**New technologies report**

Lan Xu

Lecturer Yuan Ping Jin

Course Code: CEBD 1151

16 November 2019

Since I learned big data, every day seems to be brand new!

Jack Levis (Senior director of process management at UPS) said, “Beyond knowledge is wisdom, and beyond that is clairvoyance”. The word “data” is “known” in Latin and can also be understood as “fact”. When the data is too big, most of us couldn’t understand the fact, that is the value of learning “BIG DATA”. It is a tool or technology helps us to gain insights and fact or even clairvoyance.

The field of big data is actually very broad because it involves different disciplines, technologies and industries. Personally, learning big data is to broaden my individual’s life path and create more possibilities for myself. The actual situation is, no matter which data position, there are some basic knowledge that must be learned. Because of this, in addition to attend classes in the weekends, I also learned a lot of basic knowledge in private.

Fortunately, thanks to the rapid development of the Internet, there are plenty of learning resources available. Online learning platforms, technical forums, academic papers and many local meetup events have all provided me channels to learn with convenience.

**On-line presentation**

Udemy(<https://www.udemy.com/>) is one of my most frequently used platforms to learn basic knowledge regarding Big data. It is the world's largest online learning platform.

**Pros/Cons/Learning Experience**

The online courses or presentations on Udemy are very rich. Most of the instructors teach clearly. The related course materials can be download (such as slides, datasets, programming scripts etc.). From all the courses I have learned so far, almost all the instructors arrange certain amount of exercises during the whole lecture and with solutions provided after. Before selecting the course, I can check the ratings and can also review the comments from other learners. In addition, one topic you may find courses only for beginners or for high level learners, which allows you to choose the right course according to your own level and requirements. However, Udemy has its own limitations. Most courses are not the latest technology. Thus, I normally use it for basic knowledge learning. And because it is not real time learning, so once you have questions and difficulties, it is hard to be solved. I left my questions in Q&A few times, but unfortunately there is no reply.

I have learned Python3, MySQL, Linux basics, Hadoop ecosystem (25 tools) on Udemy so far. currently, I am learning Power BI which is a straightforward visualization tool. I am going to learn more about Cloud services, like AWS cloud or Microsoft Azure one of it for my next goal.

**Today’s topic**

The topic I pick up today is Streaming Data. I chose this topic because, to a certain extent, it has witnessed the history of big data development, and it gives me chance to make a summary after learning the Hadoop ecosystem.

The large-scale data processing systems we discussed today are all derived from MapReduce in 2003 by Google engineers. The MapReduce system is designed to be innovative although it has been improved and it is no longer a popular technology. However, it provides a simple and straightforward API for building a complex, yet robust, and underlying distributed data pipeline that is sufficient to run this distributed data pipeline on a cheap, commercial server cluster. This concept laid the foundation for the subsequent development of streaming data processing.

Hadoop came out in 2005, it added a MapReduce compute layer based on HDFS. And it eventually become part of the Apache Hadoop project. After that, the Apache Hadoop project was booming through open source. Due to application of resource management systems such as Yarn and Mesos, Hadoop has become a Big Data ecosystem in the true sense. Internally, it gradually integrates a large number of data processing tools, such as Hive, Pig, HBase, etc. Externally, it can be connected to various data storage systems, such as MySQL, Cassandra and MongoDB, Other than that lots of external Query Engines (Drill, Hue, Phoenix, Presto, etc.) can also be built on top of the Hadoop platform. This is why Hadoop can be popular for a decade.

