Maintaining the Humanity of Our Models

Umang Bhatt
Carnegie Mellon University
Pittsburgh, PA 15213, USA
umang@cmu.edu

Abstract

Artificial intelligence (AI) and machine learning (ML) have been major research interests in computer science for the better part of the last few decades. However, all too recently, both AI and ML have rapidly grown to be media frenzies, pressuring companies and researchers to claim they use these technologies. As ML continues to percolate into the layman's life, we, as computer scientists and machine learning researchers, are responsible for ensuring we clearly convey the extent of our work and the humanity of our models. In our current discussion, we limit ourselves to the following three important aspects that are needed to regularize ML for mass adoption: a standard for model interpretability, a consideration for human bias in data, and an understanding of a model's societal effects.

Introduction

Mainstream media, any non-academic or non-research outlet, fawn over the tandem of machine learning (ML) and artificial intelligence (AI). Recently, technologies like AlphaGo, competitions like the Netflix Prize, and once sci-fi fantasies like self-driving cars have dominated news headlines. The media is correct in claiming that, while ML is outperforming humans at clerical and pattern-driven work, the next wave of AI will revolutionize medicine, law, finance, and transportation by processing data more efficiently than humans (Grace and Salvatier 2017). It is not wrong to be proud of and eager about the advances made in these fields annually. AI can be compared to the steam engine and electricity: powerful general-purpose technologies that can forever alter the fabric of society (Brynjolfsson and McAffee 2014). However, it is erroneous to overstate these technologies' capabilities in the immediate future, which we define hereafter as ~1-2 years. AI growth is slowly yet drastically automating aspects of the monotony in our lives (Schwab 2016).

As AI enters the limelight and displaces all, regardless of the color of their collar, researchers and practitioners of the field must poise the resultant models to be interpretable, unbiased, impactful, and thus humane (Kaplan 2015). In our discussion, we define humanity and humane to be the ethereal and emotional impact of these models on humans. We define AI as encompassing its subfield of ML. In order to build a ML system that values humanity, we consider the following questions: (1) How can researchers make their work interpretable for the end user? (2) How can researchers ensure their algorithms are not learning now unlawful or immoral patterns from antiquated data? (3) How can researchers evaluate the societal effects of their predictions?

We believe these three questions provide the foundation needed to succeed in maintaining the humanity of the models we create. To scale to the masses, ML systems need be interpretable to a non-expert. Laymen should be able to understand the sequence of steps and data points used (and their respective weights) to achieve the final result. ML systems must draw from data that researchers have vetted for potential social bias, thus ensuring the fairness of the eventual conclusion. This is an overlooked portion of current ML work: most researchers claim themselves to be data-agnostic; however, it is imperative they care about the features, source, and context of datasets (O'Neil 2016). Finally, ML systems must be aware of the user impact of each prediction made and each pattern found. Having a pointed, narrow goal with low impact is the current rule of thumb to ensure little disruption in other parts of a user's life (Armstrong and Levinstein 2017). To that end, we dive into the need for all three pillars, as the fields of AI and ML continue to evolve.

Model Interpretability

Imagine a patient visiting a doctor in 2030. They walk into an empty room filled with sensors and large screen with necessary instructions. Once the minimum readings have been made (non-invasively and implicitly), the patient can see a diagnosis (e.g. Diabetes) generated automatically by

a black box. If researchers are not cognizant of the implications of their predictions, delivering a potentially life-changing diagnosis in such an insensitive manner can stifle the adoption of AI systems, since the system lacks humanity in diagnosis. As Manuela Veloso once said, "If we don't worry about the explanation [of the result], we won't be able to trust the systems." We, as researchers and practitioners, need to ensure our current black box models gain *clear-box* access to allow end users to reason about our prediction. Therefore, researchers must prioritize exposing the inner workings of ML systems to promote interpretability - the explanation behind predictions – thus bringing the world more personable, humane models.

Current State

ML today begets a robust strength in prediction power in decision-making processes (at least in the supervised case, which we assume from here). However, due to a mismatch between prediction objectives (i.e. test set performance) and the real world costs of deployment, there is an unfulfilled demand for interpretability (Lipton 2017). As the final users of ML systems are typically non-experts, models lacking interpretability are rendered ineffective and useless. Though there exists no concrete definition of interpretability, it broadly refers to explaining a model in humanly understandable terms: many desiderata for modern ML systems, like robustness, fairness, and trust, are also commonly grouped with interpretability (Doshi-Velez and Kim 2017).

