POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Cross-domain Textual Explanation for Explainable AI

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Athira Selvan

Advisor: Prof. Mark James Carman

Academic year: 2021-2022

1. Introduction

From the development of self-driving cars to smart assistants, Artificial Intelligence has become a part of most systems we use in our everyday life.Machine learning algorithms used in these systems are black-box models, whose internal working is unknown. Explaining or interpreting the outputs of these models is not possible.Many experts remain wary of using machine learning due to this concern, especially in the domains where these predictions are crucial for decision-making. This makes explainable AI an important field as it provides tools and methods to explain these models. In this work, we develop a system that produces textual explanations for the classification problem. A grammar that was already developed as part of the initial research on this topic is used to generate new training datasets from new domains. In the previous work, GPT-2 model was fine-tuned on cardiovascular and diabetes datasets to produce the textual explanation. Even though the results were promising for explaining datasets from the said domains, the models failed to generalize for new domains. In this work, we further develop the system to make it more generalized to produce textual explanations for classification models from any domain. We add 6 new datasets from multiple domains for training the models. We introduce a modified encoding for the inputs and modified grammar for developing outputs for training. We experimented with the T5 language model for text generation. The Results of the comparative study done on GPT-2 and T5 models show that the T5 model is best suited for this task. We present a multi-domain textual explanation model fine-tuned on T5 that can produce textual explanations for classification models from any domain. We also explore ways to make the model produce more meaningful and varying natural language outputs different from the grammar.

2. Related work

As the field of artificial intelligence gets more and more popular, we find it being used in almost all technologies and domains these days. From voice assistants to self-driving cars, we see AI algorithms in every technology we use nowadays. Most of these systems are black-box models. The exact justifications for why these models produced a particular output are not known. A lot of work is being done for making these models more justifiable, as that information is very crucial when using these models in important decision-making tasks. But for that to be done, these models need to be more simple, which in turn affects their performance. This is

a reason why machine learning models are not being widely used in the field of medicine, security, etc. When making a diagnosis decision, the doctor needs to know what factors made the model make that decision. This information is very important as a wrong diagnosis can even threaten the patient's life. So explainable AI, a field that deals with methods and tools to explain the black-box models, has become more popular and important for making machine learning useful in real-life. xAI has many existing tools for explaining these models, like SHAP, LIME, counterfactuals, ceteris paribus, etc. Many works have been done in xAI to produce visual explanations. However, for normal users who are not from scientific domains, understanding these visual explanations would be another hurdle. So it is important to have these explanations in a more understandable form of natural language text. We use language models for text generation.
This section introduces 2 novel works done on the textual explanation for explainable AI.

2.1. TextVQA

A First Look: Towards Explainable TextVQA Models via Visual and Textual Explanations [6] the paper we propose MTXNet, an end-to-end trainable multimodal architecture to generate multimodal explanations, which focuses on the text in the image. The work focuses on giving explanations only for images. In our work, we deal with classification datasets that are tabular. Our model gives explanations based on all features and the other data instances present in the dataset. We make use of the xAI techniques of SHAP, counterfactuals, etc for the explanation.

2.2. Explanation algorithms for the lung cancer

The paper "The natural language explanation algorithms for the lung cancer computer-aided diagnosis system" proposes 2 algorithms for natural explanations for decisions of a lung cancer computer-aided diagnosis system. The first part of this algorithm uses LIME for selecting important features and the second part converts the important features to natural explanations. The first algorithm uses a special vocabulary of simple phrases which produce sentences and their embeddings. The second algorithm significantly simplifies some parts of the first algorithm and reduces the explanation problem to a set of simple classifiers. [5]. The work only focuses on lung scan images and the explanation is generated from the algorithm. In our thesis, we focus on classification datasets from many different domains and we make use of a language model in order to generate a textual explanation for making it more general and human-like.

