EPISODE 88

[INTRODUCTION]

[0:00:10.5] SC: Hello and welcome to another episode of TWiML Talk, the podcasts where I interview interesting people doing interesting things in machine learning and artificial intelligence. I'm your host, Sam Charrington.

This week on the podcasts we're featuring a series of conversations from the NIPS Conference in Long Beach, California. This is my first time at NIPS and I had a great time there. I attended a bunch of talks and of course learned a ton. I organized an impromptu roundtable on building AI products and I met a bunch of wonderful people, including some former TWiML Talk guests. I'll be sharing a bit more about my experiences at NIPS VMI newsletter, which you should take a second right now to subscribe to at twimlai.com/newsletter.

This week through the end of the year, we're running a special listener appreciation contest to celebrate hitting 1 million listens on the podcasts and to thank you all for being so awesome. Tweet us using the #twiml1mil to enter. Everyone who enters is a winner and we're giving away a bunch of cool TWiML swag and other mystery prizes. If you're not on Twitter or want more ways to enter, visit twimlai.com/twiml1mil for the full rundown.

Before we dive in, I'd like to thank our friends over at Intel Nervana for their sponsorship of this podcast and our NIPS Series. While Intel was very active at NIPS with a bunch of workshops, demonstrations and poster sessions, their big news this time was the first public viewing of the Intel Nervana Neural Network Processor, or NNP. The goal of the NNP architecture is to provide the flexibility needed to support deep learning primitives while making the core hardware components as efficient as possible, giving neural network designers powerful tools for solving larger and more difficult problems while minimizing data movement and maximizing data reuse. To learn more about Intel's AI products group and the Intel Nervana NNP, visit intelnervana.com.

In this episode I sit down with Timnit Gebru, postdoctoral researcher at Microsoft Research and the Fairness, Accountability, Transparency and Ethics in AI, or FATE Group. I've been following Timnit's work for a while now and was really excited to get a chance to sit down with her at the conference. We packed a ton into this conversation, especially keying in on her recently released paper using deep learning and Google Street View to estimate the demographic makeup of the U.S.

Timnit describes the pipeline she developed for this research and some of the challenges she faced building an end-to-end model based on Google Street View images, census data, and commercial car vendor data. We also discussed the role of social awareness in her work including an explanation of how domain adaptation and fairness are related and her view on the major research directions in the domain of fairness. TImnit is also one of the organizers behind the Black in AI Group, which held a very interesting symposium and poster session at NIPS. I'll link to the group page in the show notes. This was a really interesting conversation and one that I'm sure you'll enjoy.

[INTERVIEW]

[0:03:38.8] SC: Timnit, welcome to the podcasts.

[0:03:40.4] TG: Thank you. Thanks for having me.

[0:03:41.7] SC: Absolutely. What is the FATE Group at Microsoft Research?

[0:03:45.6] TG: FATE stands for fairness, accountability, transparency in ethics and AI and it's a very new group started by Kate Crawford and Hanna Wallach and there are some other people there like Jen Wortman Vaughan and some economists and some computational social scientists. It's a combination of machine learning people and social science and economics people trying to study the societal implications of AI and just make sure that we create algorithms that are fair. Our research is focused towards that.

[0:04:16.5] SC: Wow! How did you get interested in fairness in AI in particular and artificial intelligence in general?

[0:04:23.4] TG: My background is in computer vision, and as I was working on — There's a number of things I've always been interested in social justice and towards the end of my Ph.D. I saw this Propolica article about a software that was being used by judges.

[0:04:38.1] SC: I remember that one.

[0:04:39.0] TG: Yeah, crime [inaudible 0:04:39.0] to figure out a person's likelihood of committing a crime again and judges were — This was one of the inputs that they used to figure out how many years they should sentence you into prison.

[0:04:52.8] SC: This being some machine learning algorithm.

[0:04:54.7] TG: Mm-hmm. That was sold by Northpionte, I think is the name of the startup, and that was very terrifying for me knowing because I had the background to know like what kind of vices we have in the criminal justice already and what kind of vices we have in — How much discrimination there is in the data that would be trained for it.

That was one, and then while I was working on my Ph.D., I kind of figured that my work, my own work could be susceptible to this kind of bias as well, because my whole work was trying to show that we can do data mining using images, so large scale computer vision plus data. Most people use text and social networks and other kinds of textual data to do data mining and the whole point of my Ph.D. was to show that we could gain useful societal information using images.