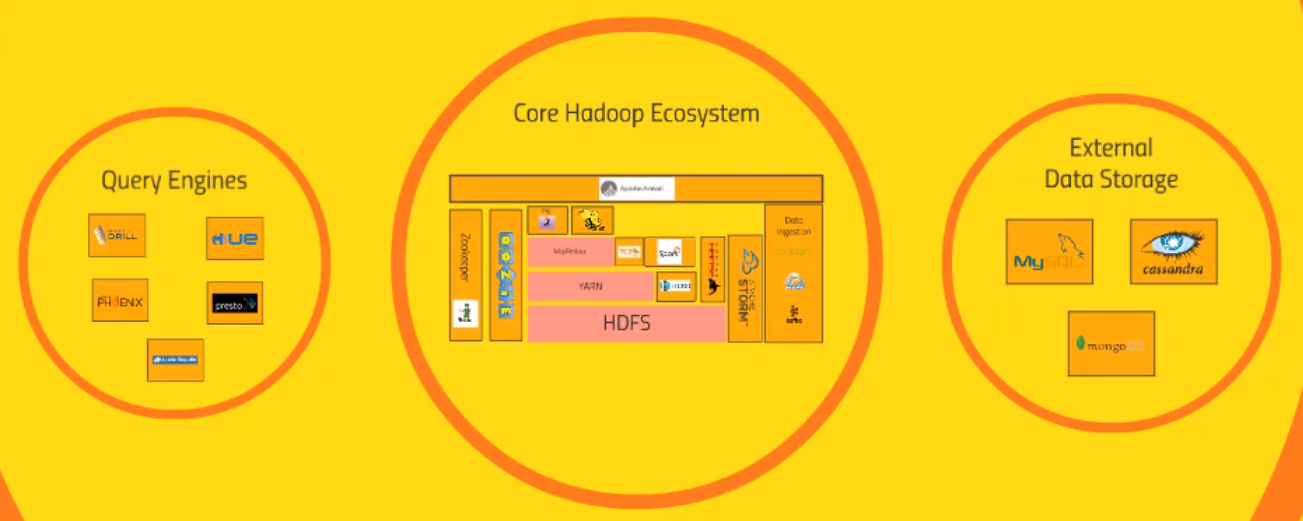
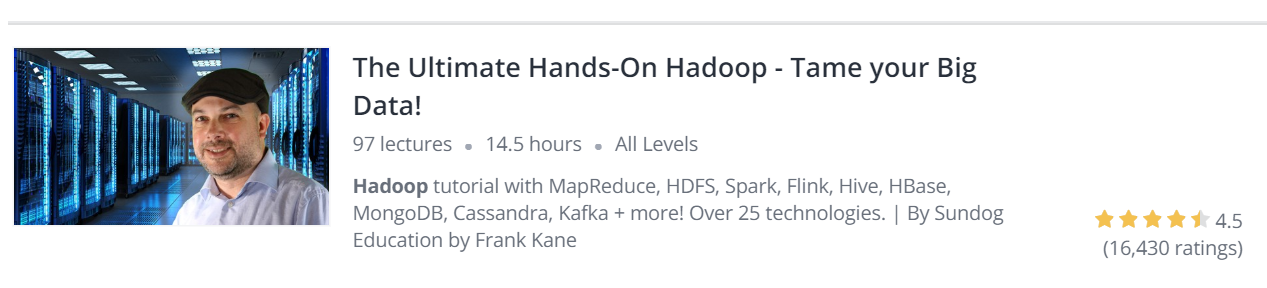


Image source: PPT from The Ultimate Hands-on Hadoop-Tame your Big Data!



As time goes, the growing demands for faster analytics and customer insights bring tremendous interest in streaming data technologies. Thus, represented by Apache Spark, Storm and Flink streaming data processing tools become a new favorite.

Today, I am taking Spark Streaming as an example. Spark uses memory and can use disk for processing, whereas MapReduce is strictly disk-based. The primary difference between MapReduce and Spark is that MapReduce uses persistent storage and Spark uses Resilient Distributed Datasets (RDDs), which means processing of RDD’s can happen in parallel on different worker nodes, thus make it 100 times faster than MapReduce.

**Spark Streaming has its own ecosystem: Spark Streaming can consume static and streaming data from various sources, process data using Spark SQL and DataFrames, apply machine learning techniques from MLlib, and finally push out results to external data storage systems.**