2.3. Textual explanation thesis

This work is a continuation of Vittorio Torri's Master's thesis "Textual explanation for Intuitive Machine Learning" [7]. The work developed a textual explanation model using GPT2. As there was no training set already available, the thesis developed a grammar to create the training dataset. The grammar was made for the cardiovascular dataset.
In his work, The input feature values, along with SHAP values of the features, mean and standard deviation values for numerical features, and one counterfactual feature were the inputs given to the model in order to produce the explanation. The input was encoded in a special format. A grammar was prepared for generating textual explanations for a given input. The encoded input along with the explanation generated by the grammar as the desired output was contexted together to be trained with GPT2. The model was tested on a new dataset of Pima diabetes.
The results from the work showed that the model performed well for the cardiovascular dataset but failed to generalize for the diabetes dataset.
To handle that, the model was fine-tuned on the diabetes dataset. The results showed that after fine-tuning the model performed well for diabetes data.
Our attempt in this thesis was to make the model generalize such that it produces textual explanations for any dataset from any domain. A detailed explanation of our experiments and progress is given in the next sections.

3. Datasets

For training, we use multiple classification datasets. Apart from Cardio and diabetes datasets already used in the previous work, we introduce 6 new datasets.

3.1. Cardiovascular

The first is a cardiovascular diseases dataset, taken from Kaggle, with 11 features and 1 target. It is a huge dataset with data from 70,000 patients collected at the moment of medical examination. Even though the dataset has 70000 values only 7000 balanced instances were used for training the model, to avoid overfitting as all the other datasets had less than 10000 instances.

3.2. Stroke

The second, also taken from Kaggle, is a stroke dataset and is used to predict whether a person is likely to get a stroke. The dataset contains 5110 instances.

3.3. Breast cancer

The third one is a breast cancer dataset. A digitalized image of a Fine Needle Aspirate of a breast mass is used to compute the features of this dataset. This breast cancer dataset was obtained from the University of Wisconsin Hospitals, Madison by Dr. William H. Wolberg. A pre-processed dataset was taken from Kaggle. [4] This dataset contains the following features mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, and diagnosis. Diagnosis is our target variable.

3.4. Mammography

Mammography is a breast cancer screening method. The dataset can be used to predict the severity (benign or malignant) of a mammographic mass. The BI-RADS attributes and the patient's age can be used for this prediction as features. The dataset contains a BI-RADS assessment, the patient's age and three BI-RADS attributes, and the diagnosis. The dataset has 516 benign and 445 malignant masses.

3.5. Stag heart disease

Similar to the cardio dataset, this dataset is used to determine the chances of having a heart disease. But this has different features than the cardio dataset. There are 270 instances with 150 positive and 120 negative predictions.

3.6. Occupancy detection

This dataset is used for predicting room occupancy based on the IoT sensor data. Understanding if the room is occupied or not is very useful for automatic electronic device functions. This Dataset contains 2666 rows and 6 columns with features: Temperature, Humidity,Light, CO2 , Humidity ratio- all 5 numerical features and the target variable Occupancy

3.7. Diabetes

This dataset was used in the previous thesis for checking the generalization capability of the trained model. In this work also keep this dataset aside and use it only for testing. This dataset has around 900 instances used for the classification of diabetes. 268 of 768 instances are positive and the others are negative

3.8. Smoke detection

This dataset is used to predict if a fire alarm will go on or not based on the values of the features. The features are values coming from different IoT sensors measuring different conditions of the room to determine if there is smoke. This dataset has about 60000 rows and 16 columns. We were not able to use this dataset for training due to its long input encoding because of the high number of features. But w used it to test the generalization capability of the model on the new data.

4. Evaluation methods

We use the 3 most used metrics for text generation-BLEU-4, METEOR, and BLEURT. The automatic matrices have their limitations when it comes to measuring the correctness of a text that can be very different from the reference sentences. In our task, as we need the explanation to be as natural and different from our grammar as possible, we cannot rely just on these metrics for evaluation. We want the model to be more generalized but at the same time, we need it to learn to use correct values and feature names given in the input. We are okay with the model modeling the feature name to something with the same meaning, but the model completely inventing new features and values is not acceptable while explaining. There are 2 types of hallucination for our task - Intrinsic and Extrinsic hallucinations. In intrinsic hallucination, the generated text contains information that is contradicted by the input data. In extrinsic hallucinations, the generated text contains extra information irrelevant to the input. [3] Even though

developed to use for in data-to-text generation evaluation, We can use the PARENT [2] evaluation metric to evaluate the hallucination in the explanations.

