NOVA - a tool for eXplainable Cooperative Machine Learning

Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, Elisabeth André

Angaben zur Veröffentlichung / Publication details:

Heimerl, Alexander, Tobias Baur, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2019. "NOVA - a tool for eXplainable Cooperative Machine Learning." In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), 3-6 September 2019, Cambridge, UK*, 109–15. Piscataway, NJ: IEEE. https://doi.org/10.1109/ACII.2019.8925519.

NOVA - A tool for eXplainable Cooperative Machine Learning

Alexander Heimerl, Tobias Baur, Florian Lingenfelser, Johannes Wagner, Elisabeth André
Human Centered Multimedia Lab
Augsburg University Germany
Augsburg, Germany
lastname@hcm-lab.de

***Abstract*—In this paper, we introduce a next-generation annotation tool called NOVA, which implements a workflow that interactively incorporates the 'human in the loop'. In particular, NOVA offers a collaborative annotation backend where multiple annotators join their workforce. A main aspect of NOVA is the possibility of applying semi-supervised active learning where Machine Learning techniques are used already during the annotation process by giving the possibility to pre-label data automatically. Furthermore, NOVA implements recent eXplainable AI (XAI) techniques to provide users with both, a confidence value of the automatically predicted annotations, as well as visual explanation. This way, annotators get to understand whether they can trust their ML models, or more annotated data is necessary.**

***Index Terms*—annotation tools, cooperative machine learning, explainable AI**

I. Introduction

In various research disciplines the annotation of social behaviours is a common task. This process includes manually identifying relevant behaviour patterns in audio-visual material and assigning descriptive labels. Generally speaking, segments in the signals are labelled using sets of discrete classes or continuous scores, e.g., a certain type of gesture, a social situation, or the emotional state of a person. In Affective Computing, a subset of these events – the so called *social signals* – are used to augment the spoken part of a message with non-verbal information to enable a more natural human-computer interaction. To automatically detect social signals from raw sensory input it is common practice to apply machine learning (ML) techniques. However, the performance of a ML-System is largely dependent on the amount and quality of the annotated training data [1], [2]. Especially in the field of affective computing, annotating enough data can be a lengthy and cumbersome task. That is, a classifier is trained on manually labelled examples to optimise a learning function. Once trained, the classifier is used to automatically predict labels on unseen data.

A solution to this problem is exploitation of computational power to accomplish some of the annotation work automatically. However, to ensure the quality of the predicted annotations this still requires human supervision to identify and correct errors. To keep the human effort as low as possible, it is useful to understand why a model makes wrong assumptions. Therefore, it is not only important to provide tools that ease the use of semi-automated labelling, but also to increase the transparency of the decision process. By visualising the predictions, for instance, even non-ML experts get an idea about the strengths and weaknesses of the underlying classification model and can immediately decide which parts of a prediction are worth keeping. If a particular label is regularly missed, a user could actively provide more training examples for this phenomenon, or redesign the ML system to capture its relevant characteristics better. Ideally, the system even guides the users' attention towards parts where manual revision is necessary. Once an annotation has been revised, the model can be retrained to improve its performance for the next cycle. This procedure can be repeated until a desired performance is reached.

In this paper, we introduce a next-generation annotation tool called NOVA, which implements the described workflow that interactively incorporates the 'human in the loop'. In particular, NOVA offers semi-automated annotations and provides visual feedback to inspect and correct machine-generated labels by incorporating eXplainable AI (XAI) techniques. In that sense, our work combines three recent topics of ML: *Explainable Artificial Intelligence*, as the transparency of the decision process is increased via visualisation of the predictions; *Semi-Supervised Active Learning*, since labels with low confidence are highlighted to guide the user towards relevant parts; and finally, *Interactive Machine Learning*, because human intelligence and machine power can cooperate and improve each other. We subsume our approach under the term *eXplainable Cooperative Machine Learning* (XCML)

II. Related Work

A. Annotation Tools

In the past, a couple of annotation tools with focus on affective computing and social signals have been developed.

This work has received funding from the BMBF under FKZ 01IS17074, FMLA, and from the DFG under project number 392401413, DEEP.

