Visual Recognition with Humans in the Loop

Authors: Steve Branson, Catherine Wah, Florian Schro, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie

Presenters: Qi Huang, Shuo Chen

Visual Recognition with Humans in the Loop

Authors: Steve Branson, Catherine Wah, Florian Schro, Boris Babenko, Peter Welinder, Pietro Perona, and Serge Belongie

Presenters: Qi Huang, Shuo Chen

Visipedia - Visual Encyclopedia

- Goal
  a. Creation of large-scale machine vision dataset
  b. Scalable representation of visual knowledge
  c. Embed interactive images with wiki articles
  d. Visual search

Visipedia - Visual Encyclopedia

- Goal
  a. Creation of large-scale machine vision dataset
  b. Scalable representation of visual knowledge
  c. Embed interactive images with wiki articles
  d. Visual search

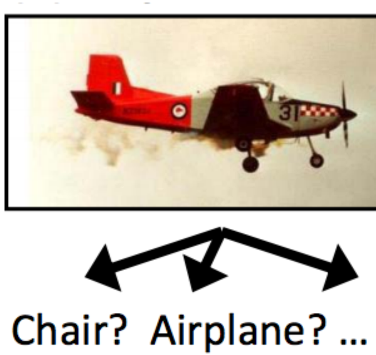
This work: fine-grained vision recognition / classification

Vision Recognition

- CV's Development
  - Good at inter-category classification (easy for human)
  - Bad at fine-grained classification (hard for human)

Vision Recognition

- CV’s Development
  - Good at inter-category classification (easy for human)
  - Bad at fine-grained classification (hard for human)

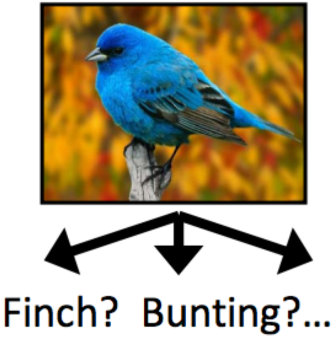


Easy for CV
Easy for human

Vision Recognition

- CV's Development
  - Good at inter-category classification (easy for human)
  - Bad at fine-grained classification (hard for human)

Easy for CV
Easy for human



Hard for CV
Hard for human

Vision Recognition

- Different Difficulties

Finch.

Bunting.

Sparrow.

Albatross.

Vision Recognition

- Different Difficulties

Human:

Lack of expertise, knowledge, memory.

Finch.

Bunting.

Sparrow.

Albatross.

Vision Recognition

- Different Difficulties

Human:
Lack of expertise, knowledge, memory.

Computer:
Lack of fundamental vision capabilities.

Finch.

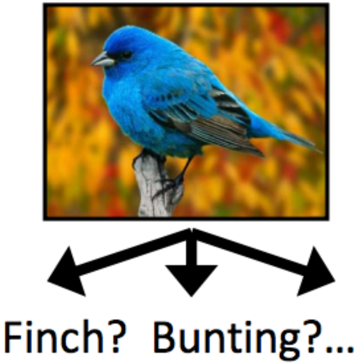
Bunting.

Sparrow.

Albatross.

Human in the loop

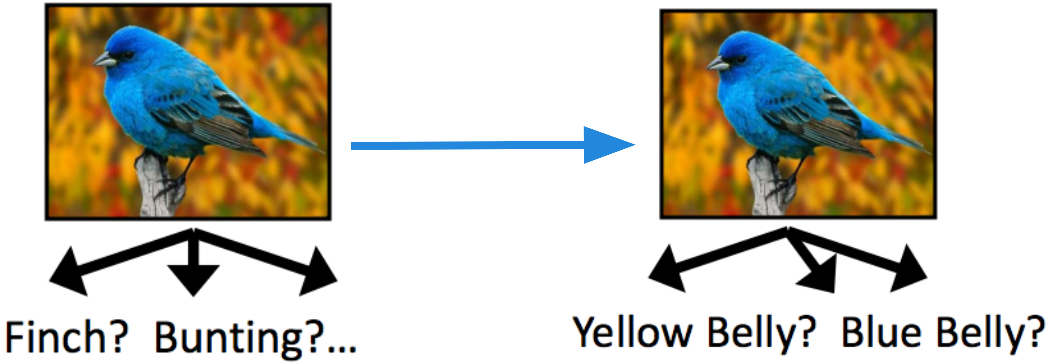
- Computer & Human contribute collaboratively



