Assignment 1

Cole Lewis

CSCE585

---

1. Pinterest deploys 5 models in parallel to handle recommendations, personalization, advertisement delivery, result tailoring and more. These models are predominantly deep neural learning, convolutional neural networks, and representation learning. [1][2][3]

2. Pinterest uses Amazon Web Services (AWS) for both storage and as computation infrastructure for deploying their machine learning models. [4]

3. Pinterest uses AWS as the sole method by which they deploy their machine learning models. Each model runs as a separate application on AWS and load balances via elastic load balancing. [5] Elastic load balancing works by automatically distributing traffic (user requests, search queries, etc.) across various models to ensure that each model is given proper data and that no model is given too much data or placed under too much strain.

4. Machine learning models are retired upon finding better ones. New machine learning models are tested offline initially before being migrated online to provide service. [6]

5. Models are updated from collections of user events collected in 15 minute intervals from their data pipelines. Every 15 minutes, the models integrate new data and evaluate their results. [7]

6. Pinterest measures their models for effectiveness by conducting comparison trials between learned representations, like with their PinSage model, for example. [8] Measuring for accuracy is done by comparing the results of the model against how users would select certain pins in studies they hold. Efficacy is further measured from offline metrics, A/B tests, and losses or gains in user traffic/engagement.

7. Most of the time, Pinterest's running models update in real-time, incorporating each new data point into their parameters. However, when this approach is not feasible due to an influx of heavy traffic or other such obstacle, models will be brute force updated with the most current data stored. [10]

8. Pinterest decreases the cost of model deployment by using previously trained models and using limited sets of data often labeled in only a single language. [9]

References

1. https://medium.com/pinterest-engineering/pinnersage-multi-modal-user-embedding-framework-for-recommendations-at-pinterest-bfd116b49475
2. https://medium.com/pinterest-engineering/how-we-use-automl-multi-task-learning-and-multi-tower-models-for-pinterest-ads-db966c3dc99e
3. https://samriddhi2958.medium.com/how-pinterest-uses-machine-learning-fb522fb667f8
4. https://aws.amazon.com/solutions/case-studies/innovators/pinterest/
5. https://aws.amazon.com/elasticloadbalancing/
6. https://medium.com/pinterest-engineering/pinnability-machine-learning-in-the-home-feed-64be2074bf60
7. https://medium.com/pinterest-engineering/real-time-experiment-analytics-at-pinterest-using-apache-flink-841c8df98dc2
8. https://medium.com/pinterest-engineering/pinsage-a-new-graph-convolutional-neural-network-for-web-scale-recommender-systems-88795a107f48
9. https://medium.com/pinterest-engineering/how-pinterest-powers-a-healthy-comment-ecosystem-with-machine-learning-9e5c3414c8ad
10. https://medium.com/pinterest-engineering/building-a-real-time-anomaly-detection-system-for-time-series-at-pinterest-a833e6856ddd