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

The general user interface of NOVA has been inspired by existing annotation tools. Prominent examples include ELAN [3], ANVIL [4], and EXMARALDA [5]. These tools offer layer-based tiers to insert time-anchored labelled segments, that we call *discrete* annotations. *Continuous* annotations, on the other hand, allow an observer to track the content of an audiovisual stimulus over time based on a continuous scale. One of the first tools that allow labellers to trace emotional content in real-time on two dimensions (activation and evaluation) was FEELTRACE [6]. Its descendant GTRACE (general trace) [7] allows the user to define their own dimensions and scales. More recent tools to accomplish continuous descriptions are CARMA (continuous affect rating and media annotation) [8] and DARMA (dual axis rating and media annotation) [9].
Though the mentioned tools are of great help to create annotations at a high level of detail, they suffer from several drawbacks. Firstly, they have been developed with a strong focus on audiovisual material, other signals like depth information, e.g. skeleton and face tracking, or physiological data streams are supported sparsely or not at all. Secondly, almost none of the tools allows different types of annotations. Since different coding types have certain pros and cons the choice depends on the observed phenomenon and should be selectable on demand. Finally, almost all of the tools offer none or only little automation. However, since labelling of several hours of interaction is an extremely time consuming task, methods to automate the coding process are highly desirable. NOVA overcomes the limitation of other tools to only playback audio and video streams, and supports the display of an arbitrary number of video and time-series tracks. Additionally, it has been advanced with features to create collaborative annotations and to apply cooperative machine learning strategies out of the box for multiple recognition problems (see section IV). To support a truly collaborative work-flow between several annotators and the machine, NOVA provides a database back-end to store, exchange, and combine annotation work. It further offers features to visualise parts of the data where a classifier is uncertain, as well as explanations of a model's decisions by incorporating explainable AI frameworks in the annotation workflow.

B. Active and Cooperative Machine Learning

A common approach to reduce human labelling effort is the selection of instances for manual annotation based on active learning techniques. The basic idea is to forward only instances with low prediction certainty or high expected error reduction to human annotators [10]. Estimation of most informative instances is an art of its own right. A whole range of options to choose from exist, such as calculation of 'meaningful' confidence measures, detecting novelty (e. g., by training auto-encoders and seeing for the deviation of input and output when new data runs through the auto-encoder), estimating the degree of model change the data instance would cause (e.g. seeing whether knowing the label of a data point would make a change to the model at all), or trying to track 'scarce' instances, e.g. trying to find those data instances that are rare in terms of the expected label.

Further, more sophisticated approaches aggregate the results of machine learning and crowd-sourcing processes to increase the efficiency of the labelling process. Kamar et al. [11]. make use of learned probabilistic models to fuse results from computational agents and human labellers. They show how to allocate tasks to coders in order to optimise crowdsourcing processes based on expected utility. Active learning has shown great potential in a large variety of areas including document mining [12], multimedia retrieval [13], activity recognition [14] and emotion recognition [15].

Most studies in this area focus on the gain obtained by the application of specific active learning techniques. However, little emphasis is given to the question of how to assist users in the application of these techniques for the creation of their own corpora. While the benefits of integrating active learning with annotation tasks has been demonstrated in a variety of experiments, annotation tools that provide users with access to active learning techniques are rare. Recent developments for audio, image and video annotation that make use of active learning include CAMOMILE [16] and iHEARu-PLAY [17]. However, systematic studies focusing on the potential benefits of the active learning approach within the annotation environment from a user's point of view have been performed only rarely [18], [19].

While techniques that enable systems to learn from human raters have become widespread, little attention has been paid to usability challenges of the remaining tasks left to end-users [20]. Rosenthal et al. [21] investigated which kind of information should be provided to users in order to reduce annotation errors in a setting for active learning. They found that contextual information and predictions of the learning algorithms were in particular useful for the annotation of activity data. In contrast, uncertainty information had no effect on the accuracy of the labels, but just indicated to the labellers that classification was hard. Amershi [22] investigated how to empower users to select samples for training by appropriate visualisation techniques. They found that a representative overview of best and worst matching examples is of higher value than a set of high-certainty images and conjecture that high-certainty images do not provide much information to the learning processing due to their similarity to already labelled images. In another paper by Amershi [23] the authors suggest an interactive visualisation technique to assess model performance by sorting samples according to their prediction scores. In their tool the user can directly inspect samples to retrieve additional information and annotate them for better performance tracking. This way, the tool allows users to monitor the performance of individual samples while the model is iteratively retrained.

Summing up, it may be said that many studies experimentally investigate the potential of novel techniques to minimise human labour. In addition, few studies were run to actually label novel data, rather than test whether such method could save effort. Also note that the prevailing choice is merely ac-

tive learning rather than the combination with semi-supervised learning, e.g. cooperative machine learning.