If the ground truth for that that you use to train is bias, you're going to have biased conclusions. I thought that I should be very cognizant of what kinds of issues could exist with that type of work given that my work lies in that type of work.

[0:05:58.2] SC: Okay. You're relatively new at Microsoft. Is that right?

[0:06:02.0] TG: I am. Yeah. I started in July. Yeah, I'm very new.

[0:06:05.3] SC: Awesome. And you just published your first paper with the group? Is that right?

[0:06:09.0] TG: No. My paper that just came out in [inaudible 0:06:11.1] is actually from my Ph.D. This is a project that took four years. I think it's more — When I give talks about it, I think people understand the level, the amount of work it went into, but it's harder to see, I think, just from that one paper. Yeah, that paper took a very long time and it just got published.

[0:06:29.6] SC: Wow! Tell us about that paper.

[0:06:31.5] TG: This paper was using Google Street View images to predict demographic characteristics. What we did was we detected and classified cars, all the cars in 15 million Google Street View images across 200 American cities. Then we were able to use the characteristics of the cars that we detected and classified in a particular zip code and precinct and associated that with certain demographic characteristics, like income or political affiliation.

An earlier paper we had, we looked at income segregation levels and even like C02 emission rates. Once we detected and classified the cars, we represented each geographic region, like basically for us it would be like a zip code or a precinct, like a voting precinct. We represented by like the features of the cars that are in that zip code.

For example, the percentage of Hondas or the percentage of each make that you have, that's one feature. Like percentage of Toyotas or Hondas or like Nissans or whatever, percentage of sedans, other metadata like the average miles per gallon, like efficiency of that particular zip code, etc. Once we had that information, then we used round truth census data or other data depending on what we're trying to predict to train another model to go from the car features to like predicting demographic characteristics. That's the project.

[0:07:58.1] SC: Interesting. You mentioned census data as part of your training dataset. When I think of the kinds of things that you're trying to predict for, like the wealth or income of geographic area, that's already in the census data. How would someone use this technique?

[0:08:17.4] TG: What we did, we had 200 cities in our dataset, right? We used a subset of our cities for training. We assume we have census data for like, let's say, I think is it 13% of — Or we sweep it to see like how much training data we need. We assume we have census data for a very small subset of the cities that we have and then we trained a model using. Then for the rest of the cities, we don't' have census data, we only have images and cars.

I guess the way someone would use it is if you have census data for some cities and you want to try to see like for other cities what the data might be, you could train our model using the census data that you have and then the cars that are detected in the other cities. That's how we used it.

We also did some experiments and trying to do this across time. So say you have data for past census data for New York, like we're using Google Street View time lapse data for New York, and you have a bunch of images as well. Can you then try to predict what's going to happen in the future or before you have the census data? We did some experiments like that as well.

For me, this is kind of — Like I don't want people to read too much into the cars. For me, it's a proof of concept. Basically like a new tool that you can use to do this kind of analysis, like demography or social science applications or work. It's like a new tool that is available to researchers and we want to show if one were — Say you wanted to study the — I don't know, relationship between trees, tree species and people's health or something. How would you go about using images to do that?

Now that you saw our paper, you could like apply a very similar pipeline to it. That's more what I want to take away to be.

[0:10:11.0] SC: It's not specifically about cars, projecting income based on cars. It's more about we've got all of these visual data from cameras and sensors and things like that. How can we use that as proxy for any other thing that we might want to —

[0:10:26.0] TG: For us, like cars — There are other things you could stay where the only data you could probably have is probably only visual data. For cars, you can argue that that's not necessarily the case. You can use DMV data. I guess Street View gives you a different perspective, which is you're not looking at the cars of the people who necessarily live there. You're just saying, "If I were to just walk around the street, what does that street look like? What kind of cars are driving? What kind of cars are parked?" What does that tell me about the people who live there?

For me I'm most excited about the tool and this pipeline, and I was very, very surprised that our thing actually worked, because there's a lot of stuff. I was very surprised. There are a lot of stuff that could go wrong, because the pipeline, there are many, many different components of it.

[0:11:18.7] SC: Can you walk us through the pipeline?

