Software Architecture of AI-enabled Systems

TECHNISCHE UNIVERSITÄT DARMSTADT

SE4AI Summer Term 2020

Overview

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Introduction
2. Challenges in AI-enabled applications
3. ML patterns
4. Distinguish Business Logic from ML Models
5. Microservice architectures for ML
6. Conclusion

Introduction

TECHNISCHE UNIVERSITÄT DARMSTADT

- What are the major challenges when designing AI-enabled applications
- What are common pitfalls encountered during development
- How can you approach these Problems with Software Engineering?
- Example architectures
- Conclusion

Challenges in AI-enabled applications

Model Deployment Location

TECHNISCHE UNIVERSITÄT DARMSTADT

- In 2019 Openai first announced their new GPT-2 model for natural language generation [1]
- Results are considered by many to be quite impressive (e.g. the unicorn example[1])
- Final model has a file size of over five gigabytes and requires special GPUs to be executed in seconds

[1] https://openai.com/blog/better-language-models#sample1

Challenges in AI-enabled applications

Model Deployment Location contd.

TECHNISCHE UNIVERSITÄT DARMSTADT

Reasons taken from “Building Intelligent Systems” [2]:

- ML models can be big and computationally heavy to execute
  - execution and update latencies
  - operation costs
- intellectual property

Challenge

AI-enabled applications might need complicated deployment setups.

TECHNISCHE UNIVERSITÄT DARMSTADT

Figure 1: Twitter profile of the Tay Chatbot [3]

- Microsoft released an intelligent chatbot `@tayandyou` in 2016
- intended to learn from user interactions
- bot started tweeting racist propaganda hours after launch [4]

Challenges in AI-enabled applications

Model Telemetry contd.

TECHNISCHE UNIVERSITÄT DARMSTADT

- AI-enabled applications need to be supervised
- Complete feedback loop needs to be monitored (input and output)
- Models can learn in production but this requires special care

Challenge

AI-enabled applications might need more/different supervision compared to traditional systems.

Challenges in AI-enabled applications

Multiple models

TECHNISCHE UNIVERSITÄT DARMSTADT

Figure 2: Google Maps Navigation

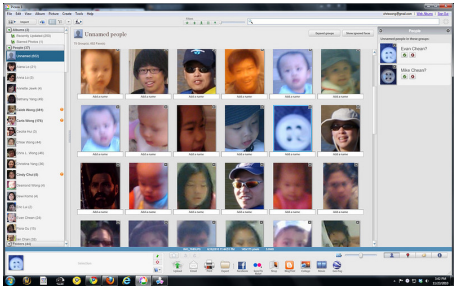

Figure 3: Picasa Face Detection

Challenges in AI-enabled applications

Multiple models contd.

TECHNISCHE UNIVERSITÄT DARMSTADT

- Different models might be necessary depending on the current usage context
- Multiple models might be tested alongside each other

Challenge

AI-enabled applications might need to be able to switch between different models in production.

Figure 4: Caching

Figure 5: Organizational Issues

ML patterns

TECHNISCHE UNIVERSITÄT DARMSTADT

Antipatterns [5]:

- Glue Code
- Pipeline Jungles
- Dead Experimental Codepaths
- Abstraction Debt
- Common Smells

What about design and architecture patterns for ML?