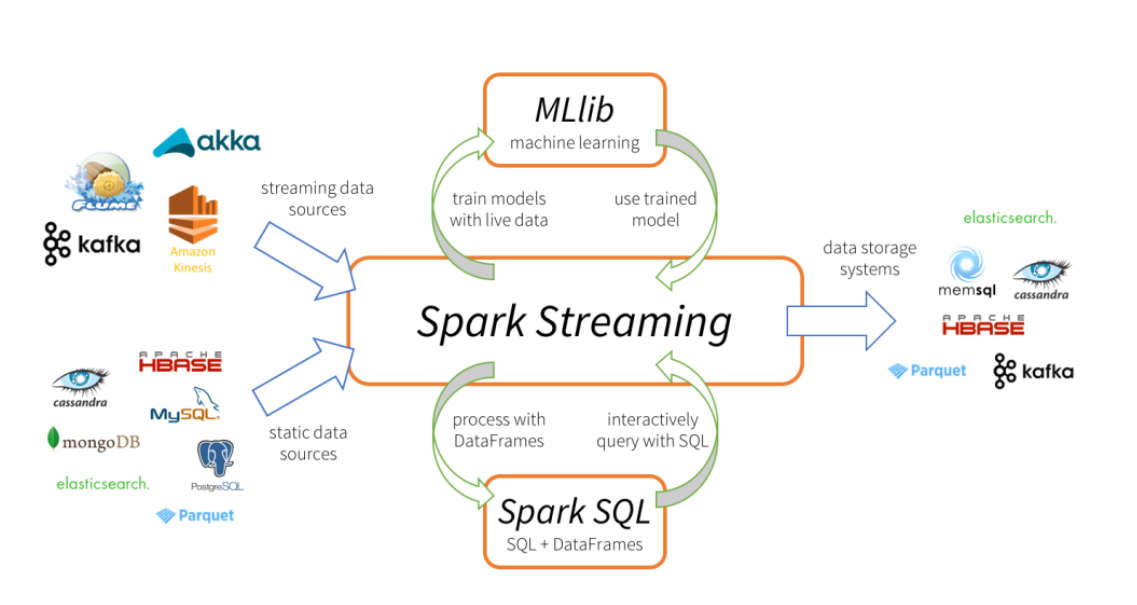


Image Source: datanami

**Apply to real world – Common Use Case for Streaming Data Analytics**

Among the growing users of Spark Streaming, Uber, Netflix and Pinterest are well-known companies on their list. And the major use case for streaming data is clickstream analytics. Clickstream analytics allows companies to track their audience’s activity on their webpages. Companies will often track what pages visitors have been frequenting and the sequence of events that lead to viewers taking a major action—like making a purchase—on their site. Similarly, companies can track what links are clicked the most often and where website visitors tend to spend most of their time on any given page.

In today’s report, I am giving a very simple use case of one of my exercise during the learning journey. It is to design a system to report web sessions per day for internal analytic purposes. I won’t show programming details here. However, I want to use below Data flow chart to show easily how we are now practically use spark streaming in our daily life.

Web Servers

Flume

Spark Streaming

Hive

HDFS

Oozie

Kafka

Step 1: Use Flume to collect and move large amounts of log data from Web Servers.

Step 2: Use Kafka to do buffer cache of log messages. Kafka is a distributed publish-subscribe messaging system.

Step 3: Use Spark Streaming to do real-time calculation.

Step 4: Write the result out to HDFS & Hive.

Step 5: Set up Oozie to run once SQL queries daily

Finally, I want to quote this instructor’s sentence: The perfect big data scenario is exactly as the designers intended - Starting from the actual business needs, don’t only follow hot tools, always choose the most appropriate one.

**Article reading**

In addition to the entire learning time, I have also small gaps in my day to read some articles associated with Big data filed. There are a couple of websites/blogs I visit regularly, such as Towards Data Science (<https://towardsdatascience.com/>), Datafloq (<https://datafloq.com/work/>), and one of my favorite websites is MEDIUM (<https://medium.com/>).

Reading articles on Medium, this habit began with a random reading behavior. After that, I receive often recommended articles by email from MEDIUM.

**Pros/Cons/Learning Experience**

The articles on MEDIUM are very readable, most of them are not purely technical papers, so it is very suitable for my current learning phase. In addition, the articles involve a very broad of fields and industries, from the use case of business management, consulting companies' research reports, and also cutting-edge explorations in the field of technology, and there are even some articles written in a way of narrative novels by which increases greatly readability and fun.

There is a feature on this site that I like, each article indicates the time it takes to read, so that before you click, you can know if you have enough time to complete it. However, this website is not free, it allows a free-trial period but you have to pay $5 per month after the free-trial.

**Today’s topic**

Recently, I read one article called “**Why big data and algorithms won’t improve business strategy**”. It attracted me by its title. Because the hype for big data strategy and machine learning built to a crescendo these years. Most of the articles I read are emphasizing the importance of big data strategy. Because of this, I read this article with great curiosity.

