FYP/FYT Essay on Professional, Ethical, Legal, Security, Social Issues and Responsibilities

Name	Jihoon Chung	Student ID	20311792
Project Code	JA1	ITSC a/c	
Project Title	Deep learning and sentiment analysis		
Essay Title	Commercial License in Machine Learning Dataset		
Date	2018.10.26		
Word Count	3915	Total Pages	7

FYP ethics essay
Commercial License in Machine Learning Dataset

Jihoon CHUNG, 20311792

October 2018

1 Introduction

Machine learning is a field of artificial intelligence that allows machine, or algorithm, to learn from the statistics of data, to get expected output. This allows the engineer the to make an algorithm without explicit programming, but with better result. Machine learning has been studied for quite a long time. But for long time, since perceptron in 1957, machine learning has never given good enough result and accuracy to used widely to solve real world problems, and they were only considered as a field of research, or used as a small portion in existing artificial-intelligence algorithm. But in 2012, Alex Krizhevsky [7], built AlexNet, with ground breaking record in ImageNet Large Scale Visual Recognition Challenge with 11 percent lower than second runner up. Now with technical breakthrough in hardware devices and with development of various techniques within machine learning, machine learning now gives promising result that can be used directly into real world problems. Machine learning is currently incorporated to solve real world problems, and also being used in lots of commercial areas. As this is new technique that was never used in product level before, new ethical problems that has not been considered has slowly been rising.

In this essay I will talk about one of the ethical dilemmas of machine learning, commercial license in machine learning dataset. I will briefly explain what dataset means to machine learning, how commercial license is used in computer engineering in section 2 and 3. Then I will state the ethical dilemma in the section 4. I will also go over some useful information related to the topic and explain why this ethical dilemma is hard to solve, and remain in gray area on section 5. Finally, I will wrap up my essay by giving two strong arguments of both sides in section 6, along with why this dilemma is important issue, and how we should see it.

2 Dataset

Dataset is a key to machine learning. Machine looks at the statistical distribution of the data, or in other words look at the pattern of them and finds a connection between input and the output. For example, in field of image recognition, dataset will be pairs of image and its label, and the machine will learn from these data and decide if the given image is a "cat" or a "dog". This is just a simple example, but dataset can vary from image, text, 3D model, or an audio. As the machine uses these datasets directly to train oneself, it is important for a researcher to provide a clean dataset in large quantity. In here, clean dataset means there are less noise, such as mislabeled data, where cat image is labeled as an elephant. Quantity is also an important factor, although the size of a dataset varies as you can see in the table 1, most of the times, it is better to have large number of data, than to have small size.

Dataset	Size	type	publicity
CelebFaces [13]	202K images	Face	Public
Google Face [11]	8M images	Face	Private
ImageNet	14M images	Images, text	Public
Zero Resource Speech Challenge 2015	120 hour	Speech	Public

Table 1: Number of data in some of the popular datasets

Moreover, some dataset requires special labels, that are hard to get by simple crawling[1]. These annotations, such as location of the face, or context of the image, are automatically added by existing algorithm, or through manual

[1] act of automatically browsing internet websites to get a particular data using a computer program

labeling. For example, ImageNet originally had classes of the image only, whether the image is cat or a dog, but now it also has context of the image, such as *The cat is drinking milk.*

As the size of the dataset has been growing, along with complexity of the dataset labels, creating dataset for the machine learning algorithm has become tough task. Some big corporates like Google, Facebook, or NVidia can make fair dataset from the services they own. Google, for example has recaptcha platform to obtain dataset for their self-driving car. These big corporates have enough budget to make a dataset from scratch. However, most start-ups or small companies, use public dataset. Public datasets are free to use, have been widely used, which proves that they give good results. Moreover, they can give good evaluation as other people also used the same dataset, it can be easily compared if new machine learning model is better or worse.