Relatively little attention has been paid, however, to the question of how to make these techniques available to human labellers. There is a high demand for annotation tools that integrate cooperative machine learning in order to reduce human effort - in particular in the area of social signal processing where human raters typically disagree on the labels [24]. In such a setting, dynamic cooperative strategies appear particularly promising, e.g. not only learning the target task, but also as much as possible about the raters and their reliability depending on the labels and the content being labelled.

C. Explainable AI Approaches

The strive for acquiring explanations for computed decisions goes way back to the 70's, with Shortliffe and Buchanan stressing the need for explanations in rule-based expert systems [25]. The current trend in machine learning tends towards deep learning and neural network architectures that go beyond human interpretability. Therefore the strive of explanation methods is experiencing a renaissance. In general, systems providing explanations can be distinguished between model-specific or model-agnostic approaches. The later are capable of providing explanations independent of the underlying model. Ribeiro et al. introduce in [26] LIME, a model-agnostic approach. Their basic idea is to approximate an interpretable model around the original model. This way they are capable of providing explanations for various problem domains like text and image classification. Their explanations come in the form of visual feedback, highlighting the sections that have been crucial for the prediction of a specific class. They showed that with the help of LIME it is easier for users to determine from a set of classifiers which one performs best for a given problem domain. This is especially useful when test-accuracy scores themselves are misleading. Moreover, they argue that LIME not only is useful for gaining additional insight about a model, but also users have been able to improve performance of classifiers by identifying unnecessary features and removing them based on the explanations generated by LIME.

Alber et al. [27] introduced iNNvestigate a library that provides implementations of common analysis methods for neural networks, e.g. PatternNet and LRP. The supported approaches have in common that they, similar to LIME, highlight regions in the image, that have been important for the classification. A broad variety of methods generating explanations is available and most of the times it is not easily comprehensible which approach suits a given problem domain the best. To address this issue Lundberg et al. [**?**] introduced SHAP. Their framework generates explanations by assigning each feature a value, that describes its importance in regard to the prediction.

Lapuschkin [28] et al. introduced a semi-automated Spectral Relevance Analysis (SpRAy). The approach is based on heatmaps and enables to detect different prediction strategies. Further it can be applied to large-scale datasets, which helps to gain globally insight on the classifiers reasoning process.

While such visual explanation systems are of great value in helping to better understand which part of the input data was relevant for a decision, they still require expert knowledge about how to setup the systems and how to incorporate them with one's own model and data. The NOVA tool implements several of these frameworks in a user interface that allows users to gather explanations about a given frame in a video and a prediction of a label or score with either a pre-trained model or a model that the user trained from the NOVA interface itself.

III. NOVA Tool

The NOVA tool aims to enhance the standard annotation process with the latest developments from contemporary research fields such as Cooperative Machine Learning and eXplainable Artificial Intelligence by giving annotators easy access to automated model training and prediction functionalities, as well as sophisticated explanation algorithms via its user interface.

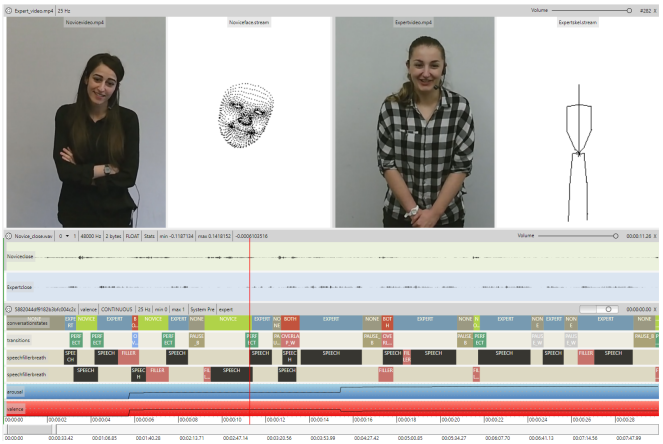

Fig. 1. NOVA allows to visualise various media and signal types and supports different annotation schemes. From top downwards: full-body videos along with skeleton and face tracking, and audio streams of two persons during an interaction. In the lower part, several discrete and continuous annotation tiers are displayed.

The NOVA user interface has been designed with a special focus on the annotation of long and continuous recordings involving multiple modalities and subjects. A screenshot of a loaded recording session is shown in Figure 1. On the top, several media tracks are visualised and ready for playback. Note that the number of tracks that can be displayed at the same time is not limited and various types of signals (video, audio, facial features, skeleton, depth images, etc.) are supported. In the lower part, we see multiple annotation tracks of different types (discrete, continuous and transcriptions) describing the visualised content. Continuous annotations have a variable sample rate and can be performed with either keyboard, mouse or gamepad joysticks.