5. Experiments

5.1. Multi model training

The first set of experiments focused on training the model with more datasets. A modified grammar was written for each new dataset. The aim was to see the effects on the performance of the model when trained on 2 datasets from different domains and tested on a dataset from a new domain. Only a few instances of the cardio dataset were used in order to limit overfitting. The GPT2 was trained and tested with cardio, stroke, and diabetes datasets. 3 models were trained with pair of datasets and were tested on the third dataset.

5.2. New Encoding

The encoding for the input was modified in order to make the input size smaller to be able to train with new datasets with more features. We also included a new percentage value confidence that shows the probability of the model prediction.

input=[name=age(52.271232876712325), shap=0.0, mean=53.3, std=6.8];...[name=BMI(21.92), shap=0.0, mean=27.4, std=5.0];target=[name=cardiovascular disease, prediction=no disease, confidence=98%];cf=[name=age(61.9)] [name=diastolic blood pressure(80.0)] [name=BMI(26.1)]; cf_pred=cardiovascular disease

5.3. T5 model

We replaced the gpt2 model with the T5 model for fine-tuning, to compare the efficiency of these models for our task. The T5 model clearly showed a rise in performance compared to GPT2. This result was particularly evident when tested with the diabetes dataset. The T5 model was able to generalize well for the new diabetes dataset compared to the GPT2 model.

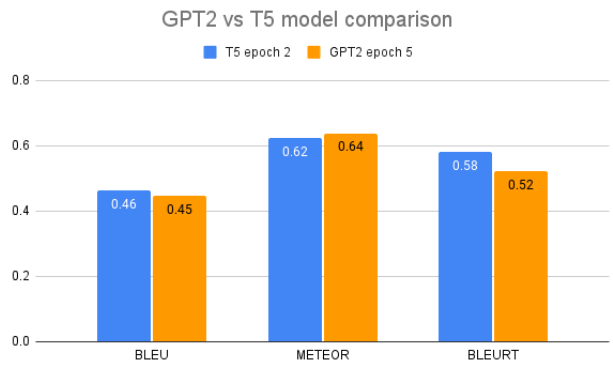

Figure 1: GPT2 vs T5 model comparison

5.4. Handling overfitting

To handle the overfitting of the models on cardio or stroke datasets, we researched by taking a subset of 50 instances from each dataset to see the effect on the model results. The subset of 50 instances was taken from all datasets. The subset performed well on generalizing for new datasets than the model trained with all data instances.

5.4.1 Output from subset(50) model

> The first element which influenced the prediction of diabetes with a confidence of 95% is the fact that the patient is young, where low values of this feature are associated with a high probability of diabetes. Moreover, the fact that the patient is middle-aged plays an important role, where low values of this feature are associated with a high probability of diabetes. If age was 49, BMI was 26, and blood pressure was 68 then the prediction would have been no diabetes.

The above output shows that even though no data on diabetes was given to the model during training, the model was able to generate a good explanation for it.

6. Paraphrasing

In order to have a more varied output and to use the full potential of T5, we experimented with different tasks T5 was pre-trained for. Paraphrasing, summarization, and changing repetition penalty values, are the 3 methods we tried for this. A pre-trained model made by fine-tuning t5 on the TaPaCo dataset is used for

paraphrasing. The model t5-base-Tapco was used for this purpose.
The T5 model is already pre-trained for the summarization tasks. By using the prefix "summarize:" the model generates the summary of the input text. This summary is more condensed and sometimes used different phrases which makes the outputs more varied.
T5 model evaluation parameter repetition value can also make the model generate more natural sentences. Experimenting with different repetition penalty values gave more natural conversation-like explanations.