[0:11:19.8] TG: Yeah. The biggest thing is a lot of people in AI don't talk about this, but its data collection is huge. If you want to do any sort of supervised machine learning and you want to do it in the real world. We're not talking about a toy dataset here. So we were saying, "What is our end goal?" Our end goal is to detect and classify all the cars in 50 million Google Street View images and then to predict demographics using that, right?

To get to that end goal we first have to figure out, "Oh, okay. How are we going — Okay, how are we going to get data? How are we going to get labeled data? How are we going to label cars in Google Street View images?" This is very hard.

[0:11:58.9] SC: You mean getting the data, is that easily accessible via an API or did you have to scrape them or —

[0:12:03.3] TG: Google images?

[0:12:04.3] SC: Yeah.

[0:12:04.9] TG: Yeah, they have an API, and these are publicly available images. Then we have to be like, "Okay. What are all the different types of cars around that we might see Google Street View images? Where do we get that list?"

We found edmunds.com and they have all the cars since 1990. There are about 15,000 types of cars. Guess what? A computer vision algorithm can only — We can only really kind of classify cars based on what they look on the outside. A subset of these 15,000 cars look the same, because they don't change them from year-to-year or from to trim-to-trim or whatever. We had to figure out how to cluster the cars that looked the same. We had a paper on this, [inaudible

0:12:47.7] paper. It was more of an ACI network. How do we get our initial subset of classes where we bucket like cars that look the same into one class?

Then that process in and of itself took a few months, by the way, because you try something, it doesn't work. You try something else. We did that.

[0:13:07.4] SC: This is unsupervised. You're just clustering the cars that you're seeing in the images or semi-supervised, because you're using Edmunds?

[0:13:13.5] TG: I guess you could say — Yeah, what we do is we use Amazon Mechanical Turk and we show people — It's a graph-based algorithm. We show people two images of cars. We have example images of cars from edumnds.com. We show people, we say, "Are these two cars the same or different?"

[0:13:31.9] SC: Meaning one of them is from Street View and the other is from Edmunds?

[0:13:34.5] TG: No. This is all from Edmunds. Right now we're not even — Okay. We haven't even gotten to getting labeled data right now. We're trying to define what our classes are. What does class one mean? Class one means 2006 Honda Accord, 2006 Honda Accord, 2007 Honda Accord, because all of them look the same. We haven't even started getting data. We're just starting to define what our classes are. That already takes a lot of time. That's the first thing you got to do.

The second one, once you define what your classes are, then for each of those classes you have to have labeled data to train your car detector, right? That's where we need the experts for Google Street View images. We also scraped data from like ecommerce sites, like cars.com and craigslist.com. This is why domain adaptation exists. If you just train a plain, like CNN or some sort of supervised machine learning algorithm on things that look like cars from Craigslist and try to test it on cars, try to detect cars in Google Street View or classify cars in Google Street View, it's not going to work. The distribution looks very, very different.

Then there's a whole like — Then I had an ICCV paper where we had — It's a domain adaptation based paper. I think actually this dataset is a very good domain adaptation dataset.

[0:14:55.4] SC: Why don't we do like an inset here on domain adaption. What's the 30-second overview of domain adaptation?

[0:15:02.6] TG: A domain adaptation is a subset of what people call transfer learning. So domain adaptation is like you have one task, one exact task, and we have something that we call a source domain and a target domain. In our example, let's say the source domain is cars from ecommerce sites, Craigslist.com, and let's say the target domain is cars from like Google Street View images.

What you try to do in domain adaptation if you assume that you have labeled data in the source domain, but in the target domain — In unsupervised adaptation, you assume you have no labeled data. In unsupervised, we would assume that we don't have any images in Google Street View that are labeled with the types of cars they contain. That's unsupervised.

Fully supervised adaptation, we assume that that in Google Street View we have labeled images for all classes. Then in semi-supervised we would assume that in Google Street View, which is our target domain, we have labeled data for a subset of our classes. So that the idea of domain adaptation is when you're training set and your test set have different distributions, how can you best use them within the different domains assuming making these different assumptions that we just talked about. How can you best use data in your social domain, I guess in conjunction with data in your target domain to maximize your accuracy on the target domain?