To support a collaborative annotation process, NOVA maintains a database back-end, which allows users to load and save annotations from and to a MongoDB[1] database running on a

[1]https://www.mongodb.com/

central server. This gives annotators the possibility to immediately commit changes and follow the annotation progress of others. Beside human annotators, a database may also be visited by one or more "machine users". Just like a human operator, they can create and access annotations. Hence, the database also functions as a mediator between human and machine. NOVA provides instruments to create and populate a database from scratch. At any time new annotators, schemes and additional sessions can be added. NOVA provides several functions to process the annotations created by multiple human or machine annotators. For instance, statistical measures such as Cronbach's α, Pearson's correlation coefficient, Spearman's correlation coefficient or Cohen's κ can be applied to identify inter-rater agreement. In the future we plan to add a python plugin interface that allows to add additional measures. Thus the foundations have been laid to fine-tune the number of labelers based on inter-rater agreement in order to further reduce work load by allocating human resources to instances that are difficult to label (see [29]).

Tasks related to machine learning (ML) are handed over and executed by our open-source Social Signal Interpretation (SSI) framework [30]. Since SSI is primarily designed to build online recognition systems, a trained model can be directly used to detect social cues in real-time. [31]. A typical ML pipeline starts by prepossessing data to input data for the learning algorithm, a step known as *feature extraction*. An XML template structure is used to define extraction chains from individual SSI components. A dialogue helps users to extract features by selecting an input stream and a number of sessions. The result of the operation is stored as a new signal in the database. This way, feature streams can be reviewed in NOVA and accessed by all users. Based on the extracted features, a classifier, which may also be added using XML templates, can be trained. Alternatively, NOVA supports Deep and Transfer Learning by providing Python interfaces to Tensorflow and Keras. This way convolutional networks may be trained, respectively retrained, based on annotations saved in NOVA's annotation database on raw video data. Such models may then be used to generate explanations as described in more detail in Section V.

IV. Cooperative Machine Learning

In this paper, we subsume learning approaches that efficiently combine human intelligence with the machine's ability of rapid computation under the term *Cooperative Machine Learning* (CML). In Figure 2, we illustrate our approach to CML, which creates a loop between a machine learned model and human annotators: an initial model is trained (1) and used to predict unseen data (2). An active learning module then decides which parts of the prediction are subject to manual revision by human annotators (3+4). Afterwards, the initial model is retrained using the revised data (5). Now the procedure is repeated until all data is annotated. By actively incorporating the user into the loop it becomes possible to interactively guide and improve the automatic predictions while simultaneously obtaining an intuition for the functionality of the classifier.

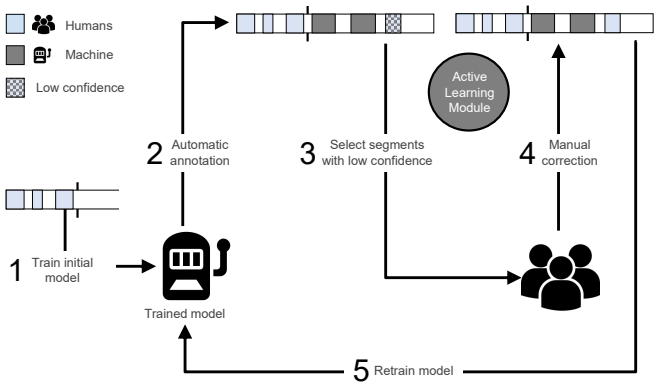

Fig. 2. The scheme depicts the general idea behind Cooperative Machine Learning (CML): (1) An initial model is trained on partially labelled data. (2) The initial model is used to automatically predict unseen data. (3) Labels with a low confidence are selected and (4) manually revised. (5) The initial model is retrained with the revised data.

However, the approach not only bears the potential to considerably cut down manual efforts, but also to come up with a better understanding of the capabilities of the classification system. For instance, the system may quickly learn to label some simple behaviours, which already facilitates the work load for human annotators at an early stage. Then, over time, it could learn to cope with more complex social signals as well, until at some point it is able to finish the task in a completely automatic manner.

