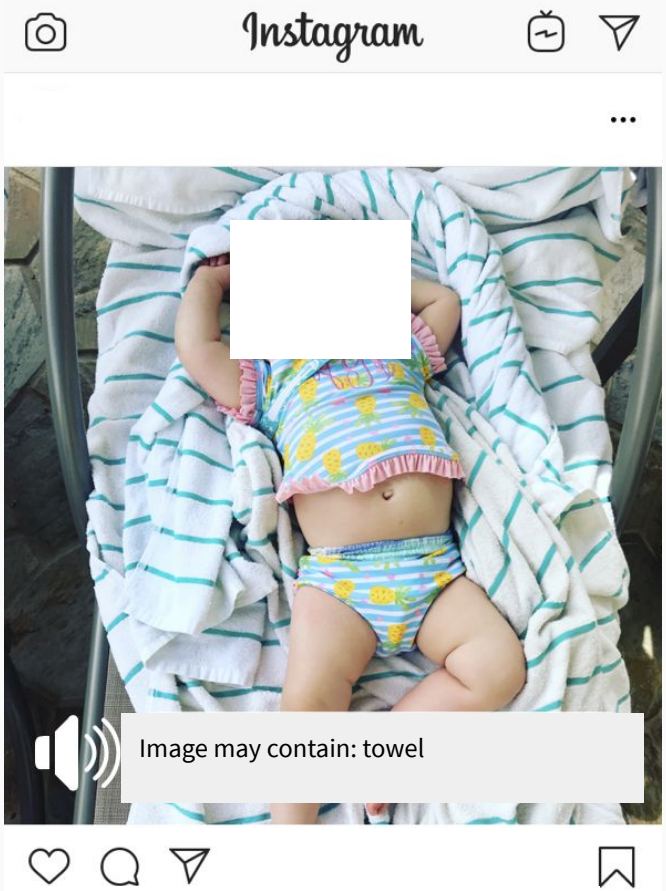
# Building Facebook's visual cortex

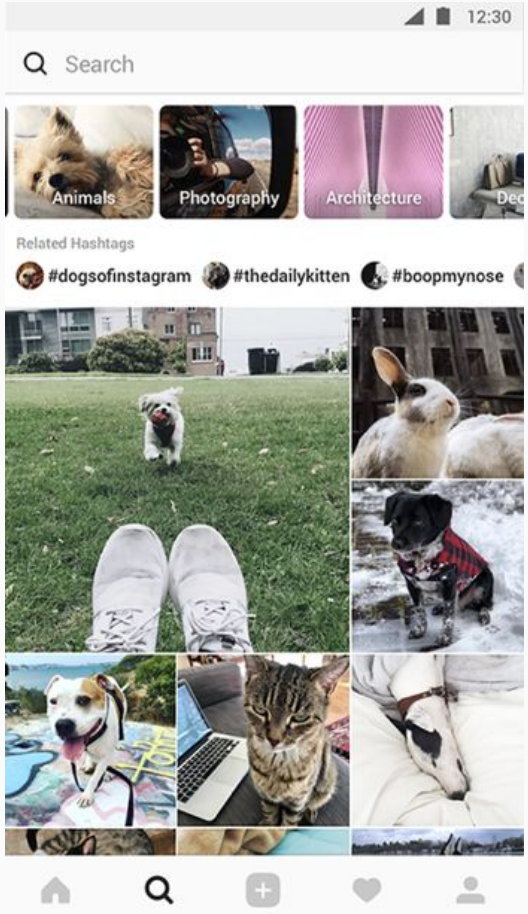
Anmol Kalia

fb.me/fbcortex



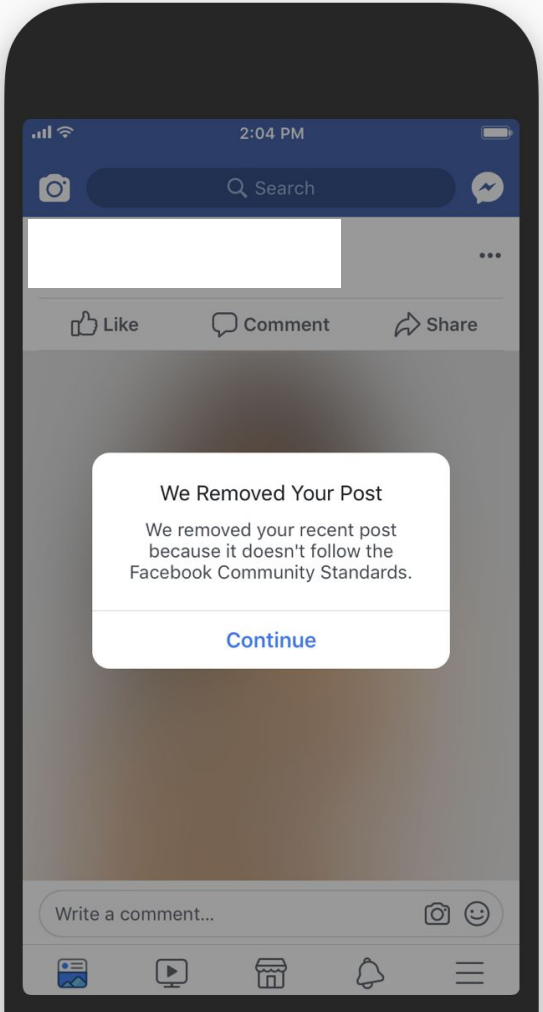

# Automatic Alternative Text

Concepts + OCR to power user-facing product




# Instagram explore ranking

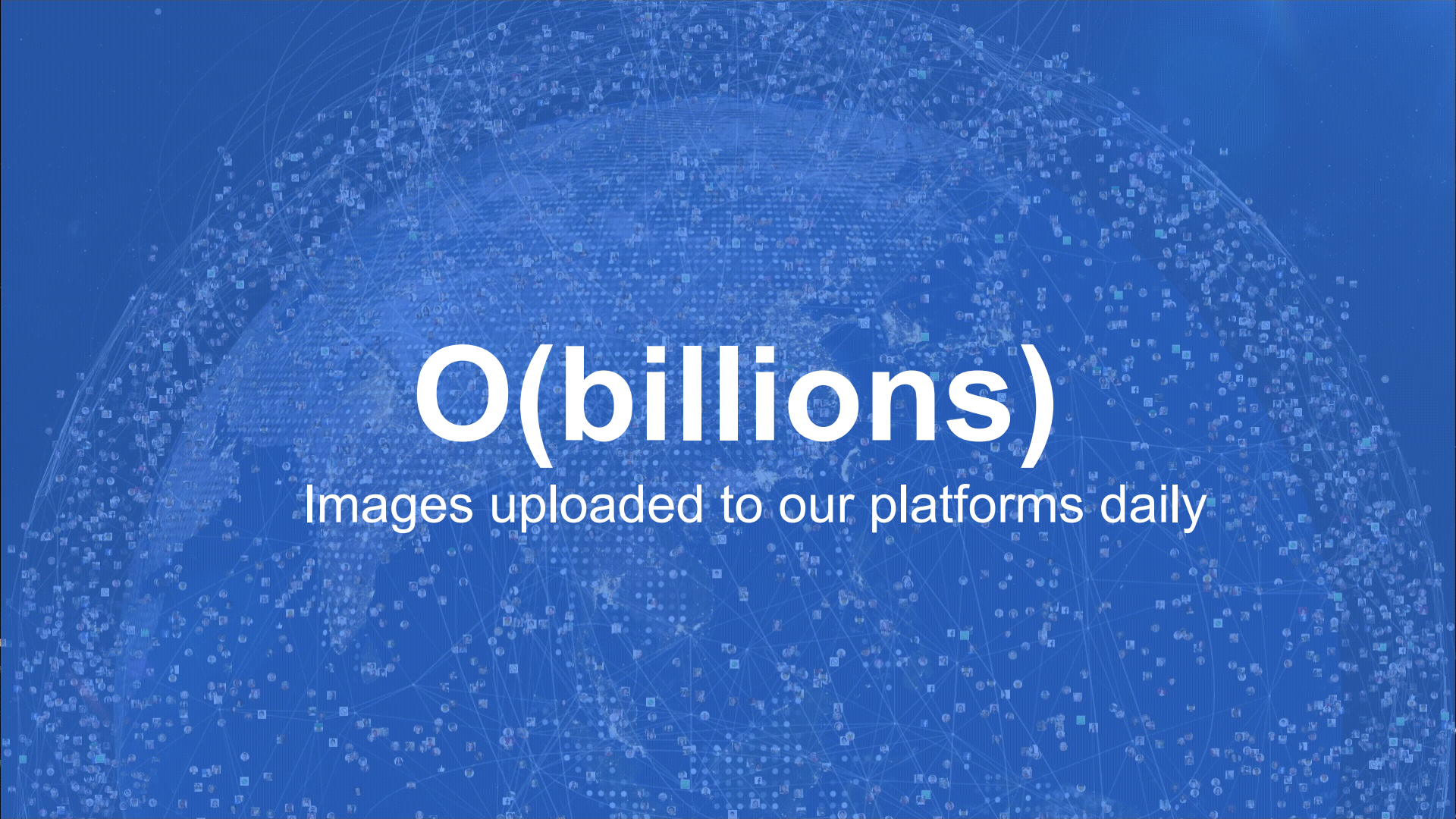
Compact representations as sparse features