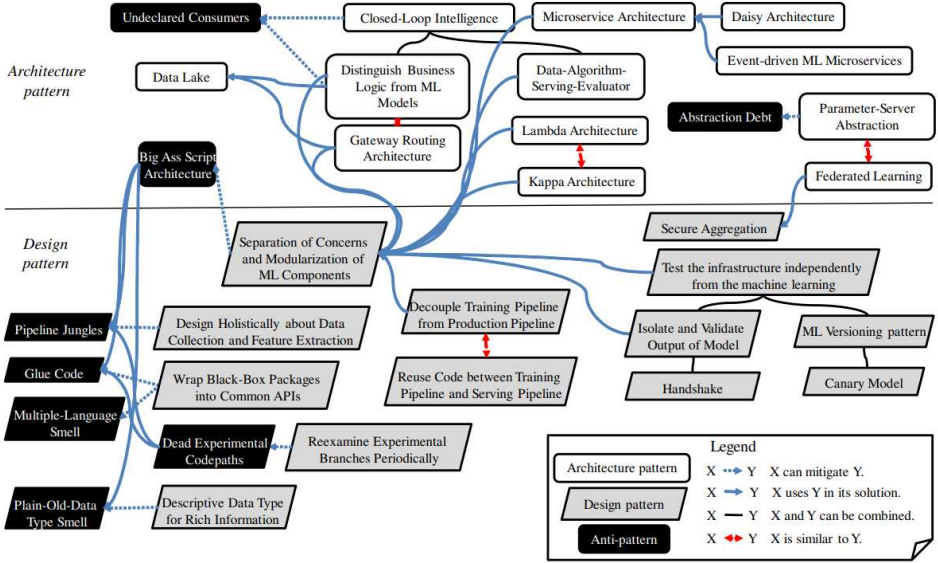

Figure 6: ML Pattern Map, Washizaki et al. [6]

Distinguish Business Logic from ML Models

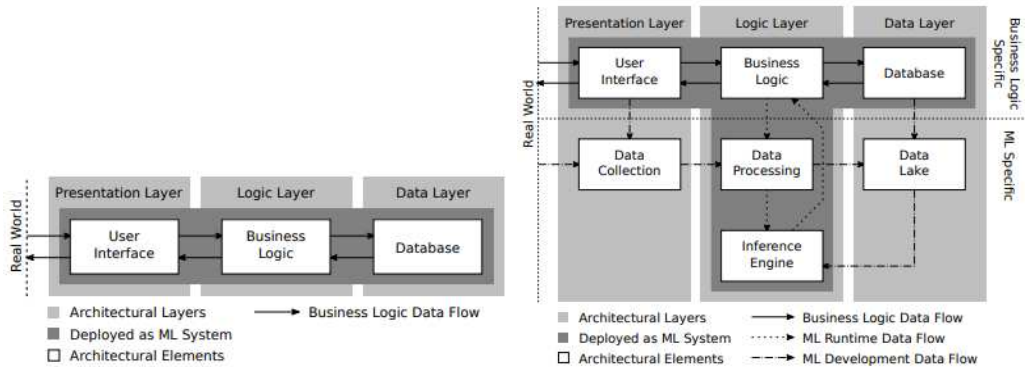

Figure 7: Distinguish Business Logic from ML Models, Yokoyama [7]

Microservice architectures for ML

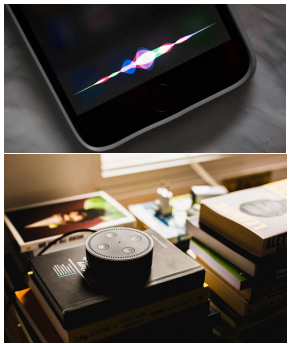

Figure 8: Siri and Alexa

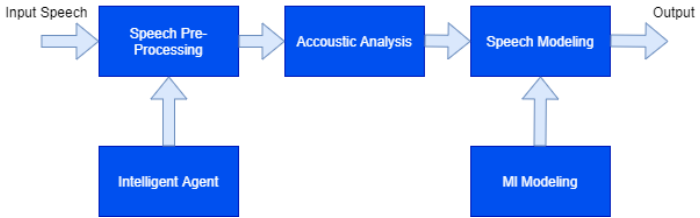

Figure 9: ML microservice architecture, [8]

Microservice architectures for ML

Usage in Production

TECHNISCHE UNIVERSITÄT DARMSTADT

- Rewe Digital uses microservices in production for their product recognition service [9]
- Netflix started deploying jupyter notebooks in production [10]

Microservice architectures for ML

How are the challenges addressed

TECHNISCHE UNIVERSITÄT DARMSTADT

- **Deployment Locations**
  - Models can be wrapped in containers → can easily be scheduled e.g. by using kubernetes
- **Model Supervision**
  - Inputs and Outputs can be closely monitored
  - Replaying of requests possible → have a "production" and a "learning" model
- **Multiple models**
  - Routing of requests is made simpler since microservice interfaces are properly defined
  - Models can easily be substituted or replaced

Capsuling machine learning models inside a microservice allows leveraging existing technology to combat AI-specific challenges.