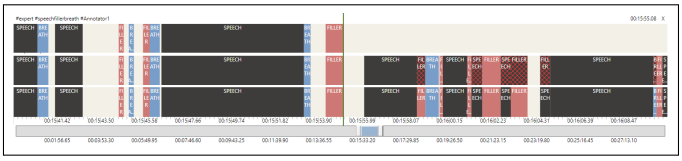

Fig. 3. The upper tier shows a partly finished annotation. ML is now used to predict the remaining part of the tier (middle), where segments with a low confidence are highlighted with a red pattern. The lower tier shows the final annotation after manual revision.

To automatically finish an annotation, the user either selects a previously trained model or temporarily builds one using the labels on the current tier. An example before and after the completion is shown in Figure 3. Note that labels with a low confidence are highlighted with a pattern. This way, the annotator can immediately see how well the prediction worked. To evaluate the efficiency of the integrated CML strategy, in our earlier work [32] we performed a simulation study on an audio-related labeling task. Following this approach we were able to reduce the initial annotation labour of 9.4h to 5.9h, which is a reduction of 37.23%.

V. eXplainable AI (XAI)

Nowadays, modern classification algorithms, such as artificial neural networks are in general able to handle very large input vectors - up to the point where raw data streams instead of descriptive features are fed into the classification system. Irrelevant features tend to be ignored and non-linear decision

boundaries are modelled by hidden layers. These advantages make neural networks a very popular classification scheme in today's environment of growing data availability. Correct network architecture is however a challenging task, as there exist mainly rules of thumb on how to determine parameters such as the number of hidden layers or layer types. Calculation load and complexity within a deep neural network is heavily increasing with the depth of the architecture and an evident relation between input data and resulting decisions becomes less comprehensible and relatable - the term *black box* is often used in this context. Given our goal to provide a cooperative learning environment between user and machine, it is nonetheless most important to provide the best possible explanations of made decisions to the user. To meet this demand we extended NOVA with the two explanation frameworks LIME [26] and iNNvestigate [27]. A brief overview for both has been given in subsection II-C. This extension allows an in-depth analysis of predictions with the help of visual explanations. The possibility to generate explanations can be beneficial for several use cases. In general whenever a model's prediction is wrong you can not only examine the prediction scores, but also take a visual explanation into account that has been generated by exploring the features most important for the classification. Moreover, this is not only the case for misclassifications. Explanations can also help to gain additional information when there are serious doubts on what the model really has learned. With the help of their explanation framework, Ribeiro et al. revealed in [26] that correct predictions are not necessarily based on semantic correct correlations.

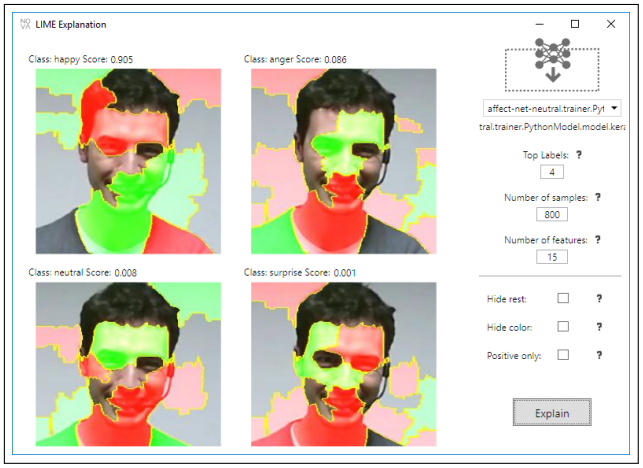

Fig. 4. Explanations for the top four classes generated in NOVA with the usage of LIME.

In section III we mentioned that NOVA provides the possibility to complete unfinished annotations automatically and highlight uncertain predictions with a confidence score. The explainable extension allows now to further investigate those particular spots and gain additional insight on the classifier's decision making. The following paragraph will outline an example for a potential work flow with explanations inside of NOVA. Figure 4 and Figure 5 display examples of generated explanations. Both show a neural network trained on the AffectNet facial expression corpus [33]. For training the model we considered following facial expressions: happy, sad, fear, anger, surprise, disgust, contempt and neutral.

Figure 4 presents the NOVA interface for generating explanations with LIME. In the present example explanations for the top four predicted classes are given. However, the number of considered classes may be changed by the user. Moreover, coherent with LIME, additional options can be altered like the number of samples or the number of features. Furthermore, for the generation of explanations the user can either choose from a list of models that have been trained with the help of NOVA for the given modality or drag and drop models from a different source.

