Dataset Augmentation with Generated Novelties

Alina Nesen
Department of Computer Science
Purdue University
West Lafayette, IN, USA
anesen@purdue.edu

K M A Solaiman
Department of Computer Science
Purdue University
West Lafayette, IN, USA
ksolaima@purdue.edu

Bharat Bhargava
Department of Computer Science
Purdue University
West Lafayette, IN, USA
bbshail@purdue.edu

2021 Third International Conference on Transdisciplinary AI (TransAI) | 978-1-6654-3412-6/21/$31.00 ©2021 IEEE | DOI: 10.1109/TransAI51903.2021.00015

***Abstract*— As machine learning models take over an increasingly larger number of domains in our lives, their accuracy, fairness, transparency and adaptability become of greater importance. In the everchanging environments, the resulting ability of the models to perform accurately depends on whether they are able to handle novel, unpredicted and unforeseen instances, examples and classes or any other novel changes in the world of model operation, such as environmental, contextual, distributional changes. The proper handling of novelties sustains the model's usefulness and adeptness in the long run. The efficiency of response to the encounter of novelties depends on the efforts that were invested at the model training, design and data collection stages. In this work, we propose a variety of approaches and methods which can be incorporated into the novelty generation techniques at the earliest stages of creating the machine learning dataset and the model to assure its robustness and reduce the bias. We revisit distinctions between novelties and anomalies to define a formal novelty generation framework that is domain-agnostic and budget efficient. Then we propose a video-specific use case and evaluate the result of the chosen methods on the video dataset. Our methods aim at making the machine learning solutions adaptable, responsible and show improvement in the accuracy and ability of the models to detect novelties.**

Keywords-novelties, novelty generation, bias, knowledge graphs

I. INTRODUCTION

The machine learning models are being deployed across many application areas: from image recognition and object detection to recommendation systems in the commerce and business personalization applications; from financial trading to healthcare sectors to Internet of Things; from natural language understanding to machine translation. The models often involve deep learning networks trained in supervised or semi-supervised methods on the previously collected and labeled datasets. Thus, prediction abilities of the model are defined by the examples it has seen in the training dataset. The novel and fluid situations that come from the open-world challenges and out of distribution examples may lead the trained models to produce erroneous outcomes with high confidence. The change of the testing instances or their context induces changes in target concepts which may leave the trained models lacking capabilities of accuracy because they are unaware how to react to changes. The framework for the theory of novelties can help to unite the possible types of novelties under common umbrella and provide necessary definitions of what it means to be a novelty [1]. Following the proposed definitions, one can apply them to detect novelties within the specific domain. Given the suite of features for modeling and measuring the novelties, the rules for generating novel examples for the specific domain can be derived, providing a method for producing novel data points on all the steps in the hierarchy.

In the fast-changing environments the quality of the model degrades over time due to the concept of data drift, when the statistical properties of the target variable change in unforeseen ways and cause the prediction to become inaccurate. Thus, the concept of the data drift represents one of many the possible scenarios when novelties cause the trained model to underperform. The menaces of failing to detect novelties are hard to overestimate. Recent research shows that the models trained on incomplete, inadequate datasets may lead to dangerous consequences in technical, technological, environmental and social aspects [2]. The datasets that encode bias have been found harmful and decrease the trust in AI systems and algorithms [3]. The importance of novelties detection and generation is amplified by the economic and societal importance of creating new objects, entities, techniques and practices.

The supervised learning models aim at identification of the well-established entities that were seen rather than the discovery and identification of new entities. Intuitively, if we provide the model with as many novel examples at the time of training as possible, it should be able to pick up the pattern and distinguish it at the time of testing with higher accuracy. In order to build the systems that can routinely acquire and extract knowledge, the datasets that are used for training must contain much greater variety of the instances. Thus, the goal of the current paper is to investigate, summarize, apply and evaluate the methods to achieve this.

The contributions of this research are two-fold:

- A roadmap for novelty generation approaches to be incorporated into machine learning solutions for various domains.
- Empirical experiments with the proposed approaches for measuring the improved result of the trained machine learning model and provide comparative analysis with other methods.

II. RELATED WORK

The novelty generation problem and expanding the dataset with novel examples covers three different topics within machine learning, namely data augmentation, novelty and anomaly detection, and novelty generation.

The success of deep neural networks is often attributed to their resemblance to human brain. While the extent of this similarity is subject to discussion, application of cognitive procedures observed in humans, animals and biological systems has gained certain popularity in deep learning experiment design. For example, approaches that combine knowledge graphs with deep neural networks are based on the intuition of using external knowledge by humans when they classify objects and extract meaning [4]. Thus, when defining a framework for novelty detection, it is important to analyze how the human brain signals and reacts to novelties.

