these we developed an ontology that describes these concepts and allows us to infer additional properties, such as facial occlusions. However, because we had so few data points, the ontology can be sparse in certain areas. For instance, there are only three tags regarding haircuts, bald, bangs, and receding hairline, all of which severely limit what we can say about hairstyle.



frma	https://tw.rpi.edu/Courses/Ontologies/2018/FRMA/FRMA/
mlmo	https://tw.rpi.edu/Courses/Ontologies/2018/FRMA/MachineLearningModelOntology/
obo	http://purl.obolibrary.org/obo/
fibo	http://www.omg.org/spec/EDMC-FIBO/FND/Arrangements/Arrangements/
lcc	http://www.omg.org/spec/LCC/Languages/LanguageRepresentation/
lio	http://purl.org/net/lio#

Fig. 1. An overview of a few key classes in the FRMA ontology.

The Face Recognition Model Analyzer Ontology is composed of five sub-ontologies for easier reuse: the image ontology; person, face, and demographic ontology; wearable things ontology; hair ontology; and the machine learning model ontology. The image sub-ontology reuses concepts from the lightweight image ontology [10] to describe general image features and the different pictorial elements depicted in an image, such as background and subject. The person, face, demographic sub-ontology focuses specifically on a person's demographic and facial features that appear in their images, including descriptions such as facial expression, age range, and nose shape. This sub-ontology reuses the Uber-anatomy ontology [11], to describe the different sections of the face, and the FIBO Agents ontology [12], to describe the attributes of a person within the image. The hair sub-ontology aims to more precisely describe the hair on the subject's head and/or face. Because human hair is incredibly varied, the hair ontology focuses on multiple traits to more completely describe a person's hair, including color, texture, and cut. The machine learning model sub-ontology allows users to describe the learning process, the structure of the learned model, and the evaluation of the model from a data-centric perspective. This ontology reuses the FIBO arrangements ontology to describe model components as collections. For example, neural networks are described as a collection of layers: fully connected, pooling, inception, etc. The wearable things sub-ontology describes what people are wearing, whether it's some type of clothing or accessory. In addition, it captures how those pieces of clothing may effectively block a part of, or occlude, someone's face or body.

Figure 1 shows how FRMA ties each of these sub-ontologies together to produce a consistent data model for the result-set viewer. Each algorithm is run against the LFW dataset to produce a Result-Set, which consists of a series of Results that have features and a tag generated by the machine learning model. For face recognition, the feature would be an image and the tag would the predicted person's name. Each image in the LFW dataset has been semantically described using the FRMA ontologies to capture image properties such as background, picture quality, and the people depicted in the images. The people depicted within the image are captured by the Person class, which describes anatomical attributes, what the person's wearing, their hairstyle and ground truth information, such as their name. Each Person class is unique to an image because even if the same person is in multiple photos, features could have changed over time, such as age, haircut, and clothes. In addition, the FRMA ontology also provides some basic inferencing capabilities, such as determining when a piece of clothing becomes an occlusion.

B. Viewer

The result-set viewer allows users to load a result-set, a file generated from the LFW dataset, and explore correct and incorrect classification results from an image feature perspective. This is accomplished by generating a tree of image features from the FRMA ontology where parents are broader features than their children. For example, face occlusion has two children, upper face occlusion and lower face occlusion; upper face occlusion contains auricle, cranial, frontal, nasal, and ocular occlusions. This tree is shown on the left-hand side of the visualization and calculates the sub-accuracy of the result-set overall images that exhibit those features. By examining these statistics, a user could identify specific features that their

TABLE I
THE SEVEN IMAGE FEATURES WITH THE LOWEST ACCURACY FROM DLIB AND FACENET.

DLib		FaceNet	
black and white image	96.34%	sunglasses	91.89%
sunglasses	97.30%	baby	92.31%
balanced lighting variation	97.32%	youth	92.31%
child	97.35%	asian	94.57%
feminine	97.51%	child	95.58%
blurry image fidelity	97.71%	blurry image fidelity	96.73%
hat	97.76%	necklace	96.83%

algorithm performed poorly on. When a user clicks on an element in the tree, the visualization loads all images that exhibit that feature, and highlights images that were correctly classified in every match in green, and images that were incorrectly matched in red. In the top left of the visualization are three buttons: All, Correct, and Incorrect. These buttons act as filters on what images are displayed. All shows all the images, Correct only show images that were correctly classified, and Incorrect, only images that were classified incorrectly.