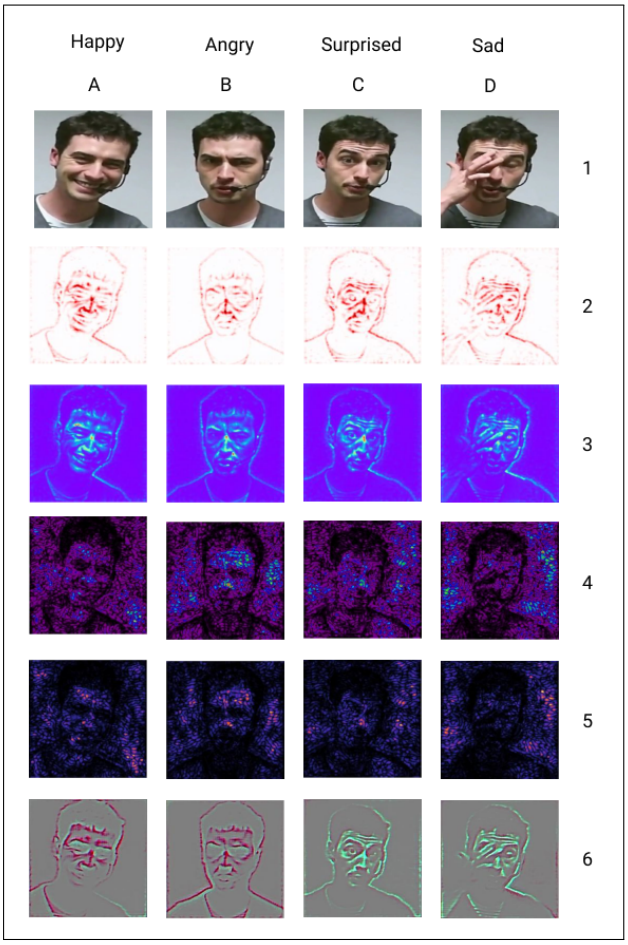

Fig. 5. Visual explanations generated with the iNNvestigate framework. Letters A to D represent the different predicted emotions noted above. The Numbers on the right map onto the following approaches: 1 .Original image, 2. Guided backpropagation, 3. Deep Taylor, 4. LRP Epsilon, 5. LRP Z, 6. LRP Alpha Beta

In the displayed case in Figure 4 the predicted top class has been happy, followed by anger, neutral and surprise. The green shapes represent areas of the original image that have been important for the prediction. In contrast to that the red shapes describe areas that spoke against a particular prediction. As

one would intuitively guess, an interesting area for recognizing whether a person is happy, is the space around the mouth to see if the person is smiling. Moreover, the same area is a strong evidence against the presence of anger, neutral and surprise, which is highlighted by a red area in the other images. Despite the fact that the used model predicted the correct class with an accuracy of 90.5% there is evidence in the explanations present that the model still has flaws. The fact that various areas of the background have been considered important for the prediction, even though there is no relevant information visible, shows that the model isn't perfectly optimized for the given use case.

Alongside the explanation generated by LIME, NOVA also offers the possibility to create explanations with iNNvestigate. The corresponding NOVA interface not only provides a variety of algorithms implemented in iNNvestigate, but also allows the user to decide between different visualization representations. Figure 5 displays an excerpt of some algorithms and visualizations for different facial expressions. The class that has been predicted by the model is written above of the original images. For Figure 5 A all algorithms highlighted the central area of the face - including the eyes, nose and mouth - as important elements regarding the prediction. In case of the angry face (Figure 5 B) the visualizations 4 and 5 show a stronger emphasis on the forehead and eyebrow area which is what would be expected as the bending of the eyebrows is a common indicator for anger. Similar is true for Figure 5 C here especially visualization 2 and 3 highlight amongst other areas the forehead which displays an intensely furrowed brow. Before covering the last facial expression we want to emphasize the fact that similar to LIME, the algorithms used in 4 and 5 all highlighted to some degree areas in the background of the original image, which corroborates the hypotheses that the model isn't fully optimized and bases the prediction to some extent on irrelevant information.

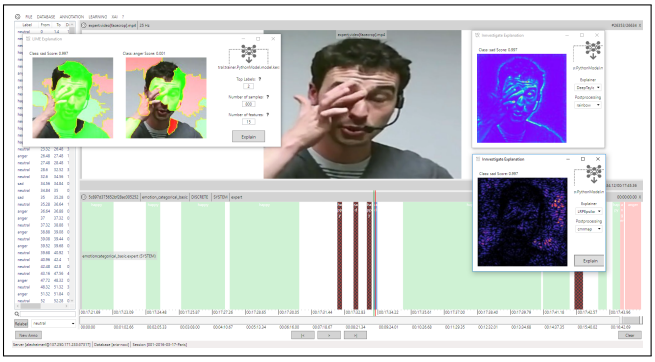

Fig. 6. An instance of the NOVA user interface with visual explanations for a particular frame.