978-1-6654-3412-6/21/$31.00 ©2021 IEEE
DOI 10.1109/TransAI51903.2021.00015

Inferotemporal cortex involved in scene and object analysis changes the firing rate in response to familiar as compared to novel stimuli, which is called repetition suppression. The human brain does not parcellate the signals of visual identity and visual novelty into distinct parts. On the contrary, the brain operates in a holistic manner and measures the novelty in a continuous rather than a binary format.

Novelty generation, in turn, is related but not entirely similar to the novelty creation, AI creativity and knowledge generation problem. Knowledge-driven creativity has been a topic of extensive research lately, and knowledge generation concept intersects with creativity. The novelty generation for enhancing the training and testing datasets is particularly related to machine learning creativity, which can be defined as generating previously unknown but meaningful and valuable new types of objects, new classes, new instances belonging to those classes or new states. This is not a mere synthesizing but doing that with a purpose of optimizing a value (reward) function. This task can be reduced to out-of-distribution generation and evaluation of the quality of the generated samples. Evaluation of the generated knowledge to measure its novelty degree is ultimately a domain-specific task since each discipline, domain, natural science has its own theory. The question of novelty can only be properly assessed for each generated case separately.

The dataset augmentation with synthetic data has been a popular routine among machine learning practitioners since it allows to create models that generalize better and offer the benefit of protecting the privacy of the data owners. However, the published works concentrate on augmenting the datasets with the examples that preserve the same properties as the initial distribution. The techniques to augment the data are mostly domain-dependent: for natural language processing one may benefit from the thesaurus and generated text corpus, for image and video apply affine, elastic, neural-based transformations, such as rotating, scaling, shift, image blending. Random erasing and inserting adversarial noise have been used for synthetic data augmentation. In [5] the speech emotion recognition dataset is augmented, in [6] information dropping for computer vision datasets is studied, in [7] the data is augmented in a way to eliminate the misrepresentation for a particular group in the dataset. Time-series data has been shown to be efficiently augmented with the ordinary differential equations models in [8]. GAN-based data augmentation techniques, on the contrary, may introduce additional bias to the dataset. The authors in [9] provide bias-mitigation techniques for such scenarios.

The novelty and anomaly detection terminology is used interchangeably in literature, similar methods can be applied for both tasks. Anomaly detection with augmented dataset that does not need labeling is discussed in [10]. Deployment and engineering details of the system that is used for real-life use cases to detect novelties, anomalies and mission-relevant information are proposed in [11, 12, 13].

III. Proposed Approach

We begin by differentiating the notions of novelties and anomalies. They are easy to confuse since they are both outliers that a machine learning model is not aware of during training. While some works do not make distinction between the two, in this work we concentrate on novelties as opposed to anomalies. We define a novel instance as the one that comes from a class previously unseen but related to the known classes. Anomaly is the outlier previously unseen and not necessarily related to the classes in the dataset. The relevance threshold may vary and can be established theoretically or empirically to classify new instances as novel or not. In broad sense this will depend on the domain and the specific classification task within it. Thus, the new type of a sportscar parked in the neighborhood may represent a novelty while a horse carriage would be an anomaly according to our definition. Note that the threshold can be adjusted in such a way that a horse carriage is considered to be a class with close proximity (as belonging to a superset of 'Transport' objects).

Another aspect that makes novelties and anomalies fundamentally different is the manner of their development over time. If outliers continue to arise at the same rate, after certain time they are not considered novel anymore, the system will have adapted and with proper tuning will declassify them into a known class (perhaps by adding a novel class to the set of the common classes). Thus, an entity with a strong degree of oddity but which can appear with some regularity on the temporal axis is defined as an anomaly. An entity which is below a certain specified threshold, but which does not belong to a known class and has not started its regular appearance in the world which we are examining is considered a novelty. Hence, we establish a 2-fold criterion for distinguishing a novelty from an anomaly:

- Degree of oddity of the particular event, instance or a state quantified with one of the distance-based methods to measure the dissimilarity between the incoming and existing examples.
- Temporal frequency of the event and its underlying distribution parameters.

In order to cover as many novel scenarios as be possible, we structure the generating novelties on the world-level or feature levels. The model of the world includes a set of possible states and the reward function that is associated with these states. The new states or new unexpected rewards can be generated. The new state can come from a new observation, i.e. collected from a sensor.

To explain novelties in terms of the hierarchy of novelties, we present two examples. Our first scenario includes a use case from the West Lafayette police department. They use surveillance camera video feeds to identify a person of interest (e.g. when looking for a missing person), while searching for distinct attributes of the person – his/her gender, race, clothes colors. Different feature extraction models [15] have been trained to achieve this goal. While considering the world state as objects, an easy novelty would be obtained by training these models in one area but getting the test data from a different distribution. A novelty which is more difficult to detect would be generated by introduction of a completely new feature for the person (such as height or build) and expecting the machine learning algorithms to identify the closest label for it or, better, assign a 'novel' label to it instead of misclassification. This second approach is only possible if the design of the neural network allows it, in other words, if the 'novel' class was introduced at the stage of training. While explaining the world states as events, an easy example of a novelty would be a football

game which is happening in town and the police detective from the example above is looking for the same types of clothes as the team jersey. Apart from exploiting the hierarchy of novelties, the variety of approaches that introduce a random perturbation or modification into the particular data point can be used as a set or individually. We list the proposed methods below.