This is a very — In the real world, this is usually the case, because you'll never going to have — Like camera statistics are different, occlusion or whatever. If you have a training set from like Google Images or search images or something and then you want to apply your model to some other thing that has a different statistic, you need to know about adaptation techniques.

[0:16:48.1] SC: When I think about Google Street View images, all those images have — From the perspective, like they have a very specific kind of look that's different from anything you'd ever see on Edmunds or any other car site.

[0:16:58.9] TG: Yeah, because Edmunds — I'm trying to sell you my cars, so I want to give you the best perspective. Like you have a really nice resolution, it's in one car in the middle. There's no other car occluding this car. There's no trees or like — I don't know. It's very, very different. There are cars in Google Street View where you only —

[0:17:19.9] SC: It's a staged image.

[0:17:21.6] TG: Yeah. There are cars in Google Street View where you only see like a couple of lights or something like that or you have this side view. That's where I'm very interested in the domain adaptation problem because of that.

This project helped me decide what core machine learning and computer vision areas I'm very interested in. Because of this, it's domain adaptation. The second one is data collection, efficient data. Basically, what some people call efficient machine learning. Data efficient machine learning. I would say domain adaption is part of it. How to efficiently collect datasets?

[0:17:56.5] SC: When I think of data efficient machine learning, I think of like one shot, few shot machine learning, that kind of stuff?

[0:18:00.1] TG: Few shot, that kind of — Anything. I haven't done too much on few shot, one shot learning, whatever. I don't know. The semantic sometimes kind of like confuse me. Because data collection does not sound fun, I don't think AI people are — In computer vision, a lot of us work on it, and there's even people doing like a hybrid computer vision HCI kind of work, especially in our labs from the very beginning. A lot of people were always working on data collection, because we don't want to work on toy problems. You can only do so much with mNest. You can only gain so much insight. You know what I mean? Into like what kind of problems we should be solving if you're always just using readily available data so that you can get to the next conference deadline or something like that.

I'm not saying you should always spend so much time collecting data, but I'm saying you should do it at least a couple of times in your life just to see where AI is right no if we want to apply it to the real world. This project really, really — I would say that — I was complaining about it the whole time I was working on it, but now that it's over, it really cemented like what I think is really important to work on. Actually, the issue of bias also came up in this project, because in an earlier paper, we also — We're just trying to say like, "Okay. We can predict this. We can predict that," and one of the things we looked into was crime, crime rates.

With crime rates, as you know the ground truth, whatever ground truth we're going to have is biased, because all we know is who got arrested for the crime and whose crime got reported. If I say, "Hey, look! With images with cars, we can do this," then that's already biased.

Basically, like if I'm going to do this type of work, if I'm going to continue to do this type of work, I have to also be working on the bias and fairness and other types of issues. What's really interesting is that domain adaptation and this whole fairness thing are very, very related actually.

[0:19:52.5] SC: How so?

[0:19:54.0] TG: Some of the techniques that you can even use that some people have even already used are very related. One of the ways in which people do domain adaption is to say — Say I want a classifier. Let's say your classifier has a primary task, which is to classify something, maybe the type of car in my image. So I want that classifier to do well regardless of what domain my image comes from. Whether or not it's Craigslist or — Whether it's Craigslist or Google Street View.

The way some people do this, and there are many variations of this, is they have another classifier, and the other classifier, all it does is it uses the features that its input is the features learned by the first classifier. The input is those features, and the output is which domain the image came from, Craigslist or Google Street View, right?

Then the first classifier — And its job is — In addition to accurately classifying the car, it's job is to also confuse the second classifier, because if the second classifier can't tell based on the features learned by the first classifier, which domain the image came from, then it means you've learned features that are sort of domain invariant. Correct?

[0:21:08.5] SC: Right.

[0:21:09.4] TG: Now, can you see how you might apply this to fairness? Say that you want to classify something, like your risk score or something like this, you want to say —

[0:21:19.3] SC: If my other classifier can identify some class that I don't want to be identified in a dataset.

[0:21:23.2] TG: Yeah. Say you want this to be like invariant to like your race or something. So then you can have another classifier that classifies the raise of the person, and then this first one confuses them. Now, there's work already that does this and it's not like a done deal. It's not solved, because then there's different fairness criteria, and then like which criteria do you use, etc., etc. But I do think that these two things, these two fields are kind of very related, and it's so weird to me, because I didn't think about that.

