Active learning

Oleksandr Obiednikov

AI Ukraine

21 Sep 2019

Agenda

- Motivation
- What is active learning?
- Faster annotation
- Smarter annotation
- Open questions and Tips & Tricks
- Q & A

Disclaimer

Likely, everything in this talk did not happen in reality; it's just a figment of my imagination.

Coincidences with real people or events are accidental.

Please don't refer to this talk in press :-)

Few words about myself

Research SDM @ Ring Ukraine

GitHub: https://github.com/alexobednikov

Facebook: https://www.facebook.com/alexander.obednikov

e-mail: obednikov.alex@gmail.com

You can talk to me on the following topics:

- Putting AI into production
- CV and Audio analysis
- Metric learning and Re-ID
- Object detection and recognition.
- GANs and domain adaptation
- just chat...

Motivation

~~Models are the bottleneck in Machine Learning~~
~~Data is the bottleneck in Machine Learning~~

Labeled data is the bottleneck in Machine Learning

Motivation

Labeled data is the bottleneck in Machine Learning

Why?

- **Most of the current “practical” ML is supervised learning**
- **Getting labeled data either**
  - **A huge amount of unlabeled data that needs to be annotated**
  - **Expensive to get even a single labeled example**
  - **Noisy annotation**
  - **...**
- **To tackle “long-tail” you need to have a lot of data**

So what to do with it?

Labeling budget or time constraints

Annotate faster / cheaper

Annotate smarter

So what to do with it?

Labeling budget or time constraints

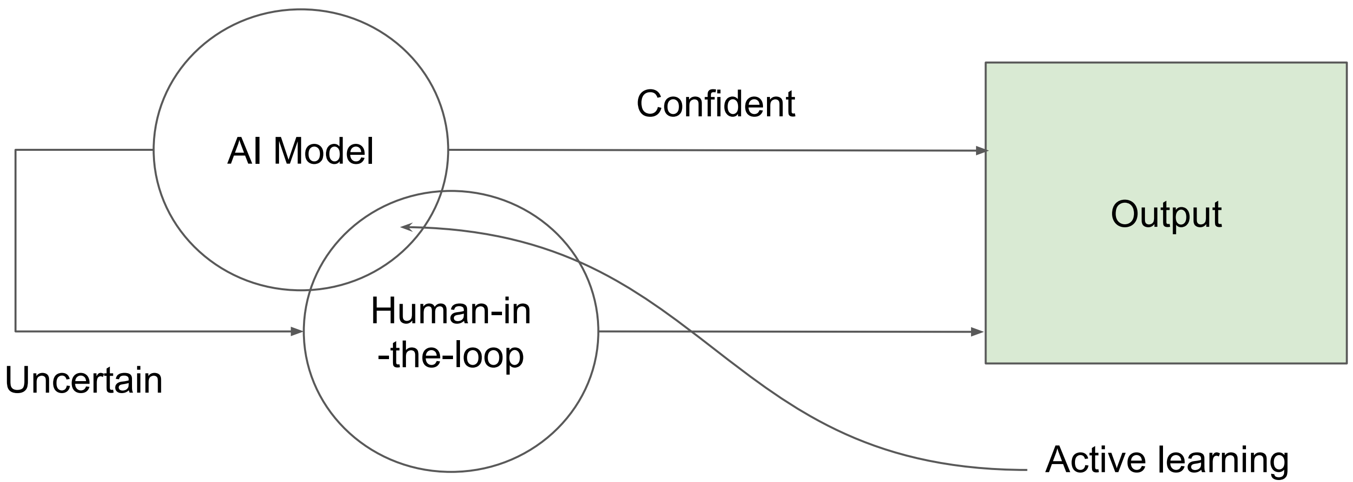
Annotate faster / cheaper

- Annotation UX
- “Machine” assisted annotation (e.g. pre-annotation via ML model, steam-like suggestions, etc)

Annotate smarter

- Select the most informative data for annotation

What does it have to do with active learning?