# Violating Content Classifiers

Classes as features in multi-modal fusion


O(millions)
Videos uploaded to our platforms daily


O(billions)
Images uploaded to our platforms daily

# Consideration 1: Data is resource intensive

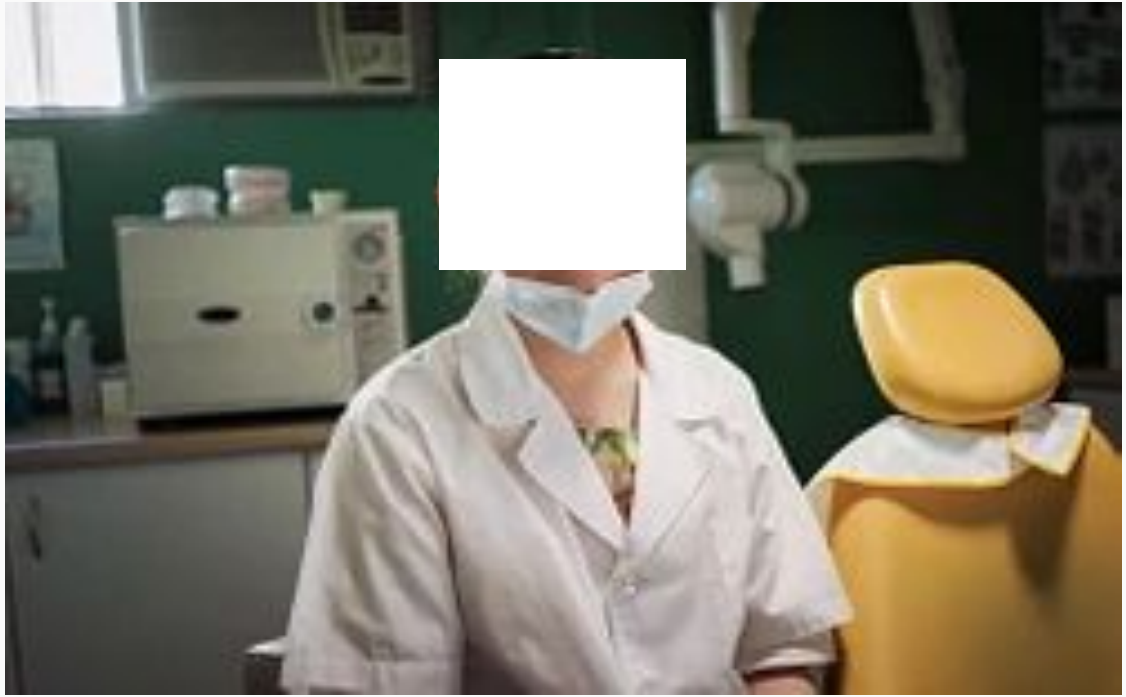
Annotating datasets is resource intensive - ImageNet-1K-ILSVRC2012 today:

| | |
|---|---|
| 1.43M | Images in dataset [1] |
| * 3 | Multi-review with 3 annotators |
| * $0.12 | Price per-image - multi-label [3] |

**~ $0.5M**

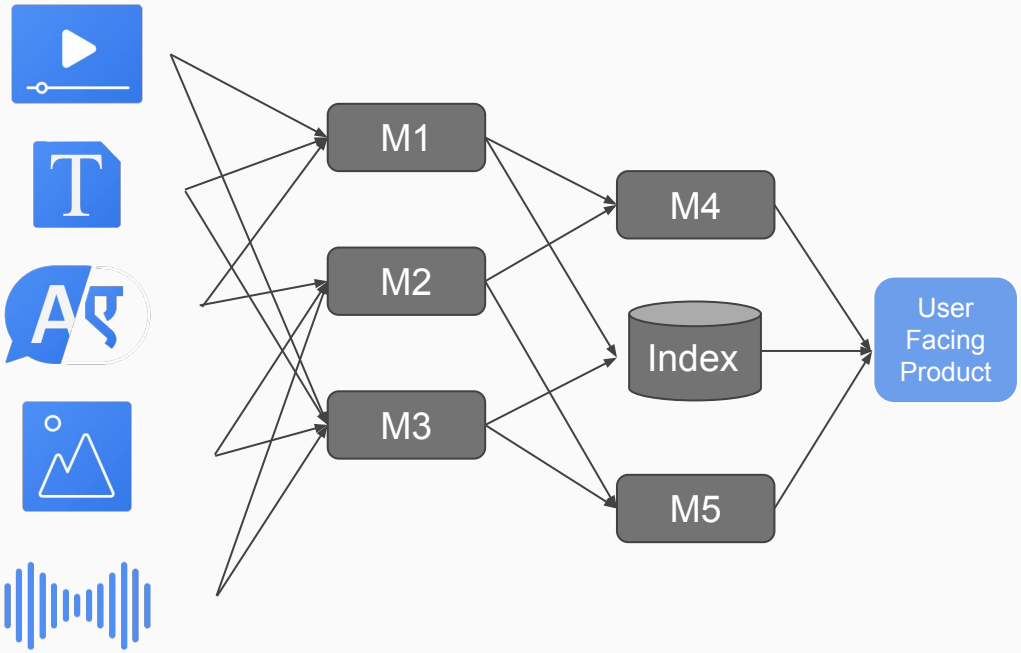
[1] - http://image-net.org/challenges/LSVRC/2012/ilsvrc2012.pdf
[2] - https://medium.com/syncedreview/data-annotation-the-billion-dollar-business-behind-ai-breakthroughs-d929b0a50d23
[3] - https://scale.com/pricing

# Consideration 2: Dynamic demand



Classifier for “mask” or “ice bucket challenge”?

# Consideration 3: Continuous improvement



# Consideration 4: Efficiency

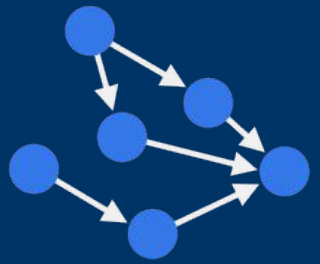
| **Backbone** | **Model Size (# params)** | **Forward pass latency (ms on Intel Skylake, 18 core, 64 GB)** |
|---|---|---|
| ResNet-50 | 25M | ~ 70 ms |
| ResNeXt-101-32x4-48 | 43 - 829M | ~ 150 ms |
| Faster-RCNN-Shuffle | 6M | ~ 600 ms |
| ResNeXt-3D-101 | 21M | ~ 4 sec |

1 billion images run on 1 machine sequentially through ResNeXt-101:

