Reading 1

(1) Title, author and where it was published

- Title: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Authors: Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova
- Published: Google AI Language

(2) Summary of the content

This paper introduces a powerful new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. BERT is a pre-trained language model that could be used to solve complicated tasks by adding an additional output layer specific to the task. The authors start the walkthrough by discussing the issues related to previous architectural designs, and pinpoint that the previous works suffer from only using unidirectional model architecture. With a brief overview of the history of pre-trained language models, including feature-based approach, fine-tuning approach and transfer learning, the authors then present the details of BERT's model architecture, where it is basically a multi-layer bidirectional Transformer encoder. Later, experimental results over four benchmarks/datasets are presented, demonstrating the effectiveness of BERT. Finally, they conducted ablation studies on the importance of bidirectionality in language modeling.

(3) Strengths of the paper

The authors save room for a more high-level overview of BERT by putting additional details of BERT in the appendix section, highlighting the important concepts that they would like to convey. Transformer itself is already a complex deep learning topic that requires pages to explain, yet bidirectionality in the language context is another academic term that needs careful interpretation. In this paper, the authors first introduce unidirectionality as a left-to-right/right-to-left architecture where every token can attend to previous tokens in the self-attention layers of the Transformer. The problem with bidirectionality is that the word to be learnt is leaked to the layer, causing the model to trivially predict the word itself, and the authors conquer this problem by adding random masking some percentage of the input tokens.
In addition, since BERT is meant to be used on a wide range of tasks, the authors demonstrate its generalizability and applicability across three different data collections and eleven NLP tasks, and pre-trained BERT on two different unsupervised language tasks, making BERT capable of dealing with word/phrase-related problems as well as sentence/paragraph-related ones.

(4) Major critiques

If possible, I think presenting the prediction details on one of the experimental results might be better in visualizing the effectiveness of BERT. Right now, the authors just include the data table showing the relevant metrics. For example, for the The Stanford Question Answering Dataset (SQuAD), the model is tasked to predict the answer text span in the passage. I would like to see what BERT_large has predicted compared with other models, and in this way I might get a deeper understanding of why BERT is a better choice.

Reading 2

(1) Title, author and where it was published

- Title: The Extent and Consequences of P-Hacking in Science
- Authors: Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, Michael D. Jennions
- Published: Blog

(2) Summary of the content

This article discusses a kind of bias called "p-hacking" that is common in scientific research. The authors argue that p-hacking does not have a significant impact as described by other literature.

This paper first provides a definition and a general context of p-hacking – also known as Inflation bias and selective reporting, it occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results. This kind of bias could lead to flawed conclusions and even negatively impact the future reference and citation by other sources, and the authors conjecture that it is problematic because there is a ongoing trend in the scientific field that researchers pursue publishing statistically significant experimental results and frequently exploit practices like stopping data exploration if an analysis yields a significant p-value. They used a technique called p-curve to evaluate the existence of p-hacking in meta-analyses in the publications. The authors first collect the extent of p-hacking with text-mining techniques in all Open Access papers available in the PubMed database, which suggests that p-hacking is widespread in the current literature. Later, the authors use this collection to examine the extent of bias p-hacking caused on the datasets used in meta-analyses, and conclude that meta-analyses might be robust to inflated effects sizes that result from p-hacking.

(3) Strengths of the paper

While covering an abundance of scientific terminology, the authors explain each of them with a clear definition and explanation of the relationship between each term. For example, p-value, p-curve, p-hacking, and evidential values are the main metrics in this discussion, and the readers could be easily confused by these similar terms. However, we are provided with a clear definition that a p-curve is the distribution of p-values for a set of studies, and it could be a helpful tool to assess the reliability of published research. On the other hand, p-values are used to determine the statistical significance of an experimental result, and p-hacking could be identified by p-curves when there is an overabundance of p-values just below 0.05. The readability of this scientific experiment is greatly improved by these interpretations.

Moreover, plots and graphs showcasing the frequency and distributions of p-values in each approach facilitate the visualization of the impact p-hacking. For example, in the text-mining approach, the authors provide the readers with a graphical representation of evidence of p-hacking obtained from the Abstracts and Results section across all disciplines, which demonstrates how frequent scientists present statistically significant results in their works in order to prove the correctness of their research.

(4) Major critiques

The issue of including/excluding the misreported p-values seem to me as an important factor in determining the effect of p-hacking on the datasets used for meta-analyses, and I think the authors could dig more into this correlation. Also, the method they used for detecting the misreporting is not specified nor justified, which is a little ambiguous and thus unconvincing to me.

Reading 3

(1) Title, author and where it was published

- Title: Playing Atari with Deep Reinforcement Learning
- Authors: Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller
- Published: DeepMind Technologies

(2) Summary of the content

This paper presents a convolutional neural network that could learn and extract important features from the control policies from raw video data in complex RL environments. The author applied their model on Atari 2600 games that provide a wide range of high-dimensional visual inputs, and tested to see if their network could learn the rules and outcompete human players. Followed by a brief discussion on the previous works about the deep nerve network's divergence on learning the action-value function and the policies, the authors state that their work successfully conquers failure and has improved the performance of an AI-game-player. With an online learning algorithm called Deep Q-learning with Experience Replay and a Deep Q-Networks, the authors conduct experiments over seven Atari 2600 games, and evaluate the performance on defined as the total reward the agent collects in an episode or game averaged over a number of games against other algorithms like Sarsa, Human experts, and HNeat Pixel, demonstrating that the DQN method is able to outperform other algorithms on six out of seven games.

(3) Strengths of the paper

This paper spends a large portion of time discussing the difficulties of learning control policies from raw pixel inputs, making it clear to the reader of the significant challenge it is dealing with. The model first needs to preprocess the computationally expensive game visual data into gray-scaled 84 X 84 square images that capture the gaming process, and then apply to the experience replay algorithm where the agents' experiences at N time-steps are stored as a reply memory. These stored experiences are then randomly pooled as samples to the Deep A-Networks in each training epoch, and the model will choose actions to execute based on an epsilon-greedy mechanism. Nevertheless, another difficulty about reinforcement learning is that it is unsupervised, meaning the metrics measuring performance has to be determined by the research team. Indeed, the most straightforward metrics – average reward over episodes – turn out to be noisy and unstable even across millions of epochs, whereas the average action-value function seems like a smoother representation.

Moreover, this paper demonstrates the effectiveness of their Deep Q-Networks with the comparison to other state-of-the-art methodologies that are heavily based on prior knowledge about their input visual problems. Since this team trained the network solely based on raw visual inputs yet still achieved highest performance over six out of seven games, it indicates that the performance could be improved by adding other information like audio signals or some prior knowledge.

(4) Major critiques

I think it is worth exploring the reason why the model performs sub-optimally in the S. Invaders game. Are there some special policies or mechanisms that are unique to S. Invaders that cause the failure of learning by this model? Such a diagnostic might produce insight into the improvement of the network architecture. Also, since this reinforcement learning incorporates an online learning algorithm, I think the

research team could examine the effect of learning rates and use discretization on N histories and find out the optimal pair of parameters.