Figure 5 (D) displays an interesting case in terms of prediction and generated explanation. Just by visually exploring the image one could easily agree that the person is sad because he just might have shed a tear and is trying to wipe it away with his hand. Also the explanations generated by the different algorithms stress the areas covering the hand and eyes. However, if one would examine the moment short before and after the specific frame it would become obvious that the person has not been sad at all and probably has just rubbed its eye. This way it becomes evident that for a correct interpretation it is vital to also consider context information. NOVA offers, besides the generation of explanations through state of the art algorithms, the possibility to investigate relevant information before and after a specific frame being part of a video or feature stream. Figure 6 shows a possible setup when working with NOVA. In the presented screenshot an annotation and the corresponding video is loaded. The frame of interest is the earlier discussed alleged sad facial expression. To gain additional insight explanations with LIME and iNNvestigate have been generated.

VI. Conclusion

In this paper we presented the latest advances in the NOVA Annotation tool. NOVA offers a collaborative workflow for multiple types of annotation tasks. Additionally it provides interfaces to machine-learning techniques that allow even non-experts to make use of these technologies in order to speed up the annotation labour. Finally, NOVA not only enables users to apply machine learning techniques, but also provides capabilities to use the latest explainable AI techniques on pre-trained as well as self-trained models, so that users get a better understanding when they can trust their model and what might cause issues, respectively when more training examples are required. Summing up, the described methodology offers transparency from two directions. By observing the output of the classifier, the user can assess its performance and also trace how it changes with new input. In addition, visualising the input to the classifier (raw media or feature streams) can provide hints why a prediction was successful in one place but failed in another. For instance, the user may find out that predictions were wrong due to failure of the tracking algorithm. This way, users also learn in which situations they can trust the model, while the model learns from the user's inputs. NOVA provides a true cooperative workflow between humans and machines. By default, it provides a wide range of machine learning algorithms and feature extraction pipelines. More technically interested users can also extend NOVA's ML tools by adding new templates. This way NOVA is not limited to current state of the art methods such as Deep Neural Networks but is also extendable in the future.

NOVA is open-source software and available on Github: https://github.com/hcmlab/nova

REFERENCES

[1] J. M. Girard, J. F. Cohn, L. A. Jeni, S. Lucey, and F. De la Torre, "How much training data for facial action unit detection?" in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1. IEEE, 2015, pp. 1–8.

[2] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes, "Do we need more training data or better models for object detection?." in *BMVC*, vol. 3. Citeseer, 2012, p. 5.

[3] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: a professional framework for multimodality research," in *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006, Genoa, Italy, May 22-28, 2006.*, N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk, and D. Tapias, Eds. European Language Resources Association (ELRA), 2006, pp. 1556–1559.

[4] M. Kipp, "Anvil: The video annotation research tool," in *Handbook of Corpus Phonology*. Oxford University Press.

[5] T. Schmidt, "Transcribing and annotating spoken language with exmaralda," in *Proceedings of the International Conference on Language Resources and Evaluation: Workshop on XML based richly annotated corpora, Lisbon 2004*. ELRA, pp. 879–896, eN.

[6] R. Cowie, E. Douglas-Cowie, S. Savvidou*, E. McMahon, M. Sawey, and M. Schröder, "'feeltrace': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

[7] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *IJSE*, vol. 3, no. 1, pp. 1–17, 2012.

[8] J. M. Girard, "Carma: Software for continuous affect rating and media annotation," vol. 2, no. 1, p. e5.

[9] J. M. Girard and A. G. C. Wright, "DARMA: Dual Axis Rating and Media Annotation."

[10] B. Settles, *Active Learning*, ser. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012.

[11] E. Kamar, S. Hacker, and E. Horvitz, "Combining human and machine intelligence in large-scale crowdsourcing," in *International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2012, Valencia, Spain, June 4-8, 2012 (3 Volumes)*, W. van der Hoek, L. Padgham, V. Conitzer, and M. Winikoff, Eds. IFAAMAS, 2012, pp. 467–474.

[12] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, pp. 45–66, 2001.

[13] M. Wang and X. Hua, "Active learning in multimedia annotation and retrieval: A survey," *ACM TIST*, vol. 2, no. 2, pp. 10:1–10:21, 2011.