3 License

License is one of the important matter in area of computer science. License allows people to use open source code in to their project freely, or with constraints. And in computer science research, commercial license and non-commercial license is also distinguished. Commercial license allows people or corporate to include the following code in to their commercial project freely, where non-commercial license does not. Similar license can be seen in computer tools. Microsoft Office has two type of license; personal and commercial. If one wants to use the Office in corporate to gain profit, they have to buy program with commercial license which are usually more expensive. License is more free in research area. section 107 of Title 17 of United States Code states that " Not withstanding the provisions of sections 106 and 106A, the fair use of a copyrighted work, including such use by reproduction in copies or phonorecords or by any other means specified by that section, for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research, is not an infringement of copyright. " [2] That is, for research, copyright of a material does not have to be taken account of, as it is part of 'fair use'. Similar law exists in different countries as well. United Kingdom has Fair dealing in United Kingdom law that gives exception to copyright infringement for purpose of research, study or other non-commercial use.

4 Commercial license in machine learning dataset

There are two type of dataset for machine learning, one is crawled dataset, and the other is generated. Crawled datasets are datasets that was attained by crawling through various internet sites. These datasets are usually posted publicly with research-only license. But this license is done by the laboratory that has collected these data, and real license, or copyright, of the original data is still remained the same. But with fair dealing/fair use law, use of any material in laboratory of any material is allowed. Most researchers use public dataset for their research as they can be used without any copyright infringement. Other type of dataset is generated, which the laboratory has made the dataset by themselves by taking photos, recording sounds, or collecting images from their websites that people has uploaded. These datasets usually include commercial license as the research organization are the owner of the image. These datasets are sometimes sold to other labs, for example, Multi-PIE [4].

Corporates tend to buy these datasets as they are the only dataset that offers commercial license, which can be used to make commercial applications out of it. However, it is tempting for companies to use public dataset for their commercial application, as they are cheaper to use, have more data, and are proven to be good dataset. In section 5, I will explain why use of public dataset for commercial use is in a grey area, and why it is an ethical dilemma to say it is wrong or good.

5 Gray Area

This license problem of machine learning is currently a gray area. As it is relatively new problem, there has not been any legal court actions related to this problem. So it is important to see this matter in different aspects. It is not possible to see this matter alone, and make a decision, but to compare with similar problems, and make logical conclusion out of it. In section 5, I will list some closely related problems and some other important things to know to understand this topic, so that the readers can have enough knowledge to make logical decision.

[2]Note that in this essay, commercial license and copyright will be considered as a similar regulation. Although they are quite different, commercial license of dataset is solely comes from copyright.

5.1 How is data stored in machine learning

To understand license problem of dataset, it is important to know how dataset is used in machine learning. Although machine learning is a big concept with different types that varies a lot, I would only take account with multi-layer perceptron, also known as deep learning, in this essay as they are widely researched nowadays and are giving results that can be used to tackle real world problems. In deep learning, model is a type of complex mathematical formula that can change its coefficient. Simply saying we can think $f(x) = 2x + 1$ to be change to $f(x) = 3x + 4$. Now, machine learning has lot more complex formula, and it does calculus computation to change its formula that the output can be something we want. As you can see, in machine learning, dataset is not stored inside but only the coefficients. These coefficients are called weights. It is called weight because machine learning is just combinations of multiplication on input, and it calculates, and decides important part of the input and unimportant part. Weight is what is stored in the result of training. If we see brain as a machine learning model, then what is inside the brain, work flow and algorithm, is called weight in machine learning. Every model has different size of weight, which can be seen in table 2.

AlexNet	62M
ImageNet	60M
VGG-16	138M

Table 2: Number of parameters for some popular model use for image recognition

These weight starts as a random number, but later with enough training, it becomes some specific values, that can perform wanted behavior. Now, weight of the model does not contain any of the original dataset, but rather it stores "method" of generating output, given the input. It is arguable if the weight of the model follows same license as the dataset, or it is independent of the dataset. It is agreed that the weight of the model is derived work from dataset, but being derived work does not mean much, for example, every painting a painter draws can be derived work of every image he has observed in his life, and we do not say his paint shares same license with images he saw. Deeper comparison with machine learning and human brain will be done in section 5.5.