There exists a need for rigorously standardizing interpretability, since the European Union will prevent automated individual decision-making this year (Goodman and Flaxman 2016). As of now, dimensionality reduction techniques like backward feature selection on a single layer perceptron or feature extraction via principle component analysis suffices to make a model interpretable in simple cases (Vellido 2012). Sparse linear classifiers and discretization methods (decision trees, rule sets, etc.) are well-known interpretable models (Kim 2015). However, much interest now lies in the nonlinear, high dimensional models and related deep learning techniques. Researchers working on joint model training techniques are exploiting known interpretable models to provide laymen with explanations for a given prediction.

More recent techniques have actually implicitly prioritized interpretability, albeit void of a standardization. Researchers working on neural modulation for semantic search in visual content are inherently making some ML models more interpretable by employing explicit reasoning and attention.

Case Study: Medicine

Returning to the 2030 scenario, the patient demands an explanation of how a complex model, like Doctor AI ("a generic predictive model that covers observed medical conditions and medication uses"), came to its diagnosis (Choi et al. 2016). Though the model might be confident about its prediction, it must expose the sequence of decisions that led to the conclusion. One option would be jointly training a recurrent neural network, a long short-term memory (LSTM) per se, with a hidden Markov model (HMM) to expose the HMM state sequences to the end user (Krakovna and Doshi-Velez 2016). This technique leverages both the predictive power of an LSTM and the explicit states of an HMM: this even unlocks transfer learning as an LSTM model trained on a sufficiently large electronic health record can be transferred to any hospital (Choi et al. 2016). However, a major shortcoming of this approach is that a domain expert must be leveraged to name the states of the HMM: it is nearly impossible for a computer scientist to attempt to name a given state sequence of symptoms and vital signs as potentially contributing to a particular diagnosis. In some simpler planning tasks, expert knowledge is taken into account in the prior distribution over the area of interest, but this does not generalize well to all situations (Kim 2015). Nonetheless, coupling combined model training with test set performance on the top-k ICD-9 codes[1] can produce accurate and interpretable results (Lipton and Kale 2015, Nigam 2016). Another such technique for making these predictively powerful LSTMs more explainable is employing input gradients to generalize decision logic, which is irrespective of the dataset (Ross, Hughes, Doshi-Velez 2017). These techniques are all means towards the end of making our ML models more interpretable and thus more humane.

Human Bias in Data

The source and features of data used as a basis for our models are essential to understanding the inherent human bias in a model's predictions. When productionalizing a model, we must divulge the exact source and features of the data used to train that model. Data, contrary to layman's thoughts, ages and grows stale. Imagine if researchers used data from the Jim Crow days to predict in which zip codes are people most likely to go to jail again (O'Neil 2016). Overtime, the data from yesteryear becomes irrelevant. So, can researchers not just create a threshold or add a layer of logistic weight to our data by recency? Well, a recency bias is just as unproductive (Abah 2016). Acknowledging the existence of and taking steps to correct

[1] The authors pick the top k most frequent ICD-9 (alphanumeric codes for patient diagnosis) and classify the accuracy of our model on those codes.

this potentially unfair data yields more humane models, as an unbiased model fed biased data gives a biased result.

Current State

When assessing the quality/recency of and reducing the human bias of a dataset, two techniques are common. One technique is debiasing, which manually severs the learned relationship between two entities. In example, gender bias in natural language generation from processing/training on text corpuses is all too common. A gender bias-free dataset of images can be created when we place constraints on certain relationships between entities within the images (Zhao et al. 2017). In a text generation algorithm, gender bias can be mitigated by identifying known gender biased words, working in a gender neutral subspace, and understanding the distance of a gender neutral world towards the preidentified gender subspaces (Bolukbasi et al. 2016). Another technique is simply omission of the stale or biased data from training; it is trivial to state, but such a decision is lossy and certain patterns in the data will be missed.

It is crucial to note that in both scenarios, researchers are imposing their own bias and morality on a given problem space. For example, if researchers think (or even empirically show) that zip code of residence is a high predictor of where crime occurs, they are then faced with a moral struggle of whether or not to patrol more in those zip codes, disadvantaging the portion of non-criminals in a zip code deemed crime prone. The legality of models matters considerably as an ounce of human bias can violate the law (Samek 2017). To that end, we show a need to remove human bias disparities with as little impact on accuracy as possible (Johndrow and Lum 2017).