[14] M. Stikic, K. V. Laerhoven, and B. Schiele, "Exploring semi-supervised and active learning for activity recognition," in *12th IEEE International Symposium on Wearable Computers (ISWC 2008), September 28 - October 1, 2008, Pittsburgh, PA, USA*, 2008, pp. 81–88.

[15] Y. Zhang, E. Coutinho, Z. Zhang, C. Quan, and B. W. Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, Z. Zhang, P. Cohen, D. Bohus, R. Horaud, and H. Meng, Eds. ACM, 2015, pp. 275–278.

[16] J. Poignant, M. Budnik, H. Bredin, C. Barras, M. Stefas, P. Bruneau, G. Adda, L. Besacier, H. K. Ekenel, G. Francopoulo, J. Hernando, J. Mariani, R. Morros, G. Quénot, S. Rosset, and T. Tamisier, "The CAMOMILE collaborative annotation platform for multi-modal, multilingual and multi-media documents," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*, N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. European Language Resources Association (ELRA), 2016.

[17] S. Hantke, F. Eyben, T. Appel, and B. W. Schuller, "ihearu-play: Introducing a game for crowdsourced data collection for affective computing," in *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015, Xi'an, China, September 21-24, 2015*. IEEE Computer Society, 2015, pp. 891–897.

[18] J. Cheng and M. S. Bernstein, "Flock: Hybrid crowd-machine learning classifiers," in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, 2015, pp. 600–611.

[19] B. Kim and B. Pardo, "I-SED: an interactive sound event detector," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI 2017, Limassol, Cyprus, March 13-16, 2017*, 2017, pp. 553–557.

[20] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," *AI Magazine*, vol. 35, no. 4, pp. 105–120, 2014.

[21] S. Rosenthal and A. K. Dey, "Towards maximizing the accuracy of human-labeled sensor data," in *Proceedings of the 15th International Conference on Intelligent User Interfaces, IUI 2010, Hong Kong, China, February 7-10, 2010*, C. Rich, Q. Yang, M. Cavazza, and M. X. Zhou, Eds. ACM, 2010, pp. 259–268.

[22] S. Amershi, J. Fogarty, A. Kapoor, and D. S. Tan, "Overview based example selection in end user interactive concept learning," in *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology, Victoria, BC, Canada, October 4-7, 2009*, A. D. Wilson and F. Guimbretière, Eds. ACM, 2009, pp. 247–256.

[23] S. Amershi, M. Chickering, S. M. Drucker, B. Lee, P. Y. Simard, and J. Suh, "Modeltracker: Redesigning performance analysis tools for machine learning," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, 2015, pp. 337–346.

[24] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," vol. PP, no. 99, pp. 1–1, 2017.

[25] E. H. Shortliffe and B. G. Buchanan, "A model of inexact reasoning in medicine," *Mathematical Biosciences*, vol. 23, no. 3, pp. 351 – 379, 1975.

[26] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[27] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K. Müller, S. Dähne, and P. Kindermans, "innvestigate neural networks!" *CoRR*, vol. abs/1808.04260, 2018. [Online]. Available: http://arxiv.org/abs/1808.04260

[28] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, p. 1096, 2019.

[29] Y. Zhang, A. Michi, J. Wagner, E. Andr, B. Schuller, and F. Weninger, "A generic human-machine annotation framework based on dynamic cooperative learning," *IEEE Transactions on Cybernetics*, pp. 1–10, 2019.

[30] J. Wagner, F. Lingenfelser, T. Baur, I. Damian, F. Kistler, and E. André, "The social signal interpretation (ssi) framework: multimodal signal processing and recognition in real-time," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 831–834.

[31] J. Wagner, T. Baur, D. Schiller, Y. Zhang, B. Schuller, M. Valstar, and E. André, "Show me what youve learned: Applying cooperative machine learning for the semi-automated annotation of social signals," *IJCAI/ECAI, Workshop on eXplainable Artificial Intelligence (XAI) 2018*, p. 171, 2018.

[32] J. Wagner, T. Baur, Y. Zhang, M. F. Valstar, B. Schuller, and E. André, "Applying cooperative machine learning to speed up the annotation of social signals in large multi-modal corpora," *arXiv preprint arXiv:1802.02565*, 2018.

[33] A. Mollahosseini, B. Hassani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *CoRR*, vol. abs/1708.03985, 2017. [Online]. Available: http://arxiv.org/abs/1708.03985