**1736 days** to process ~1 day of photos

Data collection resources

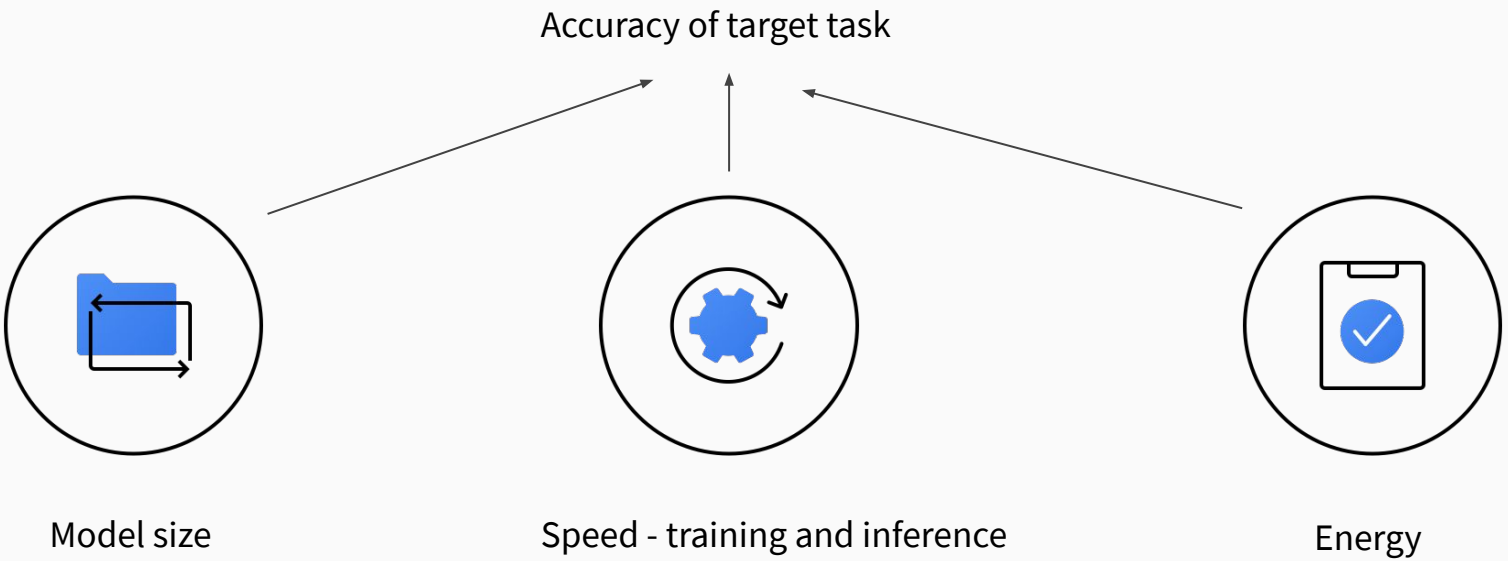
**Efficiency**



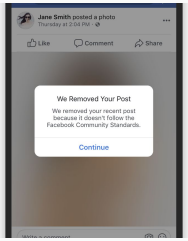
Continuous improvement

Dynamic demand

# Multi-objective optimization problem



# Idea 1: Shared backbone




Policy violating

Not policy violating

MLP + Softmax

Image backbone A

Cat Dog Pillow Towel Bear

2048 dim embedding

Image backbone B

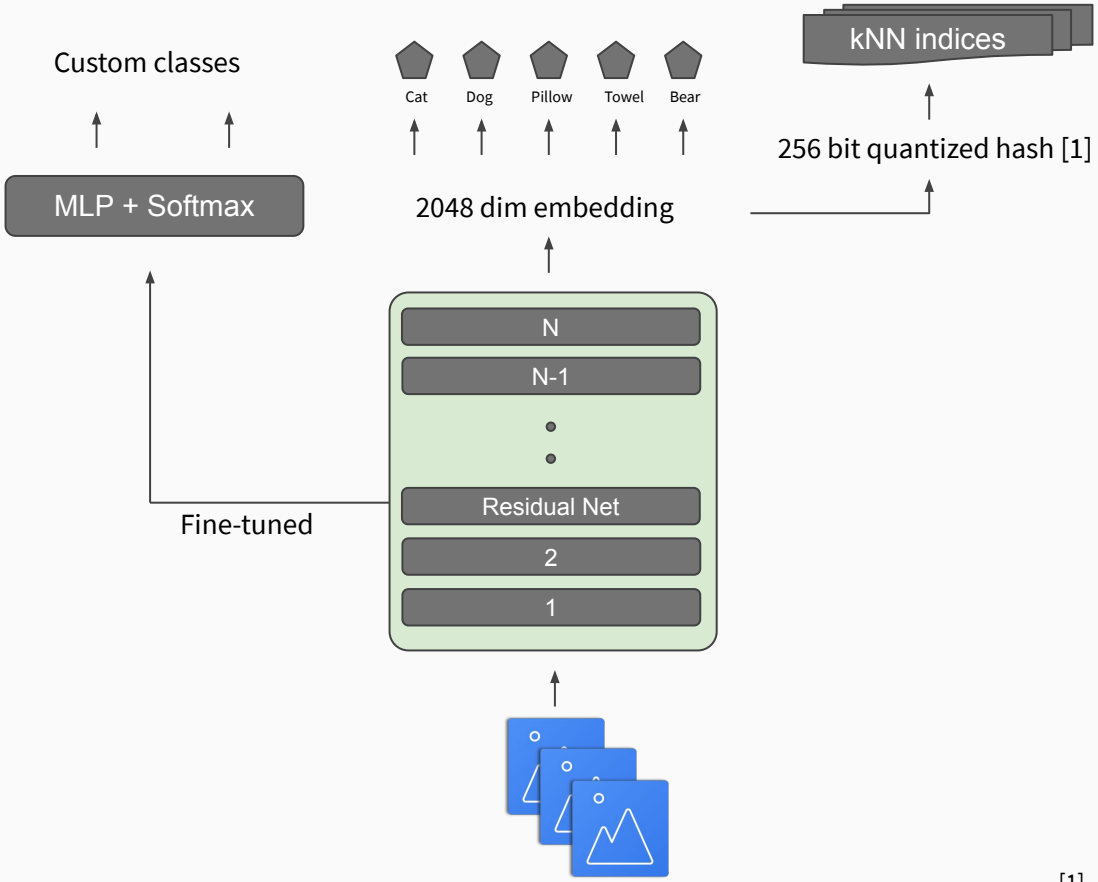
kNN indices

256 bit quantized hash

2048 dim embedding

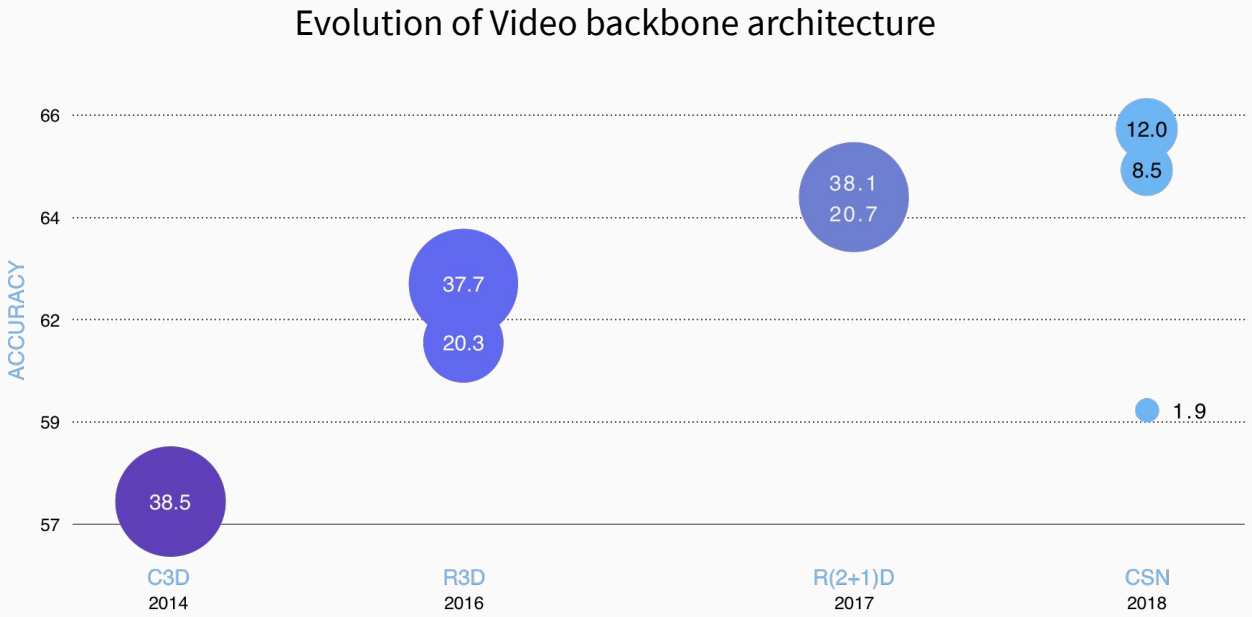
Image backbone C

# Idea 1: Shared backbone




[1] - https://arxiv.org/pdf/1806.03198.pdf

# Idea 2: Focus on efficient backbone architectures




Use 3D conv to model appearance & motion together

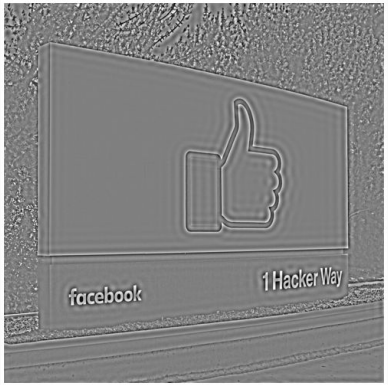
Use Residual network as backbone

Factorize 3D conv Into spatial & temporal components

Factorize 3D conv into channel and spatiotemporal interactions

# Idea 3: Develop efficiency techniques - Octave Convolution


=

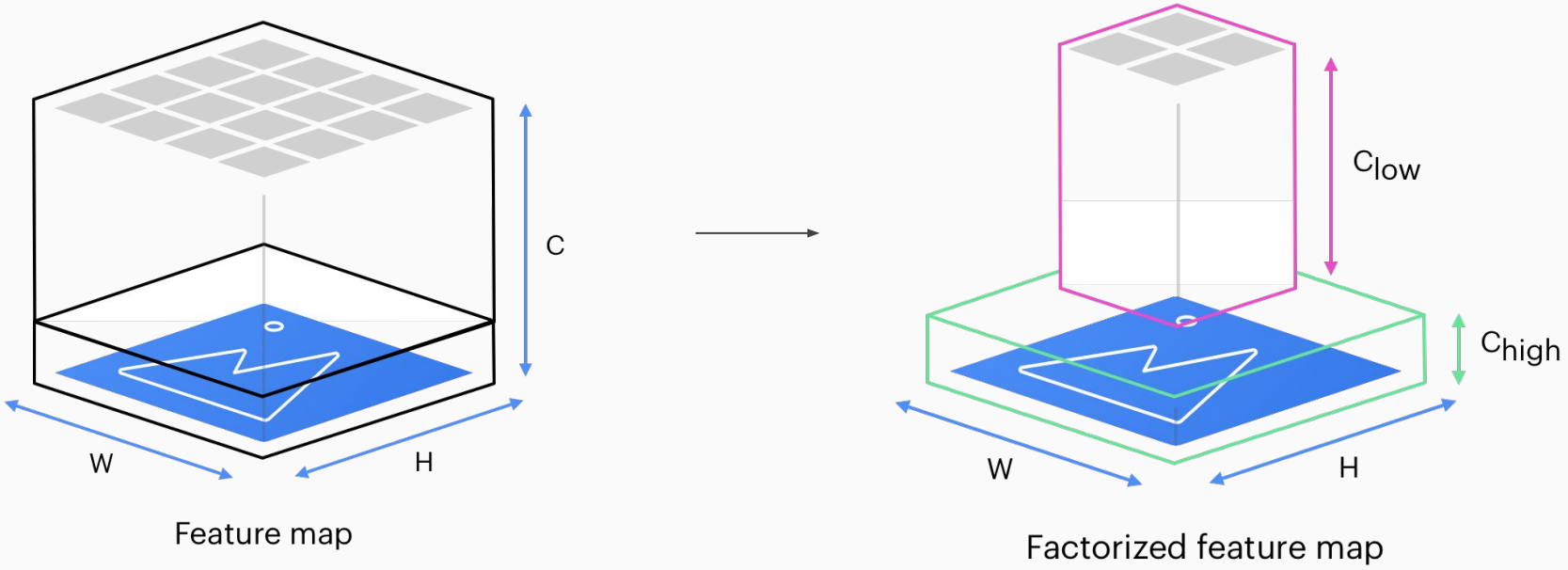



High frequency
(Fine details - edges)