Distance based methods. These methods rely on pairwise distance matrices, where a distance function is applied to each pair of inputs. These methods vary depending on the distance function chosen to calculate similarity. The Euclidian and Manhattan distances are the straightforward but still efficient ways for numerical inputs. For distance between two sequences, the longest common subsequence can be used as a metric measurement. The edit distance (or Levenshtein distance) represents the number of edit operations (insertions, deletions, substitutions) that will transform one sequence into another.

Semantic distance methods. In the presence of a semantic network, the path length between the two entities can represent their semantical similarity. Knowledge base, thesaurus or ontological graph can serve as a semantic network.

Distribution-based methods. Change in the distribution of incoming data samples can also signify presence of a novelty in the given world. If the underlying probability distribution is known, introducing changes into its parameters will influence the meta-features of the novel examples, such as their frequency and locality.

GAN-based methods. The unique characteristic of the generative artificial network based methods is that, unlike the previous approaches which reproduce the known objects, the GAN-based ones do not follow the road of adding noise to the known classes but rather capture the distinctive properties of the objects or entities and introduce practical changes into them. To ensure that GAN-based data augmentation techniques do not introduce additional bias, it is recommended not to limit with single generative network but train several GAN variants at once. The Adversarial Autoencoder GAN variant can transform high-dimensional multimodal data distribution into low-dimensional unimodal latent distributions. The samples that are generated with smaller probabilities are rare events and can be used for the training dataset augmentation.

Information theory. The information content of the dataset can be computed with measures such as entropy and relative entropy. Since the novelties will significantly alter the information content of the dataset, the novelty can be generated and later evaluated using the degree of entropy that it introduces into a dataset.

The benefits of introducing synthetic novelties into the dataset include the fact that generated novelties are interpretable and interdependent. Algorithms for generating novelties must still adhere to the complexity guidelines and be feasible and reproducible. Another benefit is the fact that synthetic examples do not intrude into privacy and do not require anonymization.

Once the novel examples have been generated, we proceed to the next step in the pipeline which can take one of the two viable routes:

Route1: Compose a novel class NC out of the generated examples and assign them the same 'novel' label. When training a deep learning model, augment the dataset with one novel class while other classes still remain the same.

Route2: Train a binary classification model that will distinguish between novel and non-novel classes only. This is essentially a simplification of the previous method and can be more appropriate when the cost of building a more complex model is quite high and the dataset is imbalanced. The choice of the method depends on a particular task and on the domain. The important criterion for selecting the generated samples is how meaningful and valuable they are with regards to the rest of the data points. Hence the mechanism of evaluation of the generated sample is necessary in the extended novelty generation framework.

IV. Experiments and Results

We have used the scenario with detection novelties in the real-life surveillance video from the Tbilisi 2015 dataset [2]. The video data contains recordings from the surveillance cameras of the city streets affected by a large flood along with the wild animals that escaped from the zoo during the flood (Figure 1). We have used the semantic distance methods for dataset augmentation and Route1 for training a deep neural network with an additional class. The procedure for data augmentation with semantic distance based method is as follows:

1. For each class in the dataset we introduce an additional vector of concepts v that are most semantically similar to the given class. The length of the vector v depends on the need to include more examples and the semantic network quality. In general, it is better to include the concepts that are within 2 or 3 steps from the given node as farther instances tend to fall too far from the initial label. We use off-the-shelf semantic network, ConceptNet [15] for identification of the most semantically similar concepts to each detected object. ConceptNet provides the word embeddings in the similar fashion to word2vec and GloVe.
2. For each added class, the dataset is augmented with the images that correspond to the newly generated labels (potential novelties). The process of augmentation can be automated: for the case of image/video dataset augmentation, we use the web crawler to extract the images of interest, i.e. for every generated class the search result of the images for that class is added. However, the classes that were added in step 1 are not added to the set for the multiclass classification as separate classes. Instead, they will be united into one class 'NC' or Novel Class, which will be used to train a Pairwise Matching Network to evaluate a probability of the example being a novel instance related to the existing instances.
3. The newly added images are assigned the *Novel class* label (NC).
4. Optionally, the initial model is re-trained with that additional class.

Figure 1. Screenshot from the video dataset of the Tbilisi city streets during the flood of 2015.