These weight can be shared among people, because with same weight, people can get same result without exhaustive training procedure. These weights are called pretrained model, and they are also used as an initial value for training the model for different task, which were proven to shorten training time dynamically. Using pretrained model is common procedure in machine learning, and essential in computer vision. However, most pretrained model are trained in public dataset, where owner of each data did not agree them to be used for commercial purpose. However, pretrained models are often overwritten with new coefficients and deleted completely. This is also related problem if use of pretrained model is violation of copyright.

5.2 Comparison with other use of commercial license

To see how commercial license is used in IT corporates, I will give an example of two main use; computer tools, and source code. Although none of them can be seen equivalent to how dataset is used in machine learning, but it can give good insights on how license has to be handled in computer science area.

Tools, such as Microsoft Office, tends to sell different version of the program to corporates. They have almost same core functionality as personal version, with only different is that they come with commercial license, which means that corporates are allowed to make profit out of it. Unity, game engine, for example, can be used freely if the profit is less than 100,000 USD per year, and developer has to buy Pro version if the profit exceeds the given value. This is logical for tool provider to charge more if the tool is used to make financial earning, rather than personal use. Key difference with these programs and a dataset is that program is a tool, like a paint brush. And this tool is something continuously being used to manufacture a product that will be sold to make profit. It is hard to see dataset to be a tool, and make direct comparison with Microsoft Office or Unity engine. In a machine learning, Tensorflow or Pytorch would be equivalent of these programs. [3]. But dataset is less of a paint brush, but more of drawing book that painter used to learn how to paint.

License of a source code is also actively checked when developer is building a program. Just like every human creation, source code is also protected by copyright. However, these codes are widely shared among community, and

[3]Tensorflow and Pytorch are libraries, that helps programmers to make machine learning model efficiently. Tensorflow is under Apache License 2.0, that allows commercial use. Pytorch is still ready for production, but, commercial use is allowed

they all have different licenses that has different limitations. Companies strictly follow this license, by stating what open source project it has referred during, and only using open source codes that allows commercial use. Dataset can be seen similar to how open source code is used in projects. Every project can be seen as a derived work of the open source code, as the source code is modified enough when it is used in the project. Weight of a deep learning model is also derived work of dataset, and they are modified enough. However, unlike source code, dataset cannot be retrieved back from the weight of a model, and dataset is used as a training material, rather than functionality of a program.

5.3 Possibility

One of the important factor in this ethical dilemma is that it is impossible to make satisfying dataset by getting all the commercial license and copyright from the public dataset.

Biggest problem of all is that the size of the dataset is too big for an organization get a commercial license of. Stated on the table 1, datasets are usually large, with image dataset containing few hundred thousand images. It is practically impossible to attain copyright for every image of these datasets, of both photographer and the model. Moreover, most of these datasets are annotated in some way, e.g. by giving a single sentence of context of the image. These annotations are usually the harder problem, as they need human labor. Some dataset does automatic annotation by running some well proven machine learning algorithm to it. Problem is that this well proven algorithm is also trained with public dataset. Making it impossible for a corporate to make 100 percent copyright-free dataset.

There are some work around way by creating dataset by themselves. These types of dataset are popular as they are known to be cleaner than web crawled dataset, and are usually annotated manually by human workers. However, size of these datasets are relatively small, being only few thousands of photos. Moreover, these datasets do not represent real world very well, as they are restricted. For example, Basel Face Model [8] was made in Basel, Switzerland, so this face model is strongly biased in ethnic group and in age. Multi-PIE dataset [4] is one of popular face dataset, as it has 750,000 images, with commercial license sold separately, and is well annotated. But the photos were taken in laboratory environment, so the deep learning trained with dataset, tends to give bad result in real life images. In figure 1, it can be seen that data from LFW dataset [5] can be used much better for real life images, than Multi-PIE, as the data does not represent images of real world.