The article compares business to a chess computer game without a chessboard. It explains very vividly and clearly that no matter how advanced the tactical analysis is, it won’t help the player to win without knowing where he is.

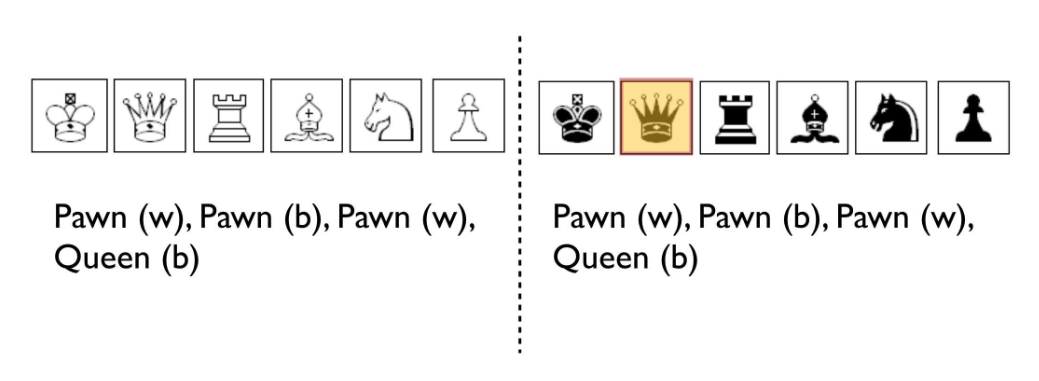


Image source: <https://medium.com/hackernoon/why-big-data-and-algorithms-wont-improve-business-strategy-54e4ebe2398>

Above Graphical description shows both players didn’t see the chessboard, all they see is the control panel and the sequence of presses. It's like a real business world where a store saw a discount on the store next door, so he started making coupons. Then the store next door also issued coupons.

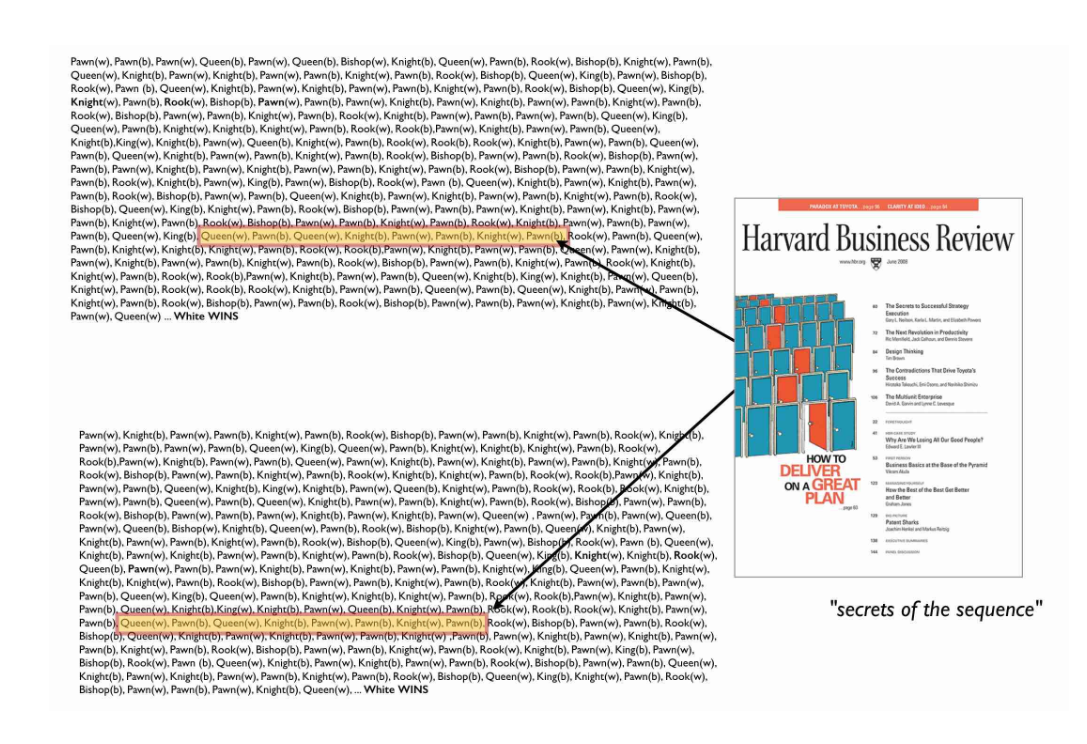


Image source: <https://medium.com/hackernoon/why-big-data-and-algorithms-wont-improve-business-strategy-54e4ebe2398>