Low frequency
(Global structure)

[1] - https://arxiv.org/abs/1904.05049

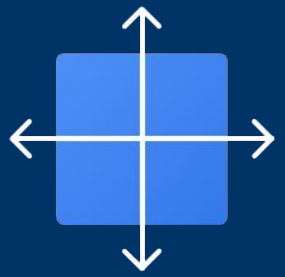
# Idea 3: Invest in efficiency techniques - Octave Convolution



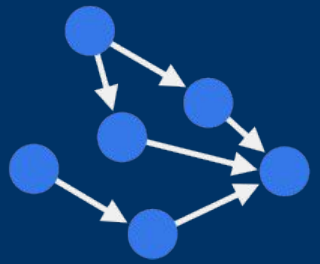

Idea: Store and process feature maps that vary spatially slower at a lower spatial resolution reducing both memory and computation
Drop-in Convolution operator giving **40% drop in GFLOPs and 50% drop in latency for ResNet-50**

[1] - https://arxiv.org/abs/1904.05049

**Data collection resources**



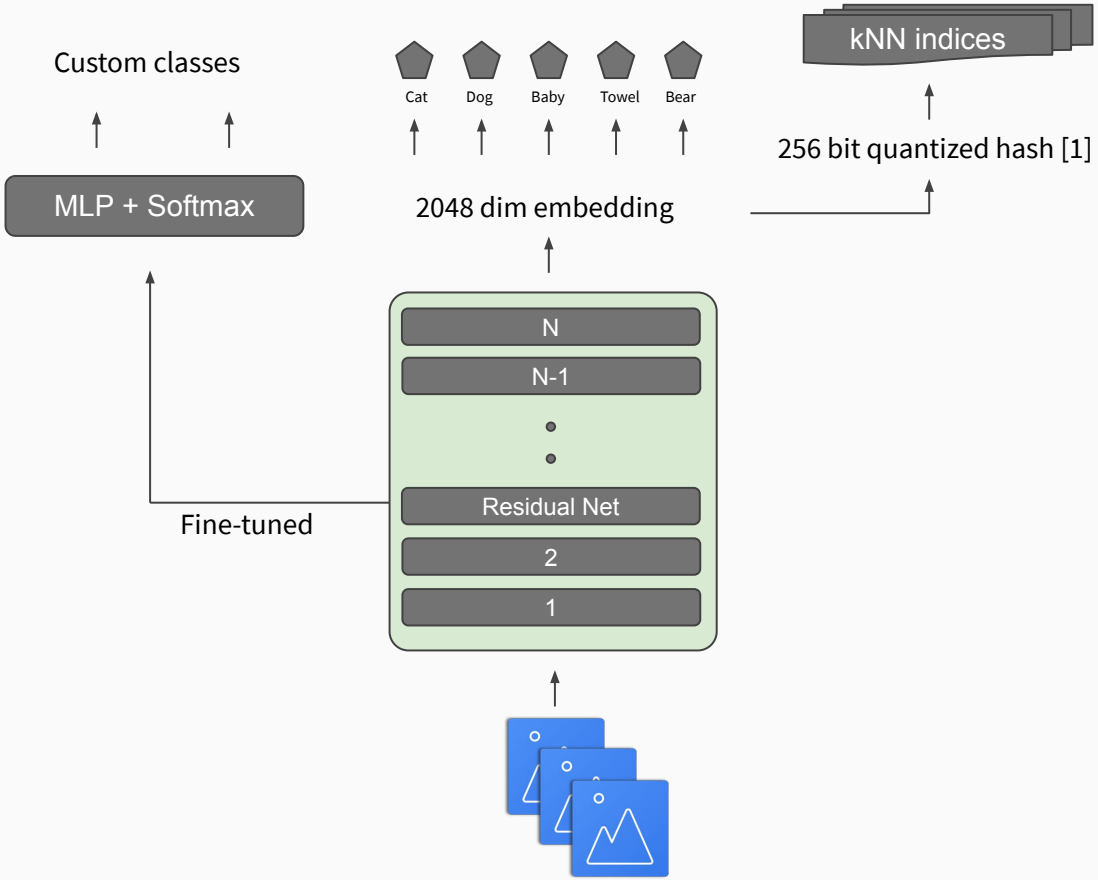
Efficiency



Continuous improvement

Dynamic demand

# Importance of backbone



# Reducing supervision for pre-trained network = less resource intensive annotation

ImageNet pre-trained
Kinetics pre-trained

Fully supervised
(“clean labels”)

# Reducing supervision for pre-trained network = less resource intensive annotation

ImageNet pre-trained
Kinetics pre-trained

Fully supervised
(“clean labels”)

Weakly supervised
(“noisy labels”)

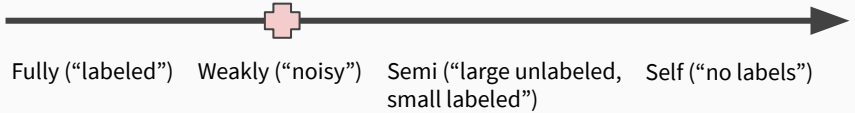
Semi supervised
(“large unlabelled
+ small labeled”)

Self supervised
(“no labels”)

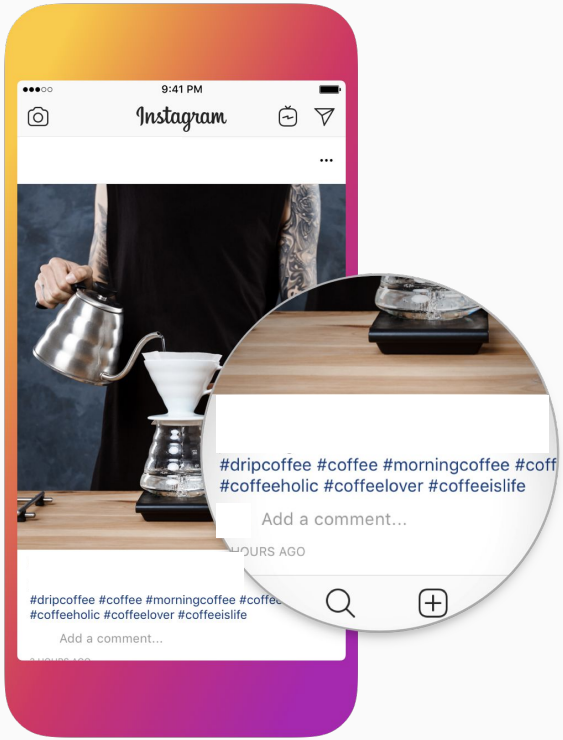
# Exploring the Limits of Weakly Supervised Pretraining

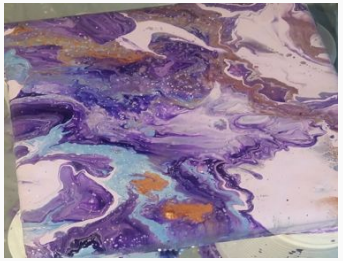
Dhruv Mahajan Ross Girshick Vignesh Ramanathan Kaiming He
Manohar Paluri Yixuan Li Ashwin Bharambe Laurens van der Maaten

Facebook



# Weakly supervised - insights



#love
**Non-visual**

#persiancat #cat
**Missing labels**

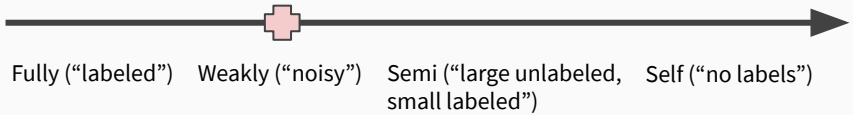
#cat
**Wrong label**

Fully (“labeled”) Weakly (“noisy”) Semi (“large unlabeled, small labeled”) Self (“no labels”)