Hard for CV
Hard for human

Human in the loop

- Computer & Human contribute collaboratively

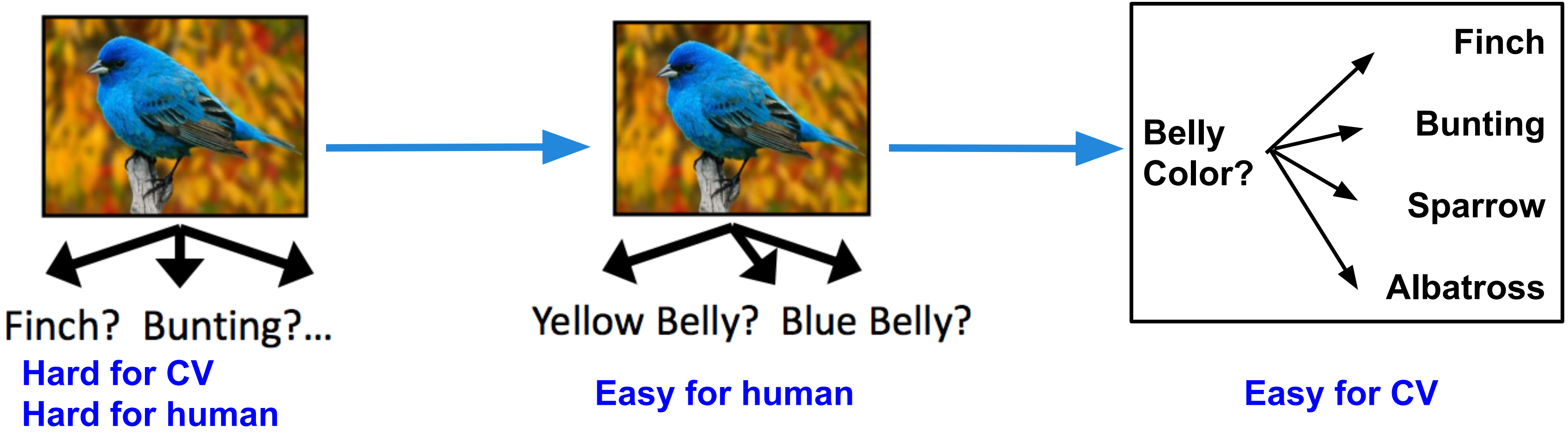


Hard for CV
Hard for human

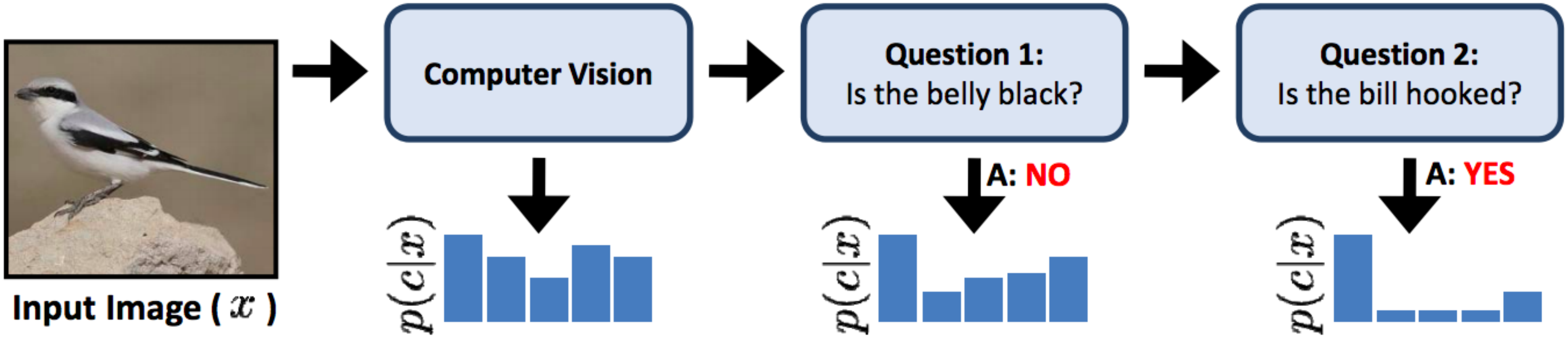
Easy for human

Human in the loop

- Computer & Human contribute collaboratively



Basic (Testing) Algorithm Flow



Three questions

1. How to incorporate computer vision?
2. How to pick the next question to ask?
3. How to update the posterior $p(c|x)$?

Three questions

1. How to incorporate computer vision?

2. How to pick the next question to ask?

3. How to update the posterior $p(c|x)$?

How to incorporate computer vision

- Basically plug in whatever you have
  - The authors used SIFT feature SVM classifier and another attribute-based classifier
- The point is to get a $p(c|x)$ before asking questions
- Doesn't even have to be a computer vision component

Three questions

1. How to incorporate computer vision?
2. How to pick the next question to ask?
3. How to update the posterior $p(c|x)$?

Some notations

- A set of possible questions $\mathcal{Q} = \{q_1 ... q_n\}$, (e. g. IsRed?, HasStripes?, BellyColor?)