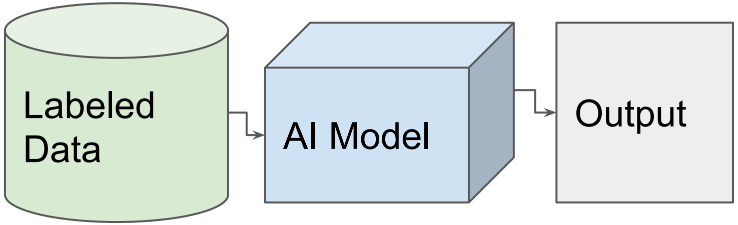


Active learning focuses on ML model / Human interactions. E.g. how to organize cooperation with human for

1. getting data
2. supporting when a model is not confident
3. etc

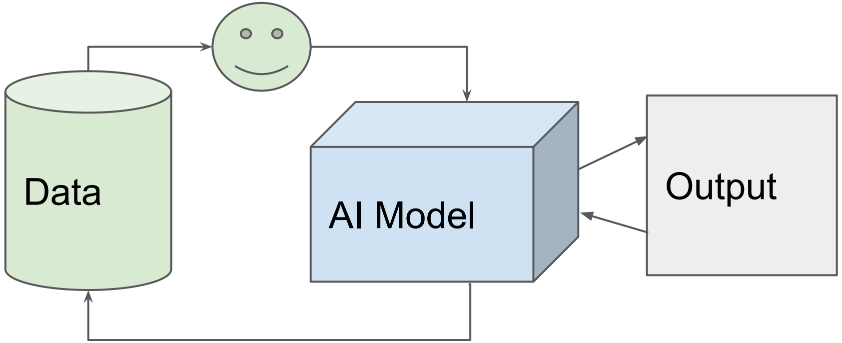
Active learning is not a model; it’s strategy / protocol

Supervised learning



Core of supervised learning is an ML model

Active learning



Core of active learning is a wrapper on top of ML model

Faster annotation

“Just add some machine learning in the annotation loop” they said...

Case study: Object detection annotation on a video

Data annotation flow:

Data annotator draws bounding boxes on **key frames** → interpolation in between → Bingo!

Faster annotation

Case study: Object detection annotation on a video

Modified data annotation flow:

Presentation with Huge Object detector → Data annotator **corrects** bounding boxes **if needed** on key frames → **tracker** → interpolation in between → **<your guess>**

Faster annotation

Case study: Object detection annotation on a video

Modified data annotation flow:

Pre-Annotation with Huge Object detector → Data annotator **corrects** bounding boxes **if needed** on key frames → interpolation in between → ~~**Bingo!**~~ **no bingo :-(**

Modified data annotation flow iteration 2:

Full replication of Data Annotator pipeline by ML → **Data annotator with usual flow** → **Bingo!**