From the scenario of the illusion above, as you aggregate more and more data from different games in ever larger “big data” systems and AI is applied then the more patterns will be proclaimed and the more secrets of success announced in popular publications. Back in the real world of business, many business management books or articles are now strongly emphasizing the importance and operability of a company's data strategy, even the data strategy is considered the company's business strategy or the core of the strategy.

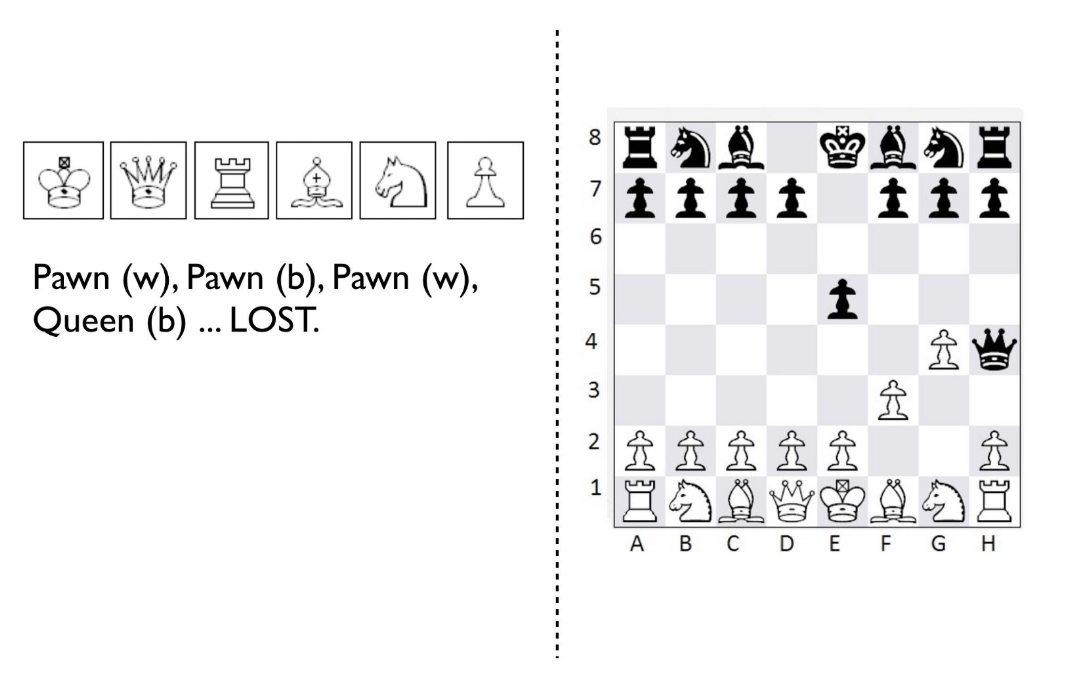


Image source: <https://medium.com/hackernoon/why-big-data-and-algorithms-wont-improve-business-strategy-54e4ebe2398>

Until one day, a chess player came and he played chess with you on the board (As shown above). You lose one after another, but you don't understand why, because you have been using advanced analysis, AI, machine learning, etc. to help you with each step. The problem for the player controlling White is they have no real understanding of the context of the game they are playing. White’s control panel is just a shadow of the landscape and the sequence of presses lacks any positional information or consistent understanding of movement on the board.

Read the text here, the author's meaning is already very clear. There are a few and rare exceptions but in general, the key is still business strategy which is first to understand the landscape of business. Then we can use large scale data analytics, AI and algorithms to discover new insights such as improving the supply chain or understanding user behavior or marketing or loyalty programs or operational performance, etc.

**Guidance to reality**

The real-life examples are proofs of the above points. The success of Apple Inc. is largely due to Steve Jobs’ genius vision. He has created an iPhone-centric production ecosystem and with the development of new products, consumers are more rely on Apple products and services. Without his vision and business strategy, Apple Inc. is impossible to stand on today’s top list.