Figure 1: example of data from LFW (left) and Multi-PIE (right)

5.4 Generative Model

One type of machine learning is generative model, and this is the type that might cause confusion when they are dealing with copyright and license. These type of model generates certain output that could be text, image, or audio. This includes creating an image from a scratch [3], or creating an image given a certain input image ([2] [6]). This type of model is lot more controversial, as they are generating result with same domain as the copyright protected data. For example, one can make a dataset full of Disney Characters, and make a model that creates new artificial Disney character that does not exist. One can also include every cartoon character in the world in the dataset instead, and generate new cartoon character. It is obvious that it will be infringement of copyright if the result looks exactly like Mickey Mouse. But if the result is fictional character that does not exist before, selling such image and making commercial value out of it would be a controversial topic.

Style transfer [2] is another type of generative model. It can be seen from figure 2 that this model transfers artistic style of a target image, and redraws source image (top left) to a new image. This type of an art will not be considered as infringement of copyright for human painters, but will just be seen as mimicking style of another

painter. But for machine painting, the fact that source image itself is feed it into the machine makes it wonder if it will be considered copyright infringement. There are already lot of commercial application in the market that uses this technique ([10] [1]).

Figure 2: Example of Style transfer

Martin Kretschmer and Thomas Margoni [9] gave good example on why copyright exception could cause problem in generative model. Below is part of their article

> "If a temporary reproduction of a Harry Potter book is made in order to produce unauthorised copies, this is clearly a copyright infringement. But if the temporary copy is made within the scope of a machine learning application to study natural language processing, only the informational value of Rowling's text is extracted. These text extracts are annotated with labels, such as named entities and sentiment tags, which are then processed, for example, to create or improve automatic translation tools.
> But the novel is not replicated as a novel, so such an extraction method should not constitute a copyright infringement.
> "

We can farther think that these extracted values can be reused to generate Harry Potter novel, but with different sets of proper noun. If such text extraction is not considered as copyright infringement, generated novel from this extraction will not be as well.

5.5 Comparison with human brain and image compression

Here I will give 2 very similar case of generative model, that is human brain and image compression. Purpose of this comparison is that even with similar mechanism, some behavior is allowed for human painter, but not for machine learning, and some are not allowed for image compression but is allowed for machine. Pointing out these similarities and differences would help to understand the ethical dilemma better.

Let's think of a human painter. A painter can train himself, by copying various work of art made by other people. During his training, he does not memorize entire painting made by other people, but he only remembers how to draw a good painting. And later on, he can draw some original work, without copying other work. Even if he has learning how to draw from copyright materials, which he did not attain any commercial license of, act of selling his original work is not considered as an intellectual property infringement, as long as his drawing does not resemble any copyright protected image. Machine learning is almost the same process as a human painter. Some argue that machines do not have creativity that human has, but there is some model that has randomness [12] included, that acts similar to creativity that human has. But, human painter has far more activities than practicing how to draw, which can also influence his style of drawing, while paint drawing of machine would be only influenced by the dataset it has received, which means that original work of the machine is not totally original, but they are all resulted from certain copyright protected material. But still, core mechanism of these two are not very different to say that they are not related.

Style transfer that I have mention above in section 5.4 is one of examples of generative model. One can argue that generative model does not output exact same image compared to input, but only something that is in same domain [4], so it is not infringement of copyright. Will it be okay as long as image is not exactly the same? We can think of lossy compression, making size of the image smaller with low resolution. Due to the compression, image has been modified with bit more blur and noisy pixel. Can this be seen as different image of from the original work? Lossy compression and style transfer shares the similarity that they are both algorithm that works on input image to slightly modify it to another image. We can farther think of style transfer that does minimal distortion, where output is almost similar to the input. It is hard to say how different they are, as in technical sense, lossy compression is not much different from style transfer.

6 Arguments

Before I make a conclusion, I would like to give arguments of both sides to point out advantages and disadvantages of giving exception of copyright and commercial license in machine learning dataset.