When I got interested to both of them, I didn't think about that at all and now I'm like, "Oh! Wait a minute."

[0:22:02.0] SC: Interesting. Now, this last thing you were describing, it sounds like the kind of thing we see in adversarial networks, is it —

[0:22:09.5] TG: Yeah, exactly. People have done it with adversarial networks. So you can implement it with adversarial networks. You don't have to implement who the adversarial networks. There are other works that use like a different laws function. There's work called — There is the gradient reversal work from way back. It was like 2014 or something like that. There is also domain confusion laws.

Judy, who is — Well, was a post-doc in our lab and I had a paper at ICCV where I used her laws from like a prior paper of hers which was not adversarial. Again, it's funny that people independently were working on this idea by the time it came with his adversarial networks thing, with his [inaudible 0:22:50.5] thing. Then once he came up with [inaudible 0:22:52.4], they're like, "Oh! Wait a minute. We can't just use [inaudible 0:22:54.8] for this." That's another thing I find interesting, is that like this idea was kind of concurrently being thought of by other people.

[0:23:02.7] SC: Interesting. How do you take this forward into your new role at Microsoft?

[0:23:09.1] TG: At Microsoft, I'm working on a whole bunch of stuff right now. Joy Buolamwini from MIT is here by the way and she and I have a paper that is hopefully going to come out at this new Fairness Conference; Fairness, Accountability, Transparency and Ethics in AI Conference. It's a whole conference in February.

[0:23:27.8] SC: When is that? In February?

[0:23:28.9] TG: In February. It's in New York.

[0:23:29.9] SC: Okay.

[0:23:30.3] TG: I'm part of [inaudible 0:23:31.3] community there. We have this paper that's basically doing algorithmic audits. It's going to come out soon, so you can read about it when it comes out. then I'm also working on this idea of like standardizing. We basically want to standardize what kind of information you should put out with your datasets or pre-trained models or whatever.

I've been telling everybody about this thing. I used to be a hardware engineer, and in hardware we have a datasheet that comes with every component. When you're a circuit designer, you would be very intimately familiar with the datasheet and you would call the designers and all these stuff before you design something into your model.

[0:24:14.3] SC: It will pull down a dataset and put it in our train model [inaudible 0:24:18.0].

[0:24:18.7] TG: You have an API that someone releases an API that you have to pay for with, really, you have no information about — You really have zero information about how are you supposed to — Is it supposed to work on this new dataset that you're using? Are there recommended applications? What's going to happen if you use it the wrong way? It's very dangerous, because in hardware at least — I think the reason things were very — It's a mature field, but things were standardized, because the failure mode is so, it's visible to most people, to everyone, right? Like your battery catches fire or something like that. Whereas here, it could be visible to some people. If face recognition doesn't work for you, because you're black or something like that, it will be visible, very visible to you. It won't be visible for other people, because they didn't test it out for something like this.

[0:25:13.8] SC: Because probability is involved, it might not even be uniformly visible in the class that's affected.

[0:25:19.8] TG: Yeah, exactly. A lot of people weren't even doing these tests. The first thing we have to start doing is like doing these audits. Stuff like these that I'm working on. I'm also working on just like I was telling you, models that are fair, or what is a fairness criteria. I've learned a lot about this field. So I'm pretty new to this field and now I've like gotten embedded in the community, which is nice. What I mean is to the fairness community.

Yeah. So kind of working from the purely machine learning perspective, like how can we have account for fairness criteria in our models. That's very broad. That's one of the things I'm working on in addition to just like my regular computer vision kind of work, like domain adaptation, blah-blah-blah. Yeah, that's how I'm taking this to my new role at MSI.

[0:26:15.1] SC: Okay. Interesting. I'm really interested in this conference. I'll need to get some info from you about it and check it out. What are some of the other kind of major research directions in the domain of fairness?

[0:26:28.1] TG: I think that uncovering bias is one. Joy and I just did this paper that's coming out where like we were auditing commercial gender classification APIs that are sold by people that you have to pay for and looking at disparities among certain groups by skin color, by gender. Then [inaudible 0:26:49.1] just came out with this paper.