The glory of Nokia's past is still faintly visible. However, there are no more Nokia phones in the market today. It was due to a wrong business strategy that led to the decline. This failure cannot be attributed to the lack of a strong data strategy.

The inspiration of this article is that: practically, data strategy should not exist alone, it should be integrated into business strategy and become a powerful tool to empower the business strategy. It is not a savior, nor should we have the idea that as long as a data lake is established in the company, or a large number of data scientists are recruited, it will become a new generation of technology companies. Companies should have a mechanism of visualizing their environment and continuously learning from it. Situational awareness should be conducted and connected with business strategy. Only with the correct business strategic settings, the optimization and insights brought about by data analysis and machine learning will have real meaning.

In conclusion, on the road to becoming a data-driven technology company, don't get lost in the ocean of technology. When a genius business strategy is presented, don't forget to pick up the powerful weapon of data analysis to make the strategic vision a reality.

**Meetup Events**

Montreal is a great place to learn big data. There are many academic seminars, public lectures, and technical exchanges held regularly.

I am very grateful to lecturer Jin for opening another door for me to learn. Because of the requirements of this technical report, it drives me to go out and communicate with people and have the opportunity to get to know the latest ideas.

Eventbrite Montreal (<https://www.eventbrite.ca/d/canada--montreal/events/>) is a Montreal-based [event management](https://en.wikipedia.org/wiki/Event_management) and ticketing website. It provides timely event information. There are frequent speeches about AI and some of them are free. The registration of the event is very simple, and the introduction of the event is also detailly described.

**My attended Meetup**

Topic: Unsupervised Video Object Segmentation for Deep Reinforcement Learning

Lecturer: Prof. Pascal Poupart of University of Waterloo

Time: Nov 14, 2019

Place: at the Borealis AI Montreal lab

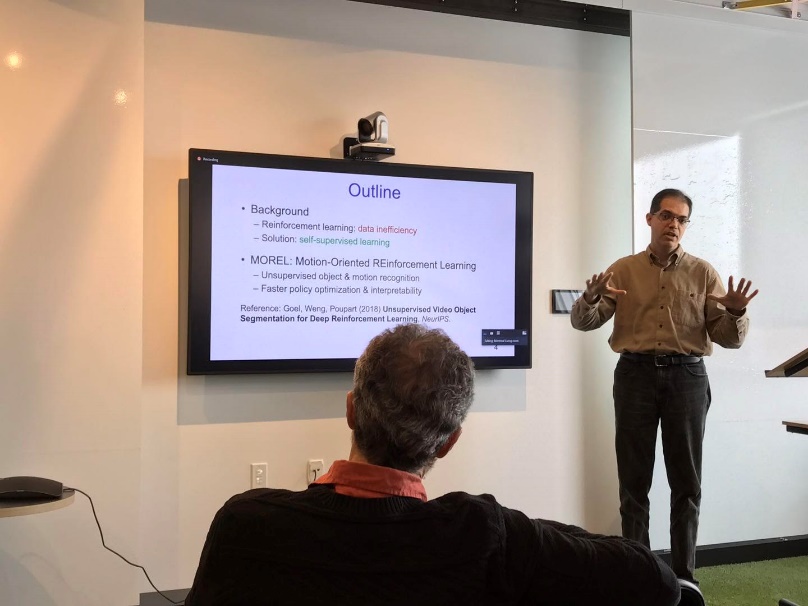


Image showing: prof. Pascal is presenting

As I have roughly learned Reinforcement learning, it has a classic algorithm Q-learning. And these years, with the prevalence of neural networks, Deep Reinforcement Learning came up. It is based on traditional reinforcement learning but merged with neural network algorithm. It uses Image as input and feed the data to Deep N.N. algorithm then take actions with sparse rewards.

Image

Feature

Extraction

Policy Optimization

Action

Sparse Reward

Deep N.N.

Extraction

Sketch of Deep Reinforcement Learning model

However, as above flow chart showing, the image data is decomposed into pixel data which fed into Deep N.N model to extract features. Since there are two sorts of objects in the picture, static and dynamic, thus the features extraction takes long calculating time with lots of uncorrelated features.

