Bringing AI from Research to Product: Lessons Learned

Tom Meyer

Personio

tom@tom.to

7 Dec 2023

About Me

Spent much of the past 14 years turning AI/ML/CV research into products.

- 20 Years at Startups, and 6.5 years at Facebook
- Roles as VP Eng, CEO, CTO, Sr EM, Principal Eng, Architect
- Patents: GANs, Steganography, AR, Interactive Ads, Affective Computing
- Publications: Hypertext, 3D User Interfaces, VR, Narrative Intelligence

Common Themes of this Talk

- Machine Learning lets you build amazing, magical prototypes. 80% is already working right away!
- And then the real work starts:
  - Sourcing, licensing, or creating more training data
  - Labeling, cleaning and curating it
  - Working around edge cases and weird bugs
  - Optimizing for compute, speed and equipment cost
    - (AI is very computationally expensive)
  - Integrating it into a smooth product experience
  - Making sure it doesn’t go off the rails, isn’t discriminatory, off-brand, etc

SNIBBEINTERACTIVE Magic Mirror (2009)

- Used depth camera and Computer Vision to recognize user's pose.
- Animated 3D character as a reflection of the user's behavior.

SNIBBEINTERACTIVE

Magic Mirror (2009)

- Before big Deep-Learning breakthroughs, so used classic Computer Vision algorithms
- Depth output from camera was extremely noisy and imprecise
- Spent most of the engineering time cleaning it up and denoising
- Simply identified a few principal points (hands, feet, head, center of body)
- People would very quickly test it, push it to the edge, and break it
  - Hopping on one leg, spinning, etc
- Didn’t deal with edge cases like multiple people, wheelchairs, etc.

oMobio A Vision Quest (2014)

- Augmented Reality Phone Camera game
- GPU-accelerated computer vision
- Recognized edges and colors in camera input, and used them to solve platform puzzles.

A Vision Quest (2014)

- Wrote preprocessors to deal with slightly different GPU syntax on iOS, different Android models, and desktop.
- Computer vision code could run on phone GPUs very fast (video frame rates), but it was slow to copy into CPU for game (collision detection, etc), so gameplay used still pictures.
- Commercializing a free-to-play game takes an enormous marketing budget.

idavatars® (2015)

- iPad-based interactive healthcare avatar
- Combined in-house and IBM Watson technologies to:
  - Animate a 3D character
  - Recognize the user's speech (ASR)
  - Used camera to recognize the user's emotions
  - Go through a script to engage the user and collect healthcare information (NLP)
  - Use text classifier to identify relevant healthcare facts from a database to share.
  - Respond verbally to the user (TTS)

idavatars® (2015)

- Extremely aggressive technology goals:
  - Voice input & output, animated avatar that can recognize emotions and collect useful information from patients.
- NLP was generally scripted, so it was consistent and followed the use case
- Having the computer ask questions and having the user answer is easier to script out

Generative Style Transfer (2016)

- First use of Generative Content at Facebook.
- Used Deep Learning to transform user's photos based on an artist's style.
- Real-time, on-phone neural network execution.
- Processed photos, and video at 12fps.