6.1 Arguments for exception of copyright for machine learning

Currently, most of machine learning research is done in corporate area, rather than public research facilities in university. State-of-the-art result in face recognition is owned by Sensetime, best machine translation and text-to-speech research is done by Google, and Nvidia also has lot of interesting results related to image and videos. As machine learning has been giving ground-breaking result in lots of areas, corporates are the leading research facilities in this field. However, strict regulation of copyright and commercial license is basically illegalizing machine learning to be used to solve real world problems, as it is impossible for a corporate to generate a dataset from strictly following copyright and commercial license. This is different story from commercial license of tools or open source problem, as these problems can be solved by spending more budget, but it is impossible, no matter how big the budget is, to get copyright of dataset. This is limiting technology that could help humanity which is greater good than license. Moreover, copyright is built to protect progressive idea from being stolen, but regulation of copyright and commercial license is limiting progressive idea.

6.2 Arguments for regulation of copyright in machine learning

These data in the dataset is still a work made by an individual, and even how big the dataset is, still each data are protected by copyright. It is ethically wrong for any corporate to make use of this data without notifying the rightful owner, to use it for commercial purpose. Moreover, since machine learning is fast spreading in various fields, giving an exception of copyright to machine learning would cause tremendous result of degrading idea of copyright. Also, with generative model creating various works from art to novel, it might end up in people making less of the original work.

7 Conclusion

I believe that information I have provided above is enough for the reader to understand this matter and why it is important. It is without a doubt, one of ethical problems of machine learning that is often ignored by the corporates. I was not able to take a stance in this dilemma as it was not something I can solely make a decision of, but I was able to make enough studies in this topic to know how important this problem is, and was able to have deeper insight on commercial license and machine learning. I believe key importance of this ethical dilemma is **awareness**. This is something that no one can make a decision, and will stay in gray area, ethically and legally for a long time. So I strongly believe having awareness in this ethical dilemma as a future-researcher in this field as also an individual might have my data being used in machine learning could prevent from bigger problem. As a researcher, I can always ask myself, if certain part of the dataset can be identified easily, that could result in violation of copyright, I also need to state which dataset I have used, if needed, and explain how I make use of the dataset in my research. As an individual, I have to keep my eyes on, if my copyright is violated by corporates, and I have to make sure these corporates do not make use of my data more than they should.

[4]Domain means same type of data, in machine learning. For example, all the images of chair belong in same domain.

References

[1] deepart. *deepart.io*. 2015. URL: https://deepart.io.

[2] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. "Image Style Transfer Using Convolutional Neural Networks". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 2016, pp. 2414–2423. DOI: 10.1109/CVPR.2016.265. URL: https://doi.org/10.1109/CVPR.2016.265.

[3] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. eprint: arXiv:1406.2661.

[4] Ralph Gross et al. "Multi-PIE". In: *Image Vision Comput.* 28.5 (May 2010), pp. 807–813. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2009.08.002. URL: http://dx.doi.org/10.1016/j.imavis.2009.08.002.

[5] Gary B. Huang et al. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Tech. rep. 07-49. University of Massachusetts, Amherst, Oct. 2007.

[6] Phillip Isola et al. "Image-to-Image Translation with Conditional Adversarial Networks". In: *CoRR* abs/1611.07004 (2016). arXiv: 1611.07004. URL: http://arxiv.org/abs/1611.07004.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[8] Marcel Lüthi et al. "Gaussian Process Morphable Models". In: *CoRR* abs/1603.07254 (2016). arXiv: 1603.07254. URL: http://arxiv.org/abs/1603.07254.

[9] Thomas Margoni Martin Kretschmer. "Data mining: why the EU's proposed copyright measures get it wrong". In: *The Conversation* (May 2018). URL: https://theconversation.com/data-mining-why-the-eus-proposed-copyright-measures-get-it-wrong-96743.

[10] prisma-ai. *prisma-ai*. 2016. URL: https://prisma-ai.com.

[11] Florian Schroff, Dmitry Kalenichenko, and James Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *CoRR* abs/1503.03832 (2015). arXiv: 1503.03832. URL: http://arxiv.org/abs/1503.03832.

[12] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[13] Yi Sun, Xiaogang Wang, and Xiaoou Tang. "Deep Learning Face Representation by Joint Identification-Verification". In: *CoRR* abs/1406.4773 (2014). arXiv: 1406.4773. URL: http://arxiv.org/abs/1406.4773.