Case Study: Recidivism

Recidivism prediction (that is, the propensity of a person to return to jail once released) is bursting with social bias. Though models like PredPol[2] exist, there is no formal feedback loop for all involved parties; thus, there exists a lack of randomness in the data (Ensign et al. 2017). Without this randomness, a human bias is propagated in the data (e.g. only patrol neighborhoods of criminals who are currently imprisoned). Unfortunately, researchers lack a method to understand the fairness of their predictions, other than the false positive rates of two subgroups within the population in question. One suggestion is to optimize parameter instability and disparity (Chouldechova and G'Sell 2017). More interpretably, one can perform a subset scan to detect if a given class has noteworthy bias for in a given subgroup (Zhang and Neill 2017). Such techniques only arise if researchers heed human biases in data, which will be of utmost importance as ML adoption continues to skyrocket.

[2] PredPol allows law enforcement to predict where crime will happen given historical/real-time data feeds and then assigns patrol units accordingly (Ensign et al. 2017).

Societal Effects

The output of ML systems affects real flesh and blood beings. Unfortunately, all too often, researchers lose sight of this reality. Some researchers focus on optimizing objectives on benchmark datasets instead of the real world applications of the code they write (Wagstaff 2012). They want to be able to transfer their expertise and models to new domains, wherein ML can augment archaic practices and automate pattern-based predictions. For example, clothing companies no longer use only intuition and actuarial science to forecast their products' performance, instead they also use models that incorporate seasonality, user preferences, and industry trends to decide what type of clothing should be designed next season (Brynjolfsson and McAffee 2014). In confluence with the proliferation of ML use cases, we must remain cognizant of the legality of our models and predictions and be alert of user intent and reception.

Current State

Society benefits from ML models daily. These models tell us what stocks to buy, how much demand a restaurant can expect next quarter, what country poses the most threat to another, whom we should date, etc. (Ross 2016). Society seems like it is subject to the output of these models, and thus mainstream media often misinterprets the power of ML.

For example, in the realm of natural language processing, many recent works report that in multi-agent environments, where agents communicate via strings of tokens to perform a given task, grounded and compositional language naturally emerges. Though this may be the case in controlled circumstances, we cannot generalize this to say: "AI agents make their languages and thus we need to shut them down," as many media claim (Lewis et al. 2017). Upon review, it becomes evident that language cannot emerge naturally and systems are shut down due to a lack of human interpretability: that is, one AI agent may say "Red man ball sit!" to another agent, who understands that to mean "Hello, how are you?" in English – without human intervention, the agents communicate in a nonsensical, incomprehensible grammar, basically gibberish, thus stressing the need for the first pillar of interpretability (Kottur et al. 2017).

As mass ML adoption is imminent, being mindful of such misinterpretations and effectively communicating the limits of ML must be kept at the top of our minds.

Case Study: Pricing

In the e-commerce world, companies optimize models to maximize profit or increase purchase frequency. One such model is a dynamic pricing engine, which prices goods based on the targeted consumer's willingness to pay. As such, these engines are used to serve the *optimal* price for a given user to maximize company profits. Plagued by sparse user level data and by legal constraints on what features can and cannot be used, dynamic pricing experts manage programs like time-limited coupons forecasted via a point-process model that makes real-time, global estimates based on transaction history and patterns (Manzoor and Akoglu 2017). Such pricing programs must be interpretable and unbiased; if they are not, the societal consequences of erroneous prices (or worse, of price discrimination) are catastrophic for a company. Being aware of and responsive to the implications of ML models is the final key towards more humane and adoptable models.

Conclusion

To be prepared for mass adoption of machine learning systems, we, as researchers and practitioners, must adopt a framework for developing humane models that ensure interpretability, unbiasedness, and practicality. By creating a rigorous standard for machine learning interpretability, we can transform the medical predictive analytics industry. By understanding the inherent human bias in the data we collect and the sample it represents, we can ensure that we build a more unbiased model for police patrol. By thinking deeply about the societal effects and ethicality of our predictions, we can ensure we deliver profitable and fair prices in the e-commerce industry. All three pillars can displace society's perception of machine learning, as the true power and beauty of how we can use autonomous agents and machine learning comes to fruition when we maintain the humanity of our models.

Acknowledgements

Many thanks to my advisors/mentors at CMU: Jose M.F. Moura, J. Zico Kolter, and David O'Hallaron for their valuable discussions and to my research partners: Satwik Kottur, Edgar Xi, Shouvik Mani, and Sam Fazel for their continued support.

References

Abah, J. 2016. Recency Bias in the Era of Big Data: The Need to Strengthen the Status of History of Mathematics In Nigerian Schools. *In Advances in Multidisciplinary and Scientific Research Journal*.