6.1. Hallucinations in output

An example of intrinsic hallucination in our case is when the model says systolic blood pressure 130, even though it was 90 in the input. An example of extrinsic hallucination is when it uses phrases like "We are all about making sure we get enough information for this prediction", this phrase was invented by the model which was not in our grammar. So we do not have a way to evaluate this kind of phrase using any metrics.
Intrinsic hallucination can be measured using certain evaluation metrics. In our task, we can use the metrics parent cite here.
Given a candidate, a reference, and a table, the parent measures intrinsic hallucination in the candidate.
Many works have been done for reducing hallucination in text generation. The best mitigation method we found in our research that can be used for our task is to introduce more error-free datasets. We can also use a post-generation technique to change the error values of features, before giving the output to the user.

7. Results & Conclusions

We present a textual explanation model that can generate explanations for datasets from different domains. We compare our new models - model trained on all datasets, trained with a subset of 50 instances from all datasets, trained with a subset of 100 instances from all datasets, with models from the previous work, on the diabetes dataset. The result is in the following figure 2. Even though the best model is FT in the graph, the FT model is the GPT2 model trained with cardio and then fine-tuned for the diabetes dataset. This fine-tuning makes the model work better. But the subset models were not trained with any dataset from the diabetes domain. Still, the models were able to get good results on an unseen dataset with scores close to the Fine-tuned one, proving its generalization capabilities.

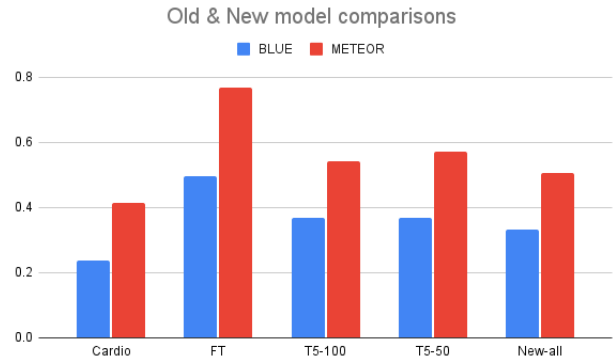

Figure 2: Final model comparison

In conclusion, our experiment shows that, by fine-tuning the T5 model with many datasets from multiple domains, we can get a model that works across all domains that are able to produce textual explanations for a new domain. We also find that more data sets from different domains give better results than more data instances from the same domain. Thus introducing more smaller datasets can clearly improve the capabilities of the model. An efficient way has to be developed to model our outputs to evaluate the hallucination using PARENT metrics. Also, more work needs to be done to understand how to make use of summarization, paraphrasing, and other generalization capabilities of T5, in order to give better outputs to the user without repetition. These can also be used for creating new data instances after some processing to correct errors. Future work has to focus more on reducing hallucinations and making the model outputs more human-like, bringing more varieties in the outputs.

8. Bibliography

The Executive Summary should contain the very essential bibliography of your study. It is suggested to use the BibTeX package [1] and save the bibliographic references in the file `bibliography.bib`.

9. Acknowledgements

Here you might want to acknowledge someone.

References

[1] CTAN. BiBTeX documentation, 2017.

[2] Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William W. Cohen. Handling divergent reference texts when evaluating table-to-text generation, 2019.

[3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, nov 2022.

[4] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming, 1990.

[5] Anna Meldo, Lev Utkin, Maxim Kovalev, and Ernest Kasimov. The natural language explanation algorithms for the lung cancer computer-aided diagnosis system. *Artificial Intelligence in Medicine*, 108:101952, 2020.

[6] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable TextVQA models via visual and textual explanations. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 19–29, Mexico City, Mexico, June 2021. Association for Computational Linguistics.

[7] Vittorio Torri. Textual explanations for intuitive machine learning. Master's thesis, Politecnico di Milano, Milano, 12 2021. http://hdl.handle.net/10589/181513.