In order to solve the issue of long calculating time, prof. Pascal presented a new technique of deep reinforcement learning that automatically detects moving objects and uses the relevant information for action selection. His idea was inspired by a paper call SfM-Net: Learning of Structure and Motion from Video published in 2017. In paper, it provided a model to extract moving objects from the images which give Pascal team’s hint to add this model to deep reinforcement learning, thus will greatly decrease the feature extraction from the raw image.

The model they created called: Motion-Oriented RL, which includes two Phases. Phase 1 is to do Object segmentation (use SfM-Net model to extract random actions). Phase 2 is a faster policy segmentation concatenated with phase 1 and use phase 1 output as input in order to narrow the features. It is showing as blow:

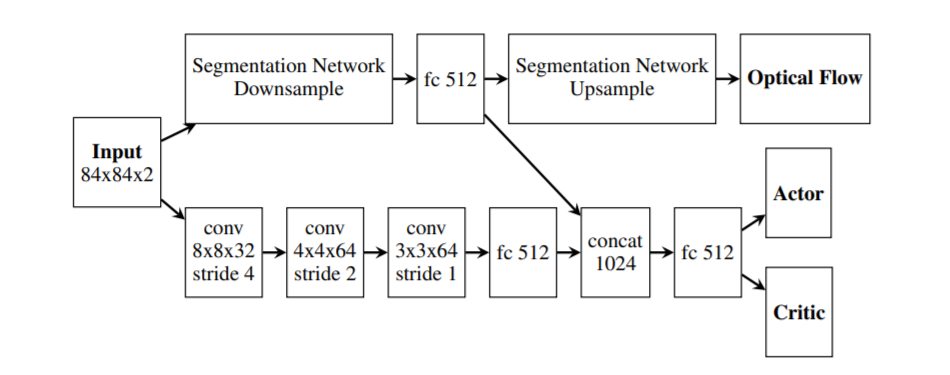


Image source: Unsupervised Video Object Segmentation for Deep Reinforcement Learning

We can simply understand the idea that instead of directly learning a policy from raw images, the agent first learns to detect and segment moving objects by exploiting flow information in video sequences. The learned representation is then used to focus the policy of the agent on the moving objects. Over time, the agent identifies which objects are critical for decision making and gradually builds a policy based on relevant moving objects.

This approach is demonstrated on a suite of Atari games where the ability to detect moving objects reduces the amount of interaction needed with the environment to obtain a good policy. Furthermore, the resulting policy is more interpretable than policies that directly map images to actions or values with a black box neural network. We can gain insight into the policy by inspecting the segmentation and motion of each object detected by the agent. This allows practitioners to confirm whether a policy is making decisions based on sensible information.

**Further discussion**

After the presentation, there was a discussion time, which I also learned a lot. This model was designed to decrease the computation time. In this sense, it is a very good idea. However, an extra layer of neural networks model added will add a high cost and make it more expensive financially. If we have to think about a trade-off, this model may need further improvement. And if this model is mostly focus on Atari games, on my own opinion, we may think of removing the static background in stead of adding another neural network. I proposed this to Pascal, he was happy to hear my idea which also gave me a motivation to continue my study.

**Application to real life**

This model is only at the stage of the thesis and there is still room for improvement from the actual application, nevertheless, it is operational and it is indeed improved in terms of calculation speed. It proposes new ideas for deep reinforcement learning. It can be used practically for vehicle tracking and traffic control.

References:

1. David Zax, Brown Down: UPS Drivers Vs. The UPS Algorithm <https://www.fastcompany.com/3004319/brown-down-ups-drivers-vs-ups-algorithm>
2. Tathagata Das, Spark Streaming: What Is It and Who’s Using It? <https://www.datanami.com/2015/11/30/spark-streaming-what-is-it-and-whos-using-it/>
3. What is streaming data, <https://www.talend.com/resources/what-streaming-data/>
4. swardley , Why big data and algorithms won’t improve business strategy, <https://medium.com/hackernoon/why-big-data-and-algorithms-wont-improve-business-strategy-54e4ebe2398>
5. Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, Katerina Fragkiadaki less, SfM-Net: Learning of Structure and Motion from Video
6. Vik Goel, Jameson Weng, Pascal Poupart, Unsupervised Video Object Segmentation for Deep Reinforcement Learning