I'm just talking about computer vision, because actually in computer vision, it's very new. People have been doing this stuff for like — There is this — I don't know. Have you heard of this man is to programmer, as woman as to a homemaker. That's the title I guess, paper —

[0:27:04.4] SC: Yeah, I've seen a few various —

[0:27:05.5] TG: Yeah. I think in computer vision that's one of the reasons that I wanted to get into it. I felt like computer vision people weren't thinking about this.

[0:27:12.7] SC: Whereas NLP, it's been a little bit more established.

[0:27:14.9] TG: [inaudible 0:27:14.2] even Hanna does some NLP and does NLP. Also theory people. I feel like in terms of technical people, theory people were starting to think about. Privacy people, like [inaudible 0:27:28.8] and also from the ethics side that I felt like, "Okay. Now, deep learning people are talking about it too and there are papers that are just like these.

Last year, I felt like people were starting to talk about it, but it wasn't a real concern. Now, in computer vision even, we're starting to see some of this work. One is uncovering bias, like what kind of bias exists. The second one is how do you mitigate it if you uncover it? There's a lot of work.

I guess the work I just talked about, the [inaudible 0:28:00.6] work, they first uncovered the bias and then they tell you some strategies of mitigating that particular bias. Then the third one, which I'm very interested in is — I'm interested in all of them, right? The third one is also just like understanding how these things are being used. If you have — And how to standardize them, how to have transparency? If you have law enforcement that's using inaccurate face recognition algorithms, where are they using it? How are they using it? We have no idea.

Also, just like — There are people who are using your social network data to like — Then selling it to other people or trying to figure out like your credit readings, and like they're startups. Cathy O'Neil talks about it in her book. Have you read this book, *Weapons of Math Destruction?*

[0:28:44.9] SC: Okay. I've heard of it, and it's on my list.

[0:28:48.5] TG: Yeah. that's a very important part of the issue actually, because as AI researchers, a lot of times, me included, like I just want to sit in a corner, read my papers and, honestly, write some code. That's what I love doing most still. Even though I do this whole like social activism stuff, what I enjoy doing is just like reading papers, thinking about ideas, writing code.

We have to understand like the implications of our work are. Like keeping track of — I just signed this extreme vetting letter against this extreme vetting initiative. I don't know if you've heard about it, by the DHS, that was trying to — I didn't even know about it until I was —

[0:29:35.1] SC: The social network data to get visa application.

[0:29:38.4] TG: Yeah. I was asked to sign this letter. I had no idea this was going on, and it's terrifying. This kind of stuff, we have to keep track of and we have to make our voices heard in addition to working on uncovering bias and mitigating bias.

[0:29:50.8] SC: Right. Awesome. I know you've got to run off to talk. So let's wrap it up here, but if you have any final words or thoughts or places that folks should be looking to keep up with all these information or finding you, now is the time.

[0:30:06.5] TG: I recently got on Twitter. I was told it's a good thing. Yeah. Tmnit is my Twitter. I have a website, so I will be releasing data soon for this monstrous work that we just discussed, the PNS paper and to look out for my new paper with Joy as well which is going to be released in a couple of weeks.

[0:30:28.2] SC: All right. We'll link you on Twitter in the show notes as well as to your homepage, and looking forward to following this line of research and the stuff you do at Microsoft.

[0:30:36.0] TG: Thank you.

[0:30:36.6] SC: Thanks, Timnit.

[END OF INTERVIEW]

[0:30:43.0] SC: All right, everyone. That's our show for today. Thanks so much for listening and for your continued feedback and support. For more information on Timnit or any of the topics covered in this episode, head on over to twimlai.com/talk/88. To follow along with the NIPS

Series, visit twimlai.com/nips2017. To enter our TWiML 1 Mil contest, visit twimlai.com/twiml1mil.

Of course, we'd be due lighted to hear from you either via comment on the show notes page or via a tweet to @twimlai or @samcharrington.

Thanks once again to Intel Nervana for their sponsorship of this series. To learn more about the Intel Nervana NNP and the other things Intel's been up to in the AI arena, visit intelnervana.com.

As I mentioned a few weeks back, this will be our final series of shows for the year, so take your time and take it all in and get caught up on any of the old pods you've been saving up. Happy holidays and happy New Year. See you in 2018. Of course, thanks once again for listening, and catch you next time.

[END]