- The response $u_i = (a_i, r_i)$ is an answer $a_i \in \mathcal{A}_i$, plus a confidence value
$r_i \in \mathcal{V}$, $(e.g.,\ \mathcal{V} = \{\text{Guessing, Probably, Definitely}\})$

Some notations (cont'd)

- At time step t
- Already have a response set

$$U^{t-1} = \{u_{j(1)} \ldots u_{j(t-1)}\}$$

- Pick a question $q_{j(t)}$ to ask

How to pick the next question?

- By maximizing information gain
- Just like a decision tree algorithm

$$\mathrm{I}(c; u_i | x, U^{t-1}) = \mathbb{E}_u \left[\mathrm{KL} \left(p(c|x, u_i \cup U^{t-1}) \parallel p(c|x, U^{t-1}) \right) \right]$$
$$= \sum_{u_i \in \mathcal{A}_i \times \mathcal{V}} \boxed{p(u_i|x, U^{t-1})} \left(\mathrm{H}(c|x, u_i \cup U^{t-1}) - \mathrm{H}(c|x, U^{t-1}) \right)$$

$$\mathrm{H}(c|x, U^{t-1}) = -\sum_{c=1}^{C} \boxed{p(c|x, U^{t-1})} \log p(c|x, U^{t-1})$$

Three questions

1. How to incorporate computer vision?

2. How to pick the next question to ask?

3. How to update the posterior $p(c|x)$?

How to update the posterior

- Bayesian rule

$$\boxed{p(c|x,U)} = \frac{p(U|c,x)p(c|x)}{Z} = \frac{\boxed{p(U|c)}p(c|x)}{Z}$$

- An assumption is made here

$$p(U|c,x) = p(U|c)$$

Terms that we need to compute

- Another assumption

$$\boxed{p(U^{t-1}|c)} = \prod_{i}^{t-1} \boxed{p(u_i|c)}$$

$$\boxed{p(u_i|x, U^{t-1})} = \sum_{c=1}^{C} \boxed{p(u_i|c)}\boxed{p(c|x, U^{t-1})}$$

What we need to compute

$$
\begin{tikzcd}
 & \boxed{p(u_i|x, U^{t-1})} & \\
 & & \boxed{p(c|x, U^{t-1})} \arrow[ul] \\
\boxed{p(u_i|c)} \arrow[uur] \arrow[r] & \boxed{p(U|c)} \arrow[ur] &
\end{tikzcd}
$$

Terms that we still need to compute

$$\boxed{p(u_i|c)} = p(\underset{\text{answer}}{a_i}, \underset{\text{confidence value}}{r_i}|c) = p(a_i|r_i, c)p(r_i|c)$$

- Yet another assumption

$$p(r_i|c) = p(r_i)$$

- Get all these numbers from training/counting

Discussion about the assumptions

1 $$p(U|c, x) = p(U|c)$$

2 $$p(U^{t-1}|c) = \prod_{i}^{t-1} p(u_i|c)$$

3 $$p(r_i|c) = p(r_i)$$

Dataset & Question selection

- Bird-200
  - 6033 images over 200 bird species
  - Hard to be identified by non-experts
- Questions extracted from whatbird.com
  - 25 question set, encompass 288 binary attributes
  - Class-attribute is “deterministic”

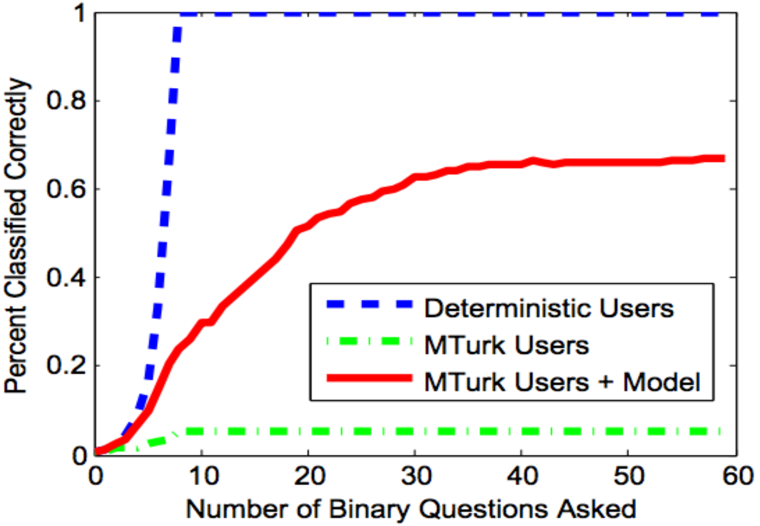
Answer collection