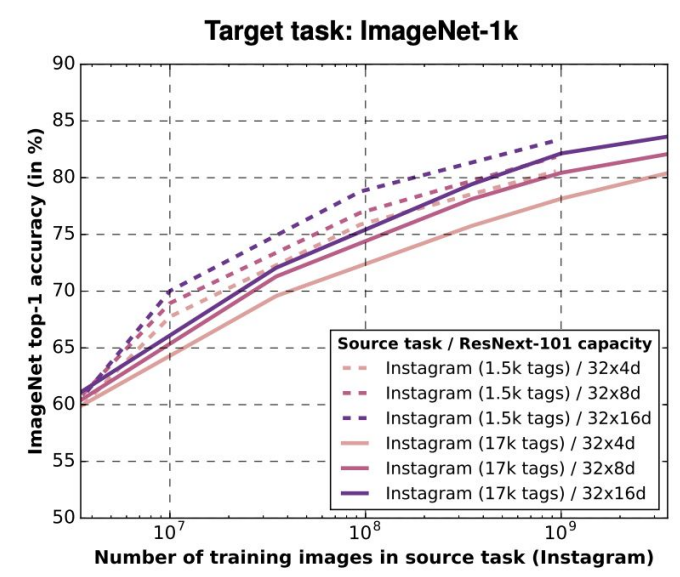
# Weakly supervised - setup

**Goal:** Pre-train to predict hashtags and then evaluate transfer to image classification (ImageNet-1K)

| | |
|---|---|
| **Dataset** | 3.5B public Instagram images, 17K hashtags<br>- Previous largest dataset: JFT-300M |
| **Pre-processing** | Replicate images from low frequency tags<br>De-dup labels based on WordNet synset hierarchy |
| **Loss** | Treat as multi-label, cross-entropy between softmax and vector of k non-zero entries each set to 1/k corresponding each hashtag |
| **Training** | 336 GPUs (42 machines) = 22 days to train |
| **Architecture** | ResNeXt-101-32x{4, 8, 16} |



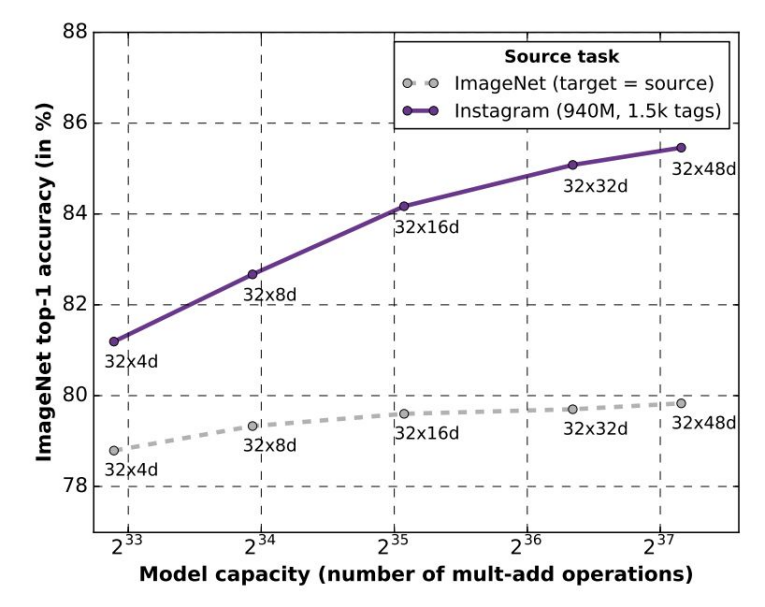
# Weakly supervised - insights




Performance of target task (logistic regression on FC) increases with pre-training image set size

Fully (“labeled”) Weakly (“noisy”) Semi (“large unlabeled, small labeled”) Self (“no labels”)

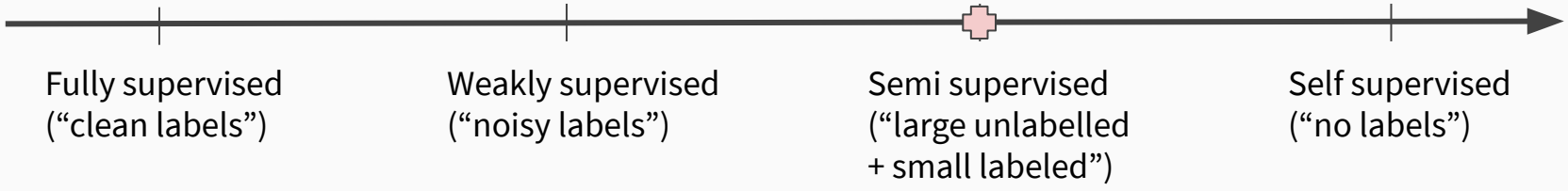
# Weakly supervised - insights




Transfer learning performance bottlenecked by model capacity

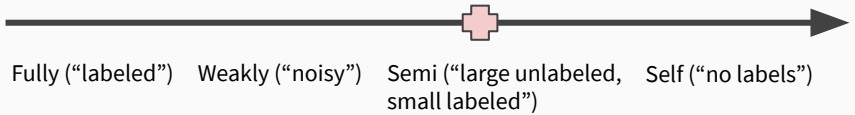
Fully (“labeled”) Weakly (“noisy”) Semi (“large unlabeled, small labeled”) Self (“no labels”)

# Reducing supervision for pre-trained network = less resource intensive annotation



# Semi supervised - setup

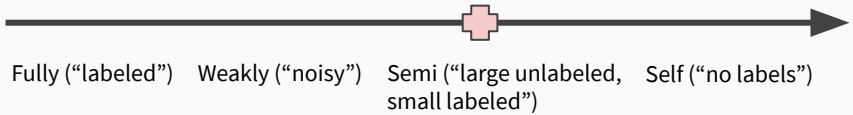
- **Consideration 1:** Hashtag data not accessible to everyone
- **Consideration 2:** Does not leverage the large amount of unlabelled data
  - Mapping it to Instagram, account for 89% of media without hashtags
- **Consideration 3:** Capacity of the models hard to deploy
  - ResNeXt-101-32x48 has **33x** more parameters than ResNet-50