The augmented dataset contains an additional novel class "NC" with all the generated labels and corresponding images labeled as 'NC.

The next step involves training Pairwise Matching Network (PM-NET) as described in [14]: the network is used to identify the probability that the detected object really belongs to the certain class. The Pairwise matching network is a convolutional neural network which is trained to detect the probabilities of the instances belonging to the same class. every object is first projected to a higher dimensional embedding space and then after concatenation the two objects will be treated as one feature vector:

$$x_{1k} \otimes x_{2k}$$

The CNN is used to predict the probability of the two instances belonging to the same class.

The confidence score is assigned to each detected object before the final decision is made regarding its belonging to a certain class from the training dataset or to a novel class 'NC' which is not in the training dataset. In the process of evaluating the probability, all the calculated probabilities that the object belongs to the class from the list of the training classes are compared with a threshold, and if it is greater then the class assignment is performed, otherwise, the object is marked as a potential novelty.

$$\hat{y} = \begin{cases} novel\ class, & if\ maxP < threshold \\ argmax\ P, otherwise \end{cases}$$

The two benchmark training datasets for the Pairwise Matching Network are created with the usage of the original dataset *D1* and augmented datasets *D2*. The two new pair-wise datasets *D3* and *D4* are constructed with randomly selecting pairs from *D1* and *D2* respectively and assigning them label 1 if they belong to the same class and label 0 otherwise. The reported accuracy of the PM-NET with original data is 52.5% and with the augmented data 54.5%.

The results of the improved detection of the novel objects in the augmented dataset are reported in Table 1. The four models trained as described in [4] are evaluated for their ability to predict the probability of an instance being a novelty, according to the Step 4 of the method described above.

Table 1. Accuracy in the detection of novelties with the augmented dataset vs. unchanged dataset with retraining a model.

	Original dataset	Augmented dataset
NN1	85.6%	92.7%
NN2	88.4%	93.3%
NN3	88.9%	95.5%
NN4	88.9%	95.5%

V. Conclusions

We have described a domain-agnostic approach for dataset augmentation with novel examples and subsequent model training to improve the ability to recognize novelties and out-of-distribution examples. The augmentation of the dataset using a line of semantically related examples was used for empirical investigation. As a future research direction, this approach can be benchmarked with other methods to determine the optimal ways of dataset augmentation with novel examples.

Acknowledgements

This work was partially funded by the Defense Advanced Research Projects Agency (DARPA) with award W911NF2020003 under the SAIL-ON program.

References

[1] Boult, T. E., et al. "Towards a Unifying Framework for Formal Theories of Novelty." *Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 35. No. 17.* 2021.

[2] Sabokrou, Mohammad, et al. "Adversarially learned one-class classifier for novelty detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 2018.

[3] Wiegand, Michael, Josef Ruppenhofer, and Thomas Kleinbauer. "Detection of abusive language: the problem of biased datasets." *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies, volume* 1. 2019.

[4] Nesen, A., Bhargava, B. "Semantic-Aware Anomaly Detection in Video with Knowledge Graphs" *IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE) (*2020*).*

[5] Pappagari, Raghavendra, et al. "CopyPaste: An Augmentation Method for Speech Emotion Recognition." *arXiv preprint arXiv:2010.14602* (2020).

[6] Chen, Pengguang, et al. "Gridmask data augmentation." *arXiv preprint arXiv:2001.04086* (2020).

[7] Sharma, Shubham, et al. "Data Augmentation for Discrimination Prevention and Bias Disambiguation." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020.

[8] Yadav, Mohit, et al. "Ode-augmented training improves anomaly detection in sensor data from machines." *arXiv preprint arXiv:1605.01534* (2016).

[9] Hu, Mengxiao, and Jinlong Li. "Exploring Bias in GAN-based Data Augmentation for Small Samples." *arXiv preprint arXiv:1905.08495* (2019).

[10] Lim, Swee Kiat, et al. "Doping: Generative data augmentation for unsupervised anomaly detection with GAN." *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018.

[11] Nesen, Alina, et al. "Towards situational awareness with multimodal streaming data fusion: serverless computing approach." *Proceedings of the International Workshop on Big Data in Emergent Distributed Environments.* 2021.

[12] Palacios, Servio, et al. "WIP-SKOD: A Framework for Situational Knowledge on Demand." *Heterogeneous Data Management, Polystores, and Analytics for Healthcare. Springer, Cham*, 2019. 154-166.

[13] Stonebraker, Michael, et al. "Surveillance Video Querying With A Human-in-the-Loop."(2020).

[14] Qin, Qi, Wenpeng Hu, and Bing Liu. "Text Classification with Novelty Detection." arXiv preprint arXiv:2009.11119

[15] Speer, Robyn, and Joanna Lowry-Duda. "ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge." arXiv preprint arXiv:1704.03560 (2017).