Behind the scenes, the result-set viewer is driven by a knowledge graph generated for the LFW dataset using Kumar image tags described by the FRMA ontology. When a new result-set is loaded into the visualization the results are integrated into the knowledge graph as a mapping between the two images being compared and how the algorithm classifies them. The visualization uses SPARQL queries against this knowledge graph to create the tree of image features, calculate sub-accuracies, and determine which images should be displayed, depending on the user's interaction.

IV. EVALUATION

We evaluated our system by loading the results of Facenet [3] and DLib [4] into our system and trying to identify features that were troublesome for face verification. We ran both algorithms on the provided training and testing set using the deep funneled LFW images [13], which have been shown to produce superior results for face verification. Using the methodology described in their documentation we achieved an accuracy of 98.1% with Facenet and 98.4% with DLib. Both of these are different than the reported accuracy because we didn't perform the full 10-fold cross-validation, but our methodology could be repeated to explore each fold. We then loaded our result-set into the viewer and began to explore the results.

The result-set viewer identified several features that are known to cause problems with image verification including sunglasses, hats, and blurry images. However, DLib also did poorly with features such as balanced lighting variation, feminine gender, and black/white images, which is a bit surprising. Features such as sunglasses, hats, blurry images, and black/white images make sense, because in all of those cases information about the person is missing via occlusion or poor image quality. However, there is no missing information for balanced lighting variation and feminine gender images. In addition, both algorithms had trouble classifying images that were tagged as being younger, which could be due to a lack of training on images from younger people.

This evaluation shows that, by using the FRMA ontology and the Result-set viewer, users can discover classic face recognition model limitations, such as trouble classifying images with occlusions. Finding and understanding the limitations of these face recognition models will make model reuse easier for future users.

V. CONCLUSION

The goal of this work was to get a better understanding of the limitations of a pre-trained face verification model. We achieved that by developing the FRMA ontology to describe the features of the images used for testing a model and a visualization to explore the results of a given model. Using these resources we analyzed the results from two well-known face recognition models and identified several classic issues with face verification, along with some other interesting shortcomings. This evaluation shows that these resources can properly discover model limitations, which will make face recognition model reuse easier for future users.

VI. FUTURE WORK

There are several avenues this project could be extended to in the future, the most important being refining and extending the FRMA ontology. Currently, the ontology only describes the attributes of images we had tags available for, but in reality, many other features could and should be modeled. In addition, the ontology needs to be extended to other data sets and machine learning tasks. Expanding the ontology to support these additional domains will lead to better representation and improve reusability. In addition, the result-set viewer could be further improved and the effectiveness of the interface verified through a series of user studies.