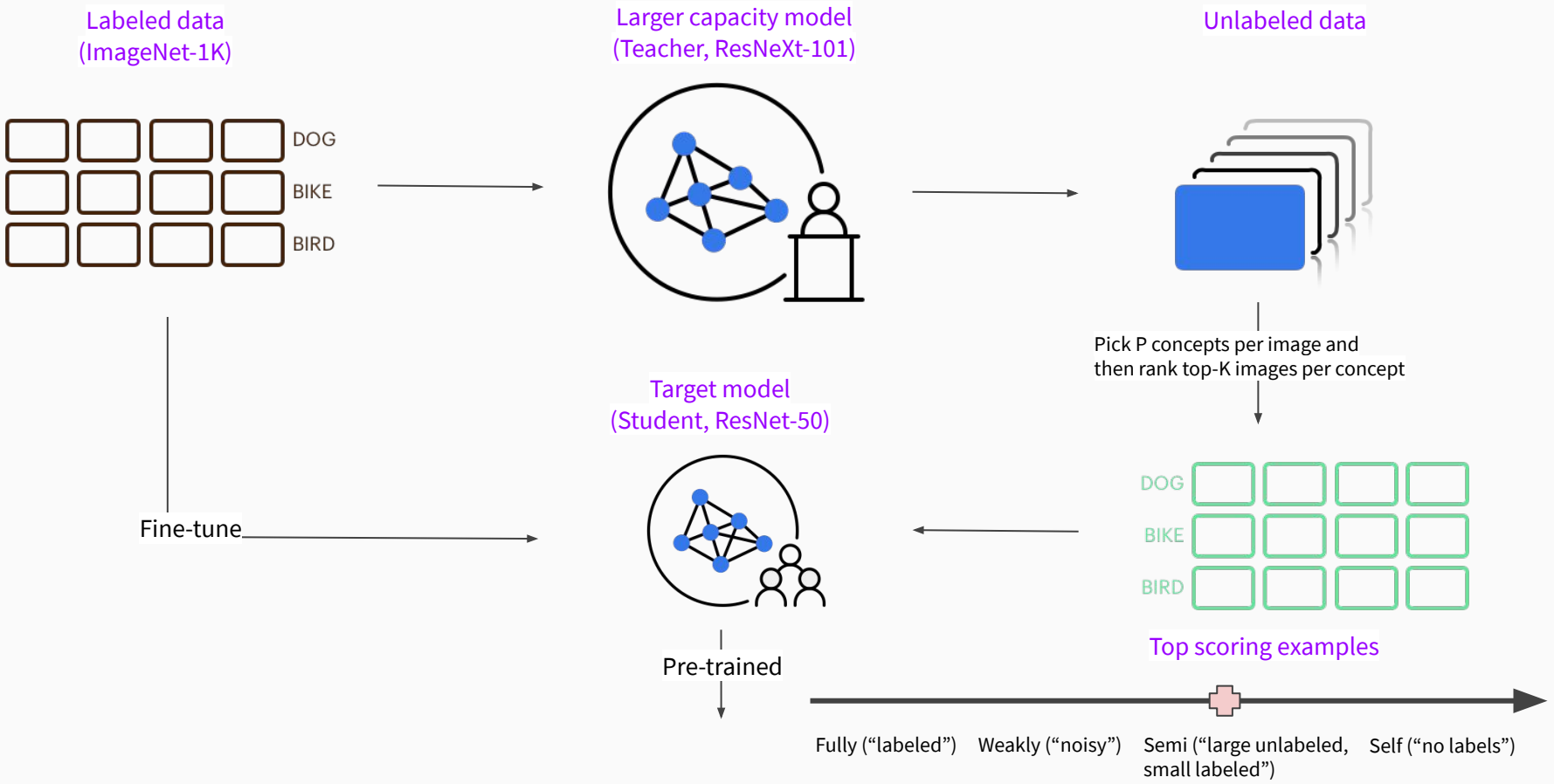
# Semi supervised - setup

**Billion-scale semi-supervised learning for image classification**

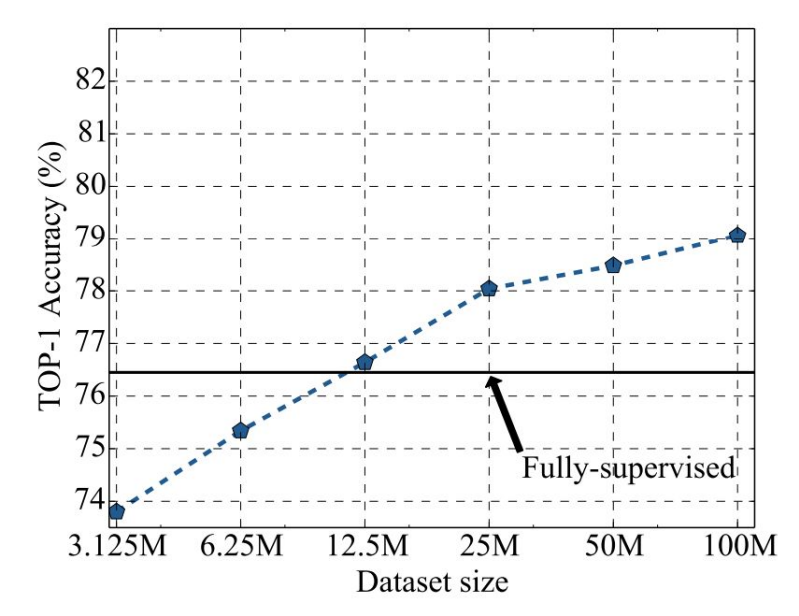
I. Zeki Yalniz  Hervé Jégou  Kan Chen  Manohar Paluri  Dhruv Mahajan
Facebook AI



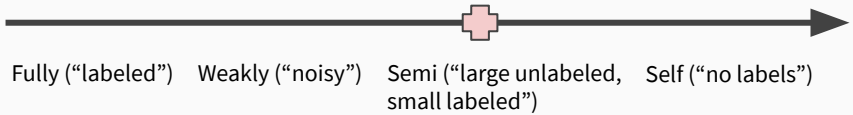
# Semi supervised - setup



# Semi supervised - insight




Value of unlabelled data - accuracy on ImageNet-1K of ResNet-50 (student)



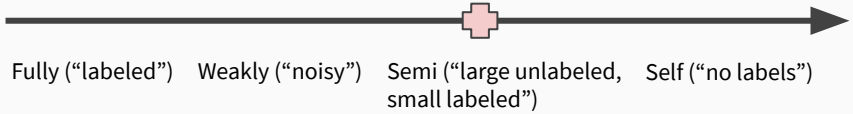
# Semi supervised - insight

| Method | # params | Fully-supervised | Semi-supervised |
|---|---|---|---|
| ResNet-50 | 25M | 70.6% | 79.1% |
| ResNeXt-50-32x4 | 25M | 77.6% | 79.9% |
| ResNeXt-101-32x8 | 88M | 79.1% | 81.2% |
| ResNeXt-101-32x48 | 829M | 79.8% | - |

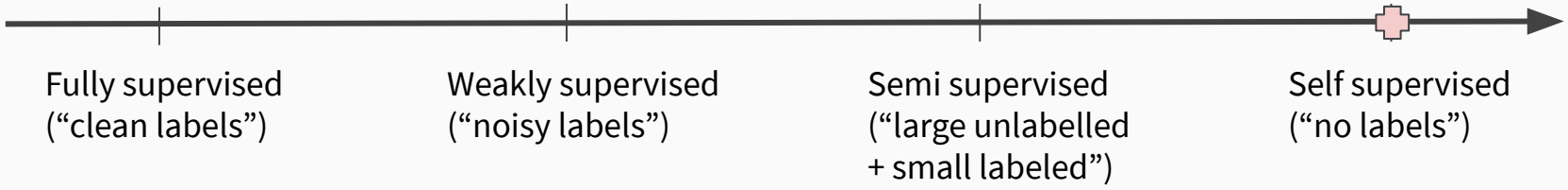
Compute saved:

Fully-supervised (RX101-32x8) ~ Semi-supervised (R50) - 4x less params

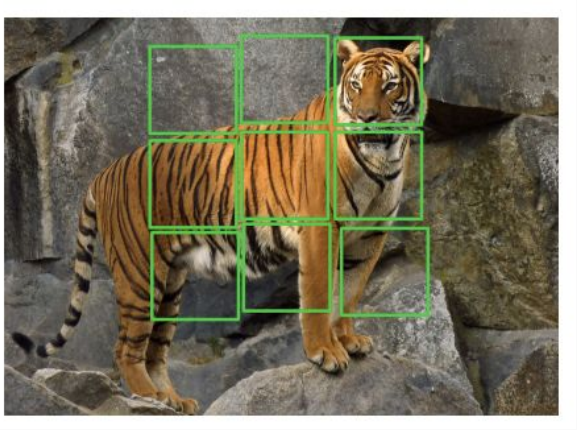
Fully-supervised (RX101-32x48) ~ Semi-supervised (RX101-32x4) - 33x less params



# Reducing supervision for pre-trained network = less costly annotation



# Self supervised - setup



Pretext Task

**Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles**

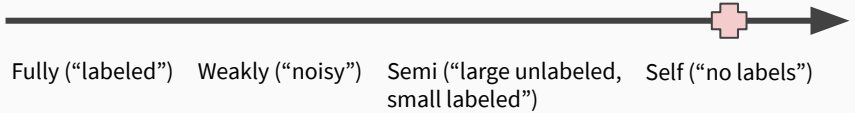
Mehdi Noroozi and Paolo Favaro

Institute for Informatiks
University of Bern
{noroozi,paolo.favaro}@inf.unibe.ch

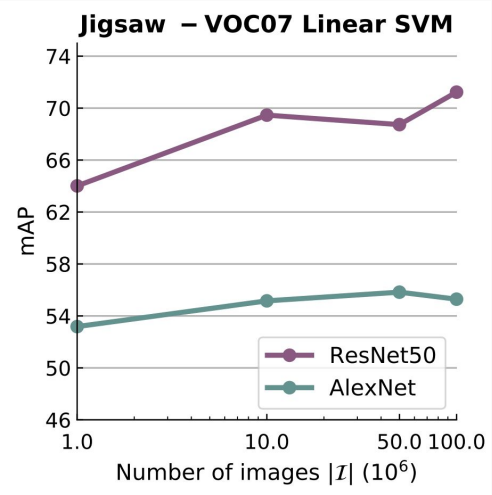
Fully (“labeled”) Weakly (“noisy”) Semi (“large unlabeled, small labeled”) Self (“no labels”)

**Scaling and Benchmarking Self-Supervised Visual Representation Learning**

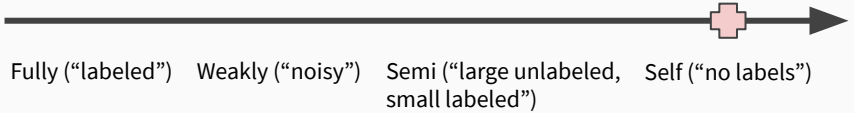
Priya Goyal Dhruv Mahajan Abhinav Gupta* Ishan Misra*
Facebook AI Research



# Self supervised - insight




Scaling pre-text task data & model capacity

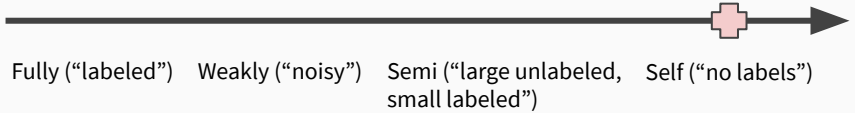


# Self supervised - insight

| Method | Top-1 accuracy of R-50 on ImageNet-1K |
|---|---|
| Fully supervised | 76.4% |
| Weakly supervised | 78.2% |
| Semi supervised | 79.1% |
| Self supervised (Jigsaw) | 45.4% |

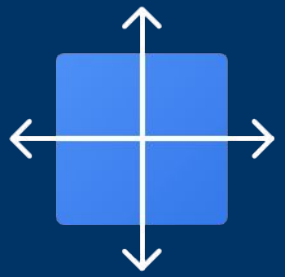
Best ResNet-50 ImageNet-1K Top-1 accuracy across all methods

