Ben Tankus

CS 519 Final Project

6-1-21

A Review of Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints

non-technical background (3 pts):

Who:

PI: Dr. Vicente Ordonez University of Virginia, Kai-Wei Chang University of Washington

Students: Jieyu Zhao, Tianlu Wang, Mark Yatskar,

Where:

This paper was published in 2017 by the Danish journal: *Association for Computational Linguistics.* It was also originally written in 2017. According to Google Scholar, it has been cited 418 times, which I would guess is pretty high impact. Looking through the citations, it seems like the impact is dirastically increasing in the past few years. When the work was published in September of 2017, there were only a few citations until 2019 when it gained traction. Most of the citations look to be from 2020 and 2021, so the impact is certainly growing. This paper received the *Best Long Paper Award* from UCLA in 2017 along with author Jieyu Zhao winning a *Microsoft PhD Fellowship* in 2020.

They reference a data source many times, but they do not share it's origin or what the data looks like beyond brief descriptions. The paper also discusses their methodology at length, but really only provides very theoretical guidelines and does not provide any code, nor a demo. With a little more digging, I was able to find their github page where the code, data, and paper are stored. It seems that this is fairly easy to reproduce with all this information, but not at my current knowledge level. I also am not able to find any slides, talk videos, nor media reports of this paper.

core (the famous what-why-how-wow template for writing abstract/intro) (8 pts):

What:

The authors are trying to solve the problem of social bias in a machine learning algorithm. Typically, the machine learning algorithm in question was being prejudice against “societal norms” by recommending items such as cooking heavily biased towards women, and alternatively things like “hunting” biased towards men. This is not a new problem in society, but it is a new problem in machines learning.

Why:

The study group chose problem because it is very important both socially, *and* economically. Companies spend fortunes ensuring their advertisements are seen by the right users, and it will be crucial to social justice in the future for machine learning models not to reenforce social stereotypes.

This is a difficult problem. The algorithm they are using is quite complex, and relies on the computer “seeing” the correct gender both based on context, and features. While humans will be fairly accurate, it’s possible even for a human to misgender someone who may be androgenous, so it’s not surprising that computers may have issues with that as well. Computer algorithms typically amplify bias by nature, so it is a difficult problem to solve to reduce computer bias.

Once this problem is finally solved I believe adds will be much more focused on who *actually* uses their product instead of the general stereotype. Companies will understand their users better, and users will feel better represented in the companies they purchase from.

How:

The research group chose to use their own developed adjustment algorithm they called the Reducing Bias Amplification (RBA) technique. The research group used RBA to “debias” the training process via adding constraints in the process. This process was iterative and improved each time through the iteration. The research group also used “Lagrangian relaxation” to validate the results. Looking at the git code for this process I really have no idea how it works. I would have to spend a large amount of time just to understand what they have done here, but it looks like they did their due-diligence.

Wow:

I think this is a great place to start looking for models to improve, but I don’t think this paper has really provided any significant results, and in-fact was a bit misleading in reporting their results.

One of the research group's claims is that they improved the agent assignment of cooking from the original model's 16% to 20%. The research group goes on to say they have improved the algorithm in this instance by 25%. While this may be correct, 20% is still a *significant* amount away from the actual 33% available in the training data (a 106% improvement by this logic). I believe reporting their model improvement as 25% is a bit misleading. I would recommend instead reporting only the 16% to 20% change without highlighting the 25% improvement. (image from the study below)