Armstrong, S., and Levinstein, B. 2017. Low Impact Artificial Intelligences. *arXiv: 1705.10720*

Bird, S., Barocas, S., Crawford, K., Diaz, F., and Wallach, H.. 2016. Exploring or Exploiting? Social and Ethical Implications of Autonomous Experimentation in AI. *Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2016*. New York, NY.

Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., Kalai, A. 2016 Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv: 1607.06520*

Brynjolfsson, E., and McAffe, A. 2014. The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. *WW Norton & Company.*

Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *In Proceedings for 2016 Machine Learning and Healthcare Conference.* Los Angeles, CA

Chouldechova, A. and G'Sell, M. 2017. Fairer and more accurate, but for whom? *In Proceedings for FAT/ML 2017.* Halifax, NS, Canada.

Doshi-Velez, F., and Kim, B. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: 1702.08608*

Doshi-Velez, F. Kortz, M, et. al. Accountability of AI Under the Law: The Role of Explanation. *arXiv: 1711.01134*

Ensign, D., Friedler, S., Neville, S., Scheidegger, C., Venkatasubramanian, S. 2017 Runaway Feedback Loops in Predictive Policing. *In Proceedings for FAT/ML 2017*. Halifax, NS, Canada.

Frank, B. September 19, 2017. You might use AI, but that doesn't mean you're an AI company. *VentureBeat*.

Goodman, B. and Flaxman, S. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". *In Proceedings for 2016 ICML Workshop on Human Interpretability in Machine Learning,* New York, NY.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O. 2017. When Will AI Exceed Human Performance? Evidence from AI Experts. *arXiv: 1705.08807*

Grbovic, M., Radosavljevic, et. al. 2016. E-commerce in Your Inbox: Product Recommendations at Scale. *In Proceedings for KDD 2015*. Sydney, Australia.

Johndrow, J. and Lum, K. 2017. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv: 1703.04957*

Kaplan, J. 2015. Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence. *Yale University Press.*

Karpathy, A. May, 31, 2017. AlphaGo, in context. *Medium.*

Kim, B. 2015. Interactive and interpretable machine learning models for human machine collaboration. PhD diss., Massachusetts Institute of Technology, 2015.

Kottur, S., Moura, J., Lee, S., Batra, D. 2017. Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog. *In Proceedings for EMNLP 2017.* Denmark.

Krakovna, V. and Doshi-Velez, F. 2016. Increasing the Interpretability of Recurrent Neural Networks Using Hidden Markov Models. *In Proceedings for NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*. Barcelona, Spain.

Lewis, M., Yarats, D., Dauphin, Y., Parikh, D., Batra, D. 2017. Deal or No Deal? End-to-End Learning for Negotiation Dialogues. *arXiv: 1706.05125*

Lipton, Z. 2017. The Mythos of Interpretability. *In Proceedings for 2016 ICML Workshop on Human Interpretability in Machine Learning,* New York, NY.

Lipton, Z., Kale, D., Elkan, C., Wetzel, R. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks. *arXiv: 1511.03677*

Manzoor, E., and Akoglu, L. 2017. RUSH! Targeted Time-limited Coupons via Purchase Forecasts. *In Proceedings for KDD 2017*. Halifax, NS, Canada.

Nigam, P. 2016. Applying Deep Learning to ICD-9 Multi-label Classification from Medical Records. *Stanford University*

O'Neil, C. 2016. Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Broadway Books.*

Reese, H. 2016. Transparent machine learning: How to create 'clear-box' AI. *Tech Republic.*

Ribeiro, M., Singh, S., Guestrin, C. 2016. Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance. *arXiv:1611.05817*

Ross, A. 2016. The Industries of the Future. *Simon & Schuster Paperbacks.*

Ross, A., Hughes, M., Doshi-Velez, F. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. *arXiv:1703.03717*

Schwab, K. 2016. The Fourth Industrial Revolution. *Crown Business.*

Samek, W., Wiegand, T., Muller, K.R. 2017. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv: 1708.08296*

Vellido, A., Martin-Guerreo, J., Lisboa, P. 2012. Making Machine Learning Models Interpretable. *In Proceedings for European Symposium on Artificial Neural Networks, Computational Intelligence, and Machine Learning 2012*. Bruges, Belgium.

Wagstaff, K. 2012. Machine Learning that Matters. *In Proceedings for the 29th International Conference on Machine Learning.* Edinburgh, Scotland, UK

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. *arXiv: 1707.09457*

Zhang, Z. and Neill, D. 2017. Identifying Significant Predictive Bias in Classifiers. *arXiv: 1611.08292*