[1] - https://arxiv.org/pdf/1901.09005.pdf

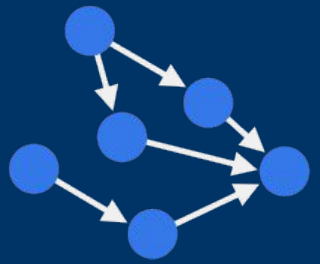


# Production state

Data collection resources



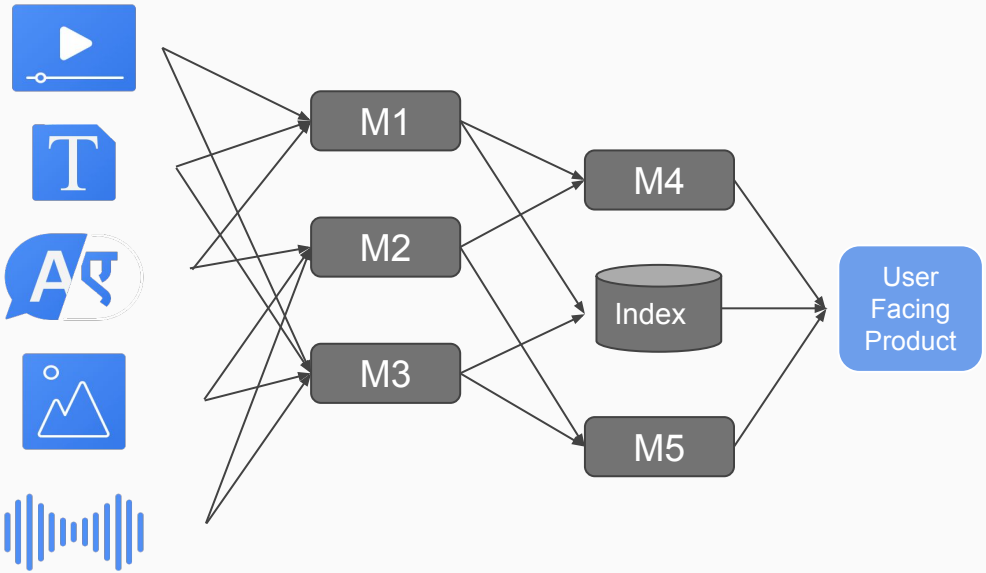
Efficiency



**Continuous improvement**

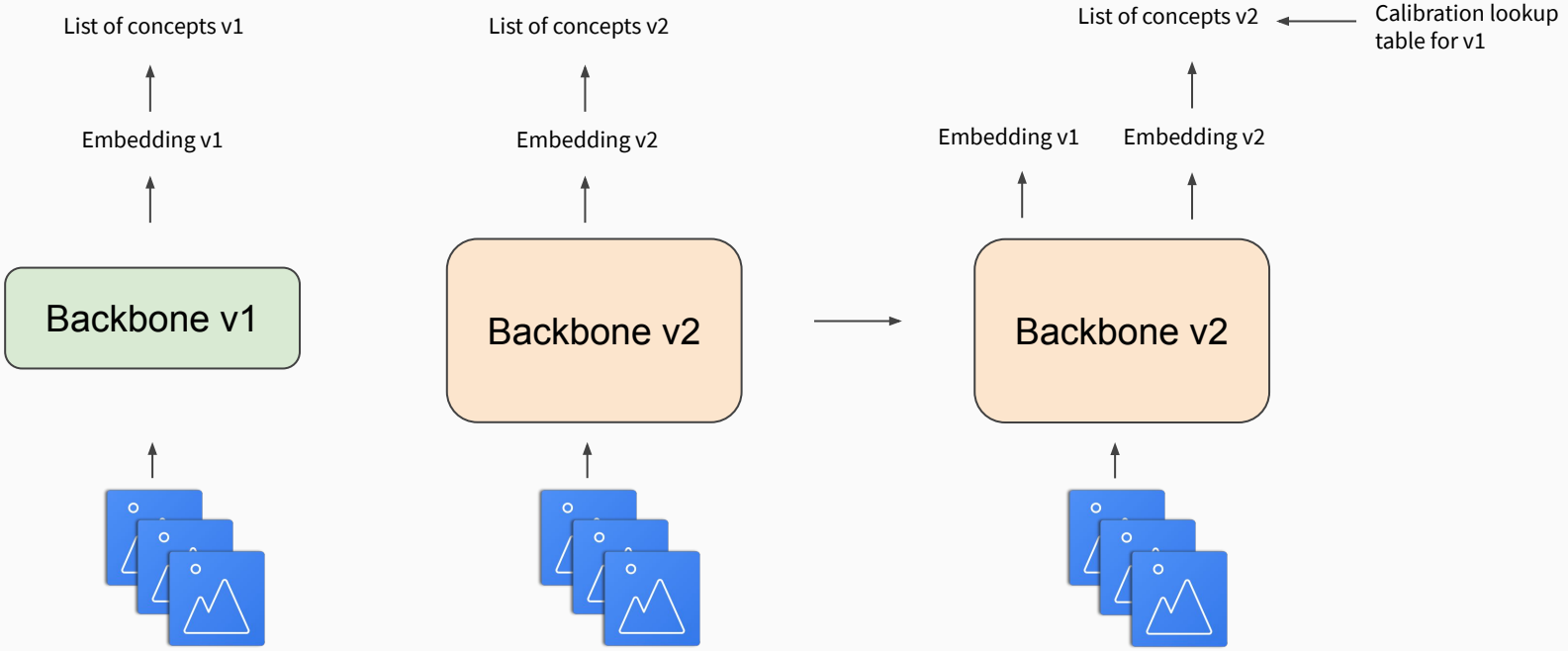
Dynamic demand

# Making upgrades easy




Contract on compatibility support

# Predictable push with N-1 backwards compatibility

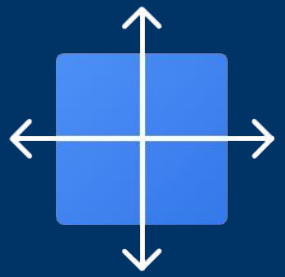



Embedding backwards compatibility via distillation (L2 loss on embedding) [1]

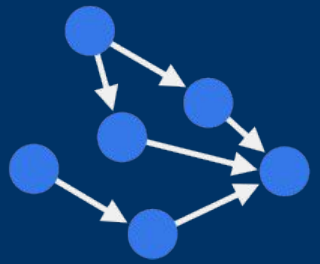
Concepts always added (concepts2 ⊇ concepts1) + impersonating old calibrated score by lookup table