Conclusion

TECHNISCHE UNIVERSITÄT DARMSTADT

- AI-enabled applications will become more prevalent in the future
- engineers might face new challenges and pitfalls when developing them
- research is currently quite sparse in this particular area of software engineering

Questions?

Bibliography I

[1] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI Blog* 1.8 (2019), p. 9.

[2] Geoff Hulten. “Building Intelligent Systems. A Guide to Machine Learning Engineering”. In: (2018).

[3] Microsoft. *TayTweets Twitter Profile*. URL: `https://twitter.com/TayandYou` (visited on 05/30/2020).

[4] Elle Hunt. “Tay, Microsoft’s AI chatbot, gets a crash course in racism from Twitter”. In: *The Guardian* 24.3 (2016), p. 2016.

[5] David Sculley et al. “Hidden technical debt in machine learning systems”. In: *Advances in neural information processing systems*. 2015, pp. 2503–2511.

Bibliography II

TECHNISCHE UNIVERSITÄT DARMSTADT

[6] Hironori Washizaki et al. “Studying software engineering patterns for designing machine learning systems”. In: *2019 10th International Workshop on Empirical Software Engineering in Practice (IWESEP)*. IEEE. 2019, pp. 49–495.

[7] Haruki Yokoyama. “Machine learning system architectural pattern for improving operational stability”. In: *2019 IEEE International Conference on Software Architecture Companion (ICSA-C)*. IEEE. 2019, pp. 267–274.

[8] Shivakumar Goniwada Rudrappa. *Microservices Architecture in Artificial Intelligence*. URL: `https://medium.com/@shivakumar.goniwada/microservices-architecture-in-artificial-intelligence-60bac5b4485d` (visited on 05/30/2020).

Bibliography III

[9] Rewe Digital. *Unexplored territory: integrate AI into a microservice architecture*. URL: https://mc.ai/integrate-ai-into-a-microservice-architecture/ (visited on 05/30/2020).

[10] Michelle Ufford et al. *Beyond Interactive: Notebook Innovation at Netflix*. URL: https://netflixtechblog.com/notebook-innovation-591ee3221233 (visited on 05/30/2020).

Licences

TECHNISCHE UNIVERSITÄT DARMSTADT

- "two girls illustrations" by "Clarisse Croset" (title slide) is licensed under the Unsplash license `https://unsplash.com/photos/-tikpxRBcsA`
- "ia-siri" by portalgda is licensed under CC BY-NC-SA 2.0 `https://www.flickr.com/photos/135518748@N08/42075167191`
- "Portrait of a lifeless Alexa." is licensed under Unsplash License `https://unsplash.com/photos/k1osF_h2fzA`
- "Database icon in the Tango style." (Figure 4) by "dracos" is licensed under CC BY-SA 3.0 `https://commons.wikimedia.org/wiki/File:Applications-database.svg`
- "Team work, work colleagues, working together" (Figure 5) by "Annie Spratt" is licensed under the Unsplash license `https://unsplash.com/photos/QckxruozjRg`
- "smartphone turned on inside vehicle" (Figure 2) by "Isaac Mehegan" is licensed under the Unsplash license `https://unsplash.com/photos/7x5V13744KM`
- "Face detection" by "Chris Wong" (Figure 3) is licensed under BY-NC-ND 2.0 2.0 `https://flic.kr/p/8VLr4A`

License

TECHNISCHE UNIVERSITÄT DARMSTADT