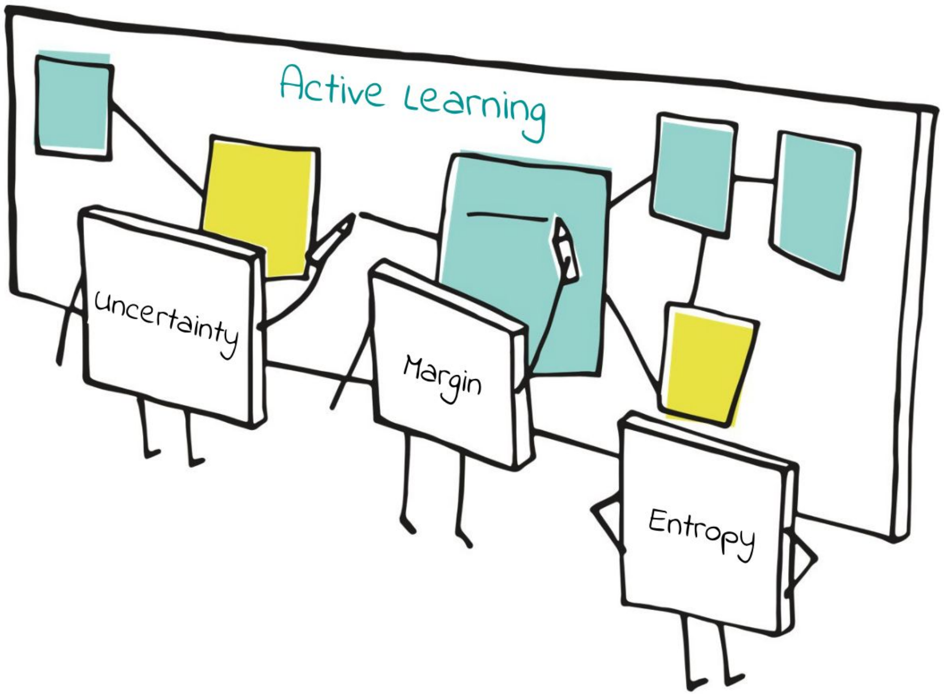
Faster annotation

Case study: Object detection annotation on a video

Lessons learned:

- Partial replacement of separate component may be not enough. Likely won’t be enough.
- Thinking about annotation UX may be more beneficial than thinking about the final result.
- Data annotation speed and statistics is more about people rather than numbers.