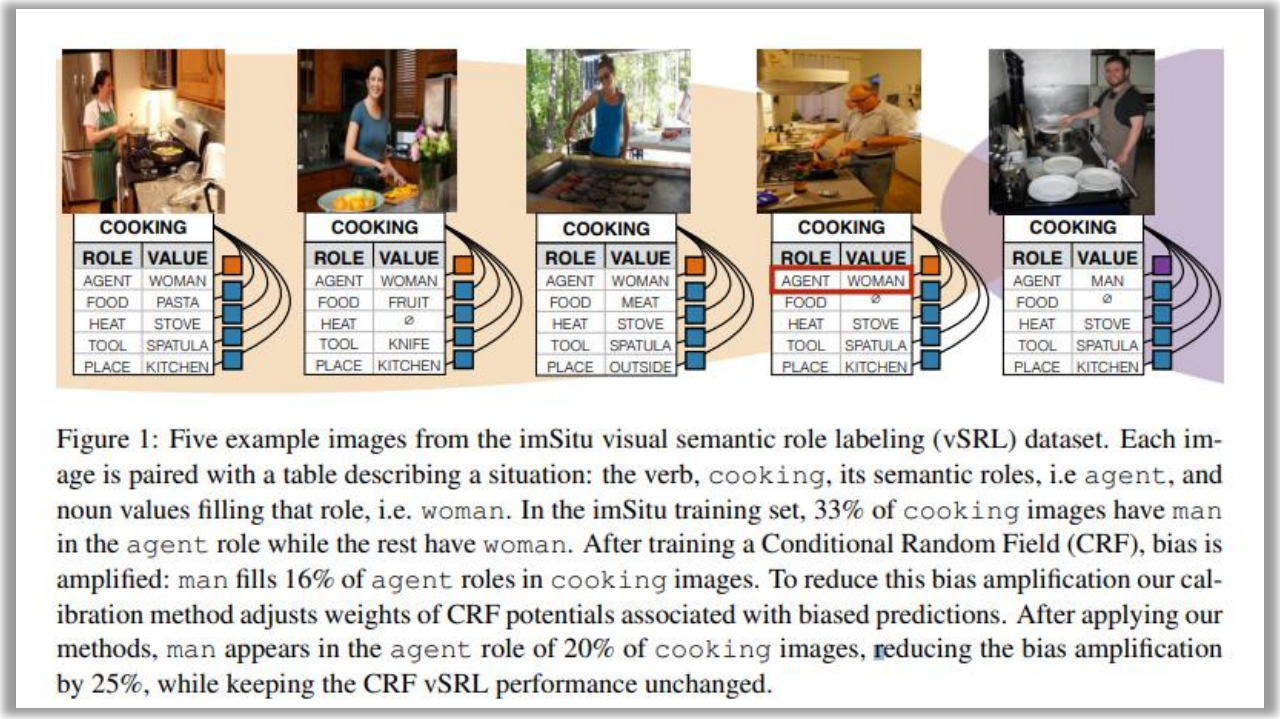

Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images have `man` in the `agent` role while the rest have `woman`. After training a Conditional Random Field (CRF), bias is amplified: `man` fills 16% of `agent` roles in `cooking` images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, `man` appears in the `agent` role of 20% of `cooking` images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

further: (3 pts)

But:

It's difficult to critique other's work without seeing the whole process from start to finish, especially with such a complex problem. If I were to recommend something to improve on this paper, it would be how they report results, as stated above, and I would like to see more due-diligence in proving independence within the data.

On this same topic, the study group claims on page four, "Lagrangian relaxation guarantees the solution is optimal if the algorithm converges and all constraints are satisfied". While that may be understood as common knowledge for the research audience, as this paper has been circulated heavily I would like to see a citation for this *very* strong claim.

Machine learning algorithms can generate dependent data fairly easily by nature as they learn. On page four of the paper the research groups states the data can be “decomposed into small components based on an independence assumption”. This is a crucial assumption in the analysis that has no citation, nor is it discussed further either in the paper, or the github code.

If this study were to be redone I would like to do more work to validate data independence in the study. I would like to at least understand a bit more about how the data was gathered, and where the data is accessed.

More:

This research group brings up a very important point, gender bias. I definitely don’t think it has solved the problem, but the 400+ papers that cite this study may have come closer. If I were to conduct research following this study I would like to take their developed RBA algorithm and apply it to other datasets to see if it improves their models as well.

In addition to trying out the model on new data sets, I would like to iteratively improve this model as it is exposed to more and more data. As the model is exposed to more and more data the researchers can identify gaps in the model and improve it as needed.

All:

My overall feeling of this paper is the research group made a big *first step* in researching the societal bias inherent in machine learning models. Machine learning models are great to model the intelligence of humans, but they only understand at a very shallow level. They do not (currently) understand implications of their actions and are only programed to assign based off of effectively stereotypes. This inherently balloons bias in a social classification setting as it rewards the algorythim for a successful match more than it penalizes for a poor match.

relevance: (1 pt)

Did materials covered in this course help you understand this paper? How relevant is this paper to this course?

The materials in this course did give me a good understanding to know where this paper fit in machine learning. I found it interesting and likely would not have been able to understand it before taking this class. I believe this paper is fairly relevant to this course, but really that could be said about any machine learning paper.

What extra materials did you study in order to understand this paper?

I would like to have more understanding in the area of machine learning “general knowledge”. The lack of Lagrangian Relaxation “guarantee” citation leads me to believe this is common knowledge in the machine learning industry and I have not heard about it *at all* before I read this article. It’s difficult to get this knowledge without constant exposure over time though.

Things I looked up:

- Lagrangian relaxation, approximates a difficult problem. Simply put, they made it less complicated so they could solve it.
- Corpus – previous academic works