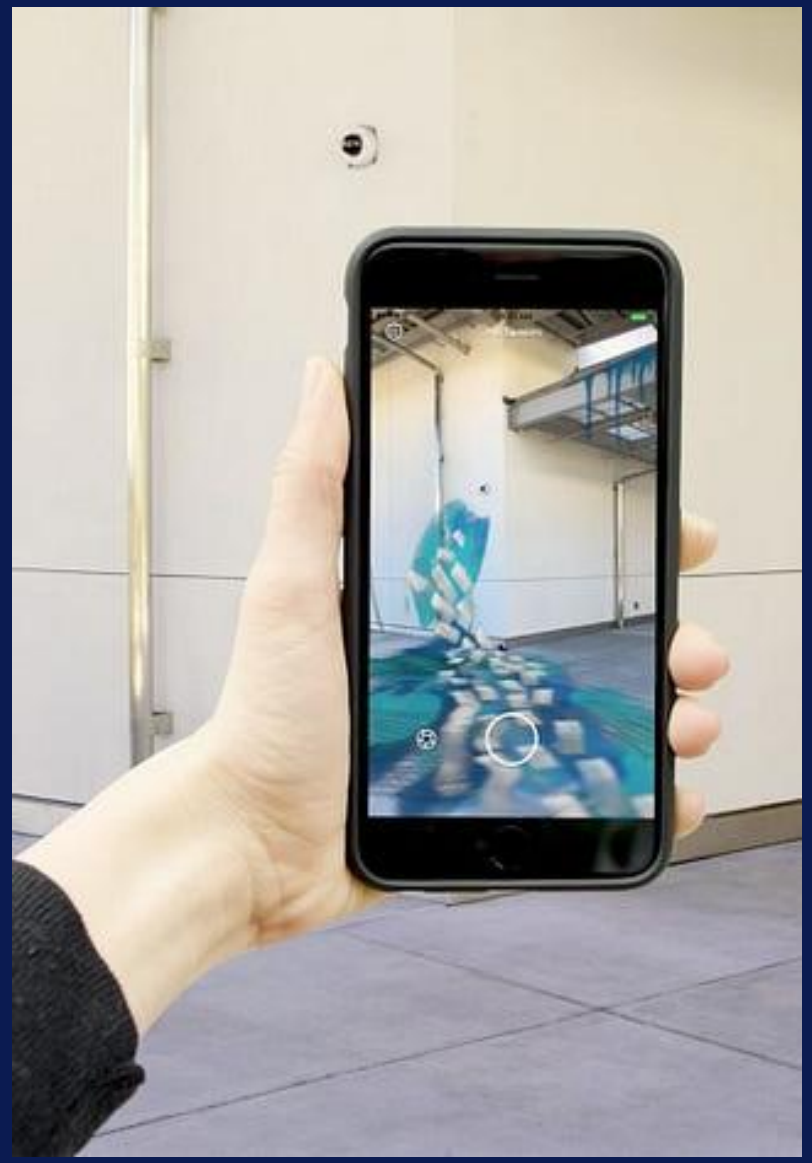


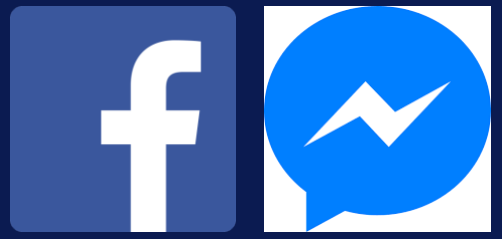
Generative Style Transfer (2016)

- Video-rate style transfer on mobile phones required shortcuts:
  - Lower resolutions (256x256)
  - Simpler networks (not as faithful to the artist's style)
- Hand-picked examples looked great (especially if the subject was posed similarly, etc), but took a lot of user experimentation to make good style transfer photos
- Set off an optimization war between CPU & GPU neural network execution people, where every week one or the other was drastically sped up
- Model compression and simplification became super-important to move toward shipping
- Optimizing for older phones and reducing runtime size became major efforts

Mobile SLAM (2017)

- Adapted Oculus' technology for inside-out tracking used in Meta Quest
- Facebook's version of Apple's ARKit and Google's ARCore
- Could place 3D content in the world as a camera filter





Mobile SLAM (2017)

- Basic demo worked very quickly
- Lots needed to get working on most common phones
  - Needed camera lens parameters for every phone
  - Phone accelerometers & gyroscopes very different or buggy
  - Auto-focus would break the algorithm
  - Had to work on older or medium-end phones
- Memory usage over time, losing tracking occasionally
- Hard keeping code for Oculus VR & mobile phone AR use cases in sync
- Apple & Google both released their own AR frameworks as we prepared to release ours

Person Segmentation (2018)

- Identify a person in the foreground, so you can replace or blur the background
- Real-time Deep-Learning on Handset
- Used a common ML runtime shared with all NN visual effects
- Part of Spark AR, so users could create their own effects using foreground/background segmentation, blurring, etc

Person Segmentation (2018)

- Couldn’t use FB user selfies due to privacy rules
- Training dataset took a lot of work:
  - Selfies with different kinds of consumer phones
  - Different camera quality & resolution
  - Different poses and lighting conditions
  - Gender, age & racially balanced group of people
- Data set labeling:
  - Needed a custom labeling flow to outline the person
  - Hair detail was very hard
- Hats, headphones, funky hairstyles, multiple people, animals, confusing backgrounds
- Very high quality standard – almost didn’t ship, with same artifacts that other products shipped with and were fine.

Identity Attack Discovery (2020)

- Created an Ensemble model to predict fake ID attacks
- Used signals from:
  - Visual AI forgery detection models
  - Content-based AI forgery models
  - Human review of ID documents
  - Temporal clustering by locations, devices, countries, etc.
- Sped up data pipeline from 24 hours to 15 minutes so we could get useful warnings

Identity Attack Discovery (2020)

- Highly regulated and complex space:
  - Creating fake ID documents is illegal, so couldn’t create fake data
  - Possessing fake IDs or images of them is illegal in most countries, so no ground truth
  - Possession of real IDs is highly regulated
  - Rules vary from country to country, and even within a country
- Some fakes are obvious (especially by 11-year olds trying to pass age verification), some are made by nation-state actors
- Never able to measure true false-positive and false-negative, due to no ground-truth data
- Legal, privacy, and policy reviews
  - Explainability of model was very important to be able to launch

Fast Creation of Classifiers for New Integrity Policies (2022)

- New types of hate speech and other violating content can arise nearly overnight (Jan 6, Ukraine/Russia, Israel/Hamas, etc)
- Need to quickly find examples to train classifiers on, in a matter of days or weeks. Was taking months to adapt.
- Used LLMs with prompt engineering to identify probably violating content
- Manually reviewed and used to bootstrap train traditional NLP classifiers
- Zero-shot -> Few-shot -> Full classifier

Fast Creation of Classifiers for New Integrity Policies (2022)

- Even with LLMs, required a lot of initial work to set up new policy:
  - search for keywords
  - try prompt engineering variations
  - manual review of possible violations
- LLMs are super expensive to run at FB scale, so needed to aggressively pre-filter candidates with regular expressions or existing simpler classifiers
- No ready pool of manual reviewers that could be used on short notice, since the need was “bursty,” and required reviewers to be fully trained on the new policy’s guidelines

Recap: Common Themes of this Talk

- Machine Learning lets you build amazing, magical prototypes. 80% is already working right away!
- And then the real work starts:
  - Sourcing, licensing, or creating more training data
  - Labeling, cleaning and curating it
  - Working around edge cases and weird bugs
  - Optimizing for compute, speed and equipment cost
    - (AI is very computationally expensive)
  - Integrating it into a smooth product experience
  - Making sure it doesn’t go off the rails, isn’t discriminatory, off-brand, etc

Thank you

Any Questions?

Feel free to reach out to me: tom@tom.to