Smarter annotation



$$u(x) = 1 - p_1(x)$$

Uncertainty

$$m(x) = p_2(x) - p_1(x)$$

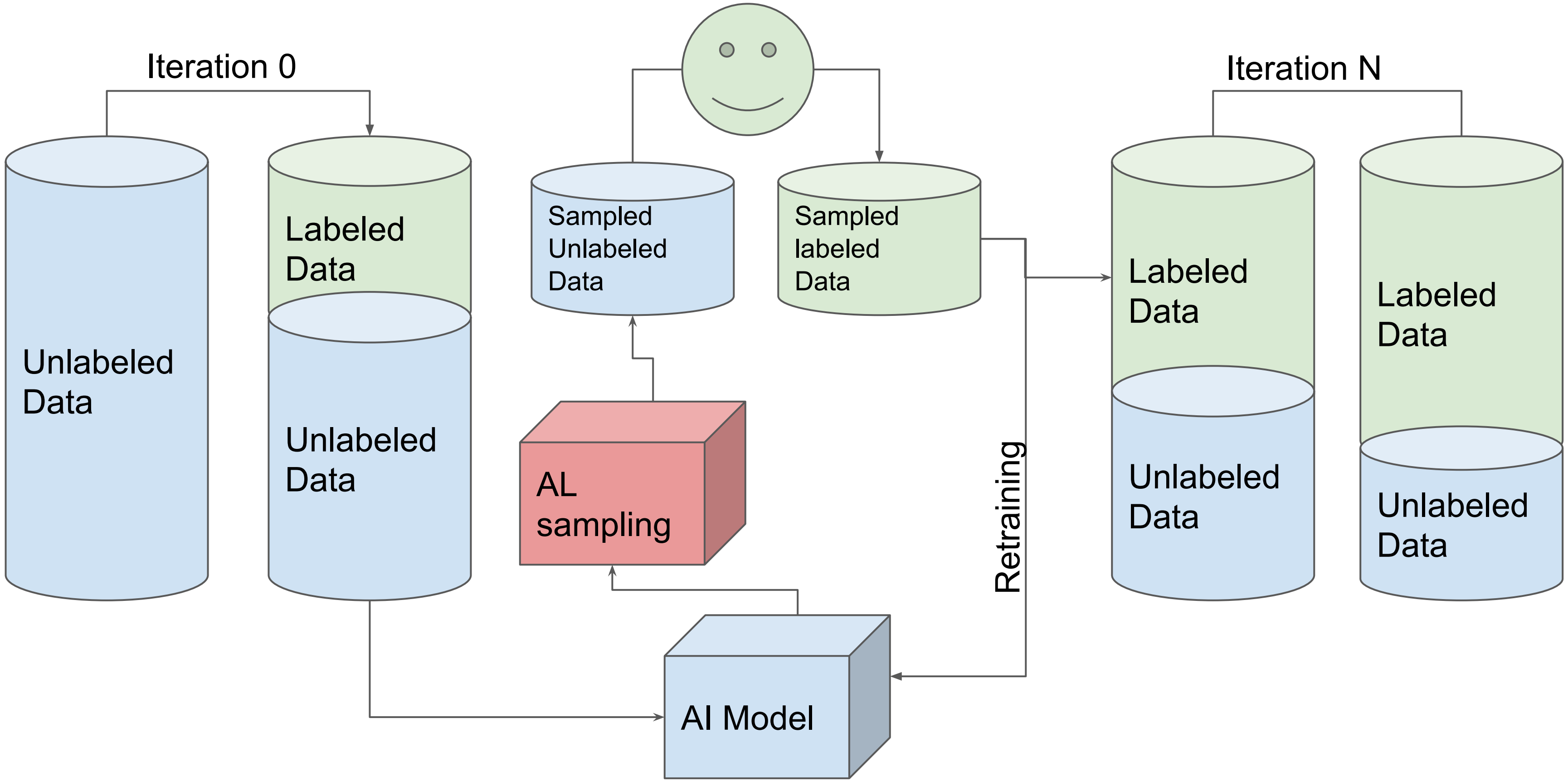
Margin

$$e(x) = \sum_{i=1}^{K} p_i(x) * \log p_i(x)$$

Entropy

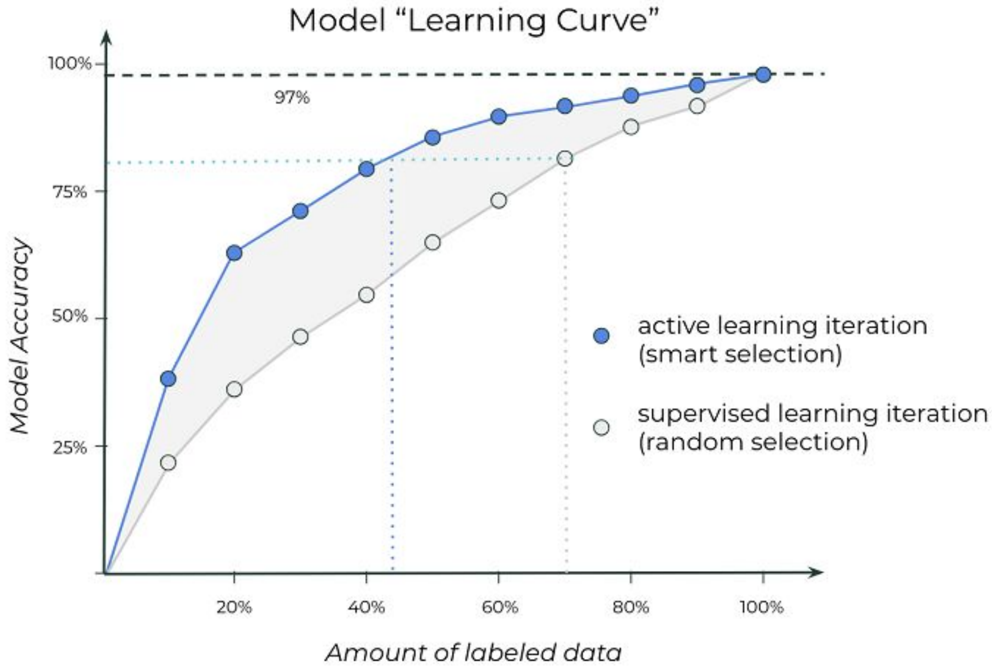
*https://towardsdatascience.com/learn-faster-with-smarter-data-labeling-15d0272614c4

Active learning step-by-step



Smarter annotation. What it gives to me?