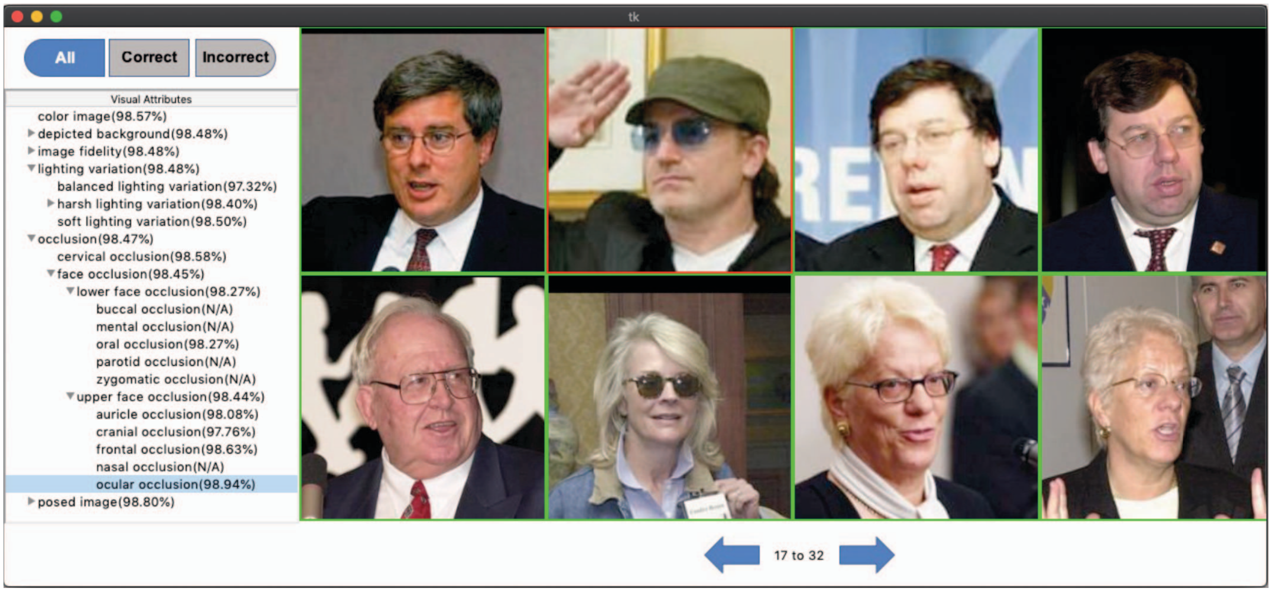
VII. ACKNOWLEDGEMENTS

This work grew out of a class project for Ontologies 6340 at Rensselaer Polytechnic Institute. We would like to thank our original team Jordan Faas-Bush, Andrew Gaudet, and Alexander Schwartzberg, along with our instructors Rebecca Cowan, Elisa Kendall, and Deborah L. McGuinness, who provided insight and expertise that greatly assisted the research. In addition Matthew Johnson acknowledges financial support from the Air Force Research Lab (AFRL) through the Civilian Academic Degree Payment (CADP) program [Case Number 88ABW-2019-5370].

REFERENCES

[1] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[2] N. Kumar, A. Berg, P. N. Belhumeur, and S. Nayar, "Describable visual attributes for face verification and image search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1962–1977, 2011.



Visual Attributes

- color image(98.57%)
- ▸ depicted background(98.48%)
- ▸ image fidelity(98.48%)
- ▾ lighting variation(98.48%)
  - balanced lighting variation(97.32%)
  - ▸ harsh lighting variation(98.40%)
  - soft lighting variation(98.50%)
- ▾ occlusion(98.47%)
  - cervical occlusion(98.58%)
  - ▾ face occlusion(98.45%)
    - ▾ lower face occlusion(98.27%)
      - buccal occlusion(N/A)
      - mental occlusion(N/A)
      - oral occlusion(98.27%)
      - parotid occlusion(N/A)
      - zygomatic occlusion(N/A)
    - ▾ upper face occlusion(98.44%)
      - auricle occlusion(98.08%)
      - cranial occlusion(97.76%)
      - frontal occlusion(98.63%)
      - nasal occlusion(N/A)
      - ocular occlusion(98.94%)
- ▸ posed image(98.80%)

Fig. 2. Result-set viewer showing images from the Dlib test that have occular occlusions.

[3] F. Schroff, D. Kalenichenko, and J. Philbin, “Facenet: A unified embedding for face recognition and clustering,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[4] D. E. King, “Dlib-ml: A machine learning toolkit,” *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1755–1758, 2009.

[5] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint arXiv:1708.08296*, 2017.

[6] “Labeled faces in the wild home,” 2019. [Online]. Available: http://vis-www.cs.umass.edu/lfw/index.html

[7] “What if...” 2019. [Online]. Available: https://pair-code.github.io/what-if-tool/index.html

[8] W. Samek, T. Wiegand, and K.-R. Müller, “Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models,” *arXiv preprint arXiv:1708.08296*, 2017.

[9] M. K. Sarker, N. Xie, D. Doran, M. Raymer, and P. Hitzler, “Explaining trained neural networks with semantic web technologies: First steps,” *arXiv preprint arXiv:1710.04324*, 2017.

[10] P. Hayes and M. Warren, “A lightweight ontology for describing images,” in *2010 AAAI Spring Symposium Series*, 2010.

[11] C. J. Mungall, C. Torniai, G. V. Gkoutos, S. E. Lewis, and M. A. Haendel, “Uberon, an integrative multi-species anatomy ontology,” *Genome biology*, vol. 13, no. 1, p. R5, 2012.

[12] M. Bennett, “The financial industry business ontology: Best practice for big data,” *Journal of Banking Regulation*, vol. 14, no. 3-4, pp. 255–268, 2013.

[13] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, “Learning to align from scratch,” in *NIPS*, 2012.