[1] - https://arxiv.org/abs/1503.02531

Data collection resources



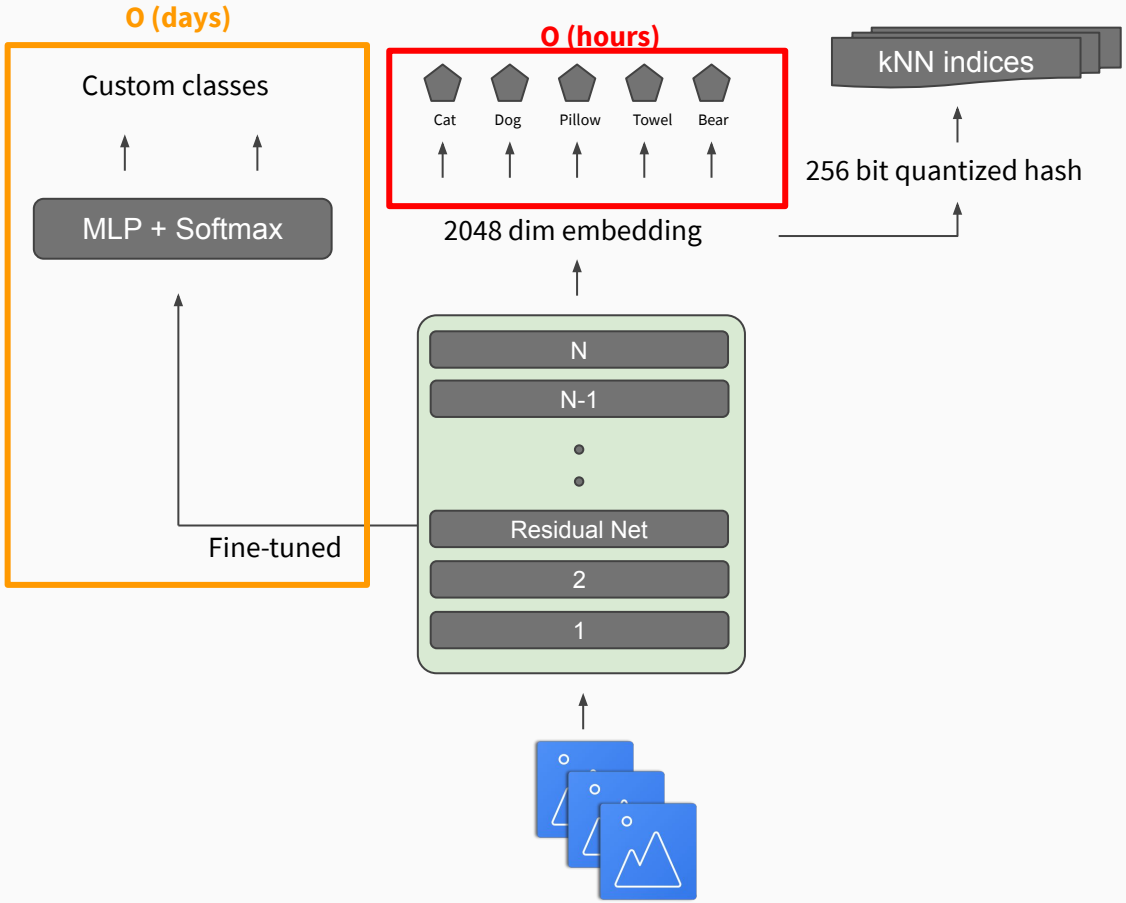
Efficiency



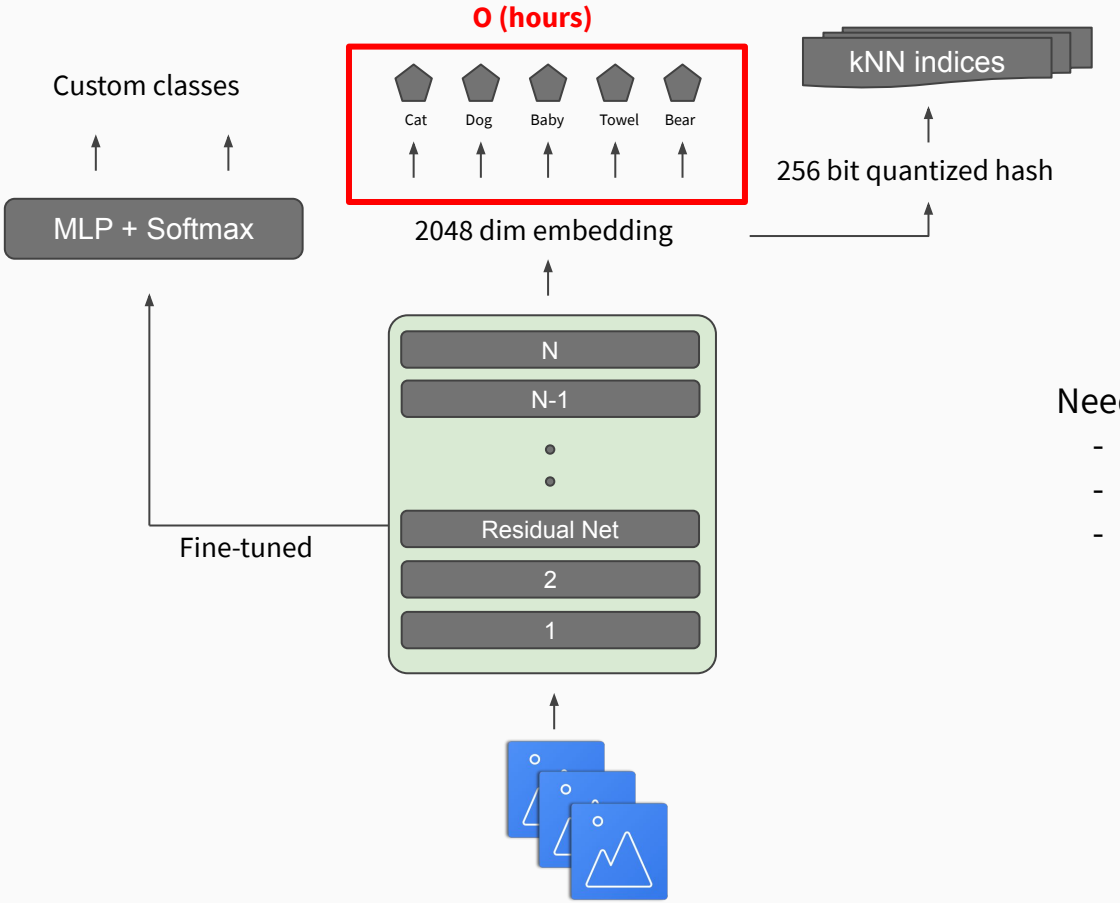
Continuous improvement

**Dynamic demand**

# Fast classifiers for concepts



# Fast classifiers for concepts



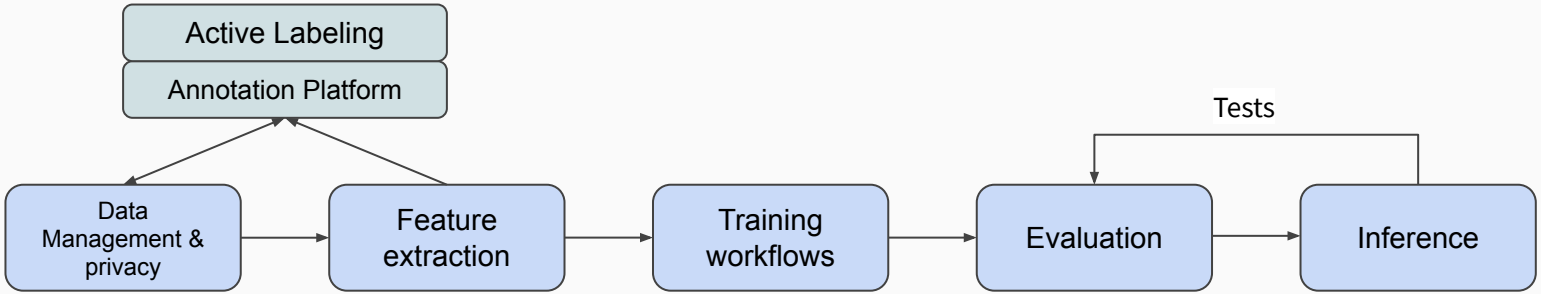

Need platform to quickly train linear classifiers:

- Reproducible & automated
- Monitored
- O(hours) to bring online

# Concept Factory

Manifold sampling using FAISS [1] indices

Concept drift & model unit tests




**O (hours)**

**O (minutes)**

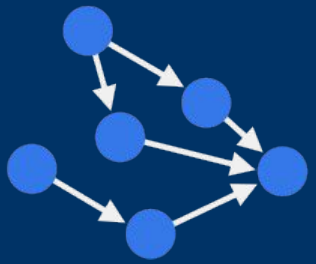
[1] - https://github.com/facebookresearch/faiss

Data collection resources

Reduce supervision

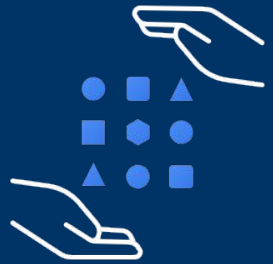
Efficiency

Attack from various angles - operators, architectures



Continuous improvement

Compatibility contracts for fast upgrades



Dynamic demand

O(hours) workflow for trending concepts

# A few pointers…

fb.me/fbcortex

Costly data
- Image weakly supervised
  - **Paper:** “Exploring the Limits of Weakly Supervised Pretraining” https://arxiv.org/abs/1805.00932
  - **Pre-trained model:** https://github.com/facebookresearch/WSL-Images
- Video weakly supervised
  - **Paper:** “Large-scale weakly-supervised pre-training for video action recognition” https://arxiv.org/abs/1905.00561
  - **Pre-trained model:** https://github.com/facebookresearch/VMZ
- Image semi-supervised
  - **Paper:** “Billion-scale semi-supervised learning for image classification” https://arxiv.org/abs/1905.00546
  - **Pre-trained model:** Coming soon!
- Image self-supervised
  - **Paper:** “Scaling and Benchmarking Self-Supervised Visual Representation Learning” https://arxiv.org/abs/1905.01235
  - **Benchmark:** https://github.com/facebookresearch/fair_self_supervision_benchmark

Efficiency
- Architecture Search
  - FBNet**:** “FBNet: Hardware-Aware Efficient ConvNet Design via Differentiable Neural Architecture Search” https://arxiv.org/abs/1812.03443
  - ChamNet: “ChamNet: Towards Efficient Network Design through Platform-Aware Model Adaptation” https://arxiv.org/abs/1812.08934
- Video architecture evolution
  - **2014 (C3D) -** “Learning Spatiotemporal Features with 3D Convolutional Networks” https://arxiv.org/abs/1412.0767
  - **2017 (R(2+1)D)** - “A Closer Look at Spatiotemporal Convolutions for Action Recognition” https://arxiv.org/abs/1711.11248
  - **2019 (CSN)** - “Video Classification with Channel-Separated Convolutional Networks” https://arxiv.org/abs/1904.02811
- Octave Convolution
  - **Paper:** “Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution” https://arxiv.org/abs/1904.05049
  - **Code:** https://github.com/facebookresearch/OctConv
- Optimized kernels
  - FBGEMM (server): https://github.com/pytorch/FBGEMM, https://engineering.fb.com/ml-applications/fbgemm/
  - QNNPACK (mobile): https://github.com/pytorch/QNNPACK, https://engineering.fb.com/ml-applications/qnnpack/
- Catalyzer hash
  - **Paper:** “Spreading Vectors for Similarity Search” https://arxiv.org/abs/1806.03198
  - **Code:** https://github.com/facebookresearch/spreadingvectors

Dynamic Demand
- FAISS
  - **Code:** https://github.com/facebookresearch/faiss

# Thank you!

fb.me/fbcortex