- Reasonable result faster
- Better learning curves
- Helps with “long-tail”
- Likely outperform supervised learning



*https://www.kdnuggets.com/2018/10/introduction-active-learning.html

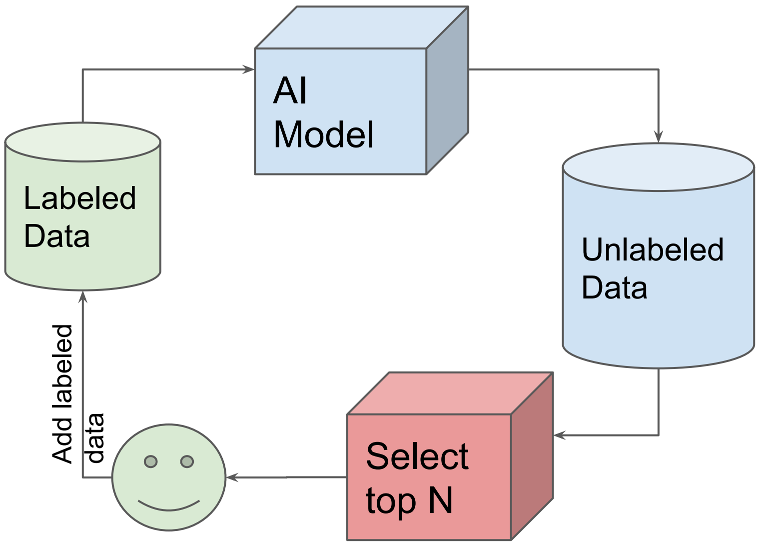
Some examples of sampling approaches

- Model uncertainty based: *criteria(x) = 1 - p(x)*
- Max class margin based: *criteria(x) = p(x | y = dog) - p(x | y = cat)*
- Entropy based: *criteria(x) = ∑p(x) * log p(x)*
- Information density: *criteria(x) = 1/N * ∑||f(x) - f(xj)||*
- etc

Captain note: you cannot just annotate data points that has the largest uncertainty.

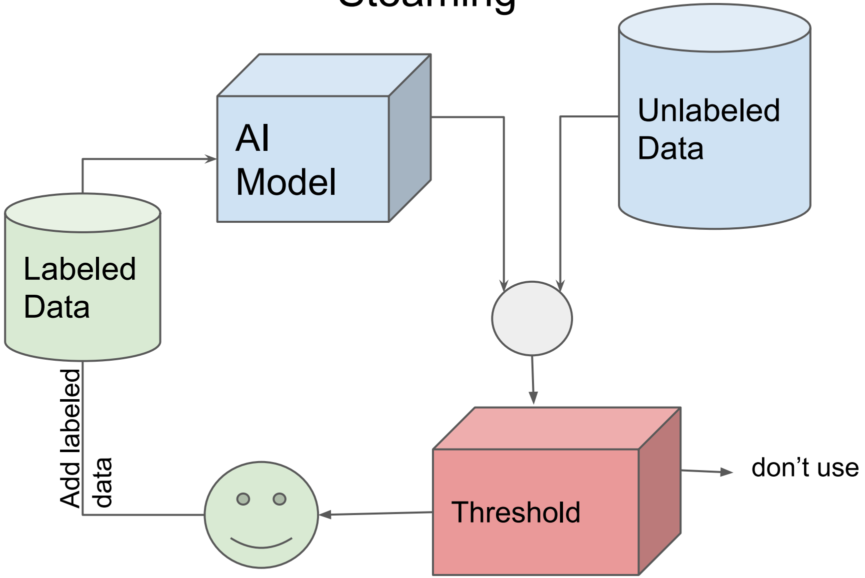
Active learning approaches

Pooling



+ Budget can be guaranteed
+ Real optimization
- all data need to be evaluated

Steaming



+ Lightweight
- budget not guaranteed

What can we tune? What are hyperparameters?

- Pooling, streaming or custom protocol?
- Sampling strategies
- Pool size for pooling and threshold for streaming
- When we want to stop

Problems and open questions

- Bias and fairness
- Easy to understand, hard to implement
- In real scenarios active learning pipelines ~~often~~ sometimes collapses to
  - finding incorrect labeled data
  - finding corner cases where even a human is highly unconfident

References

- https://www.kdnuggets.com/2018/10/introduction-active-learning.html
- https://towardsdatascience.com/learn-faster-with-smarter-data-labeling-15d0272614c4
- https://arxiv.org/abs/1801.05124
- http://parnec.nuaa.edu.cn/huangsj/alipy/
- https://www.youtube.com/watch?v=V33Ut36eUsY

Questions

We are hiring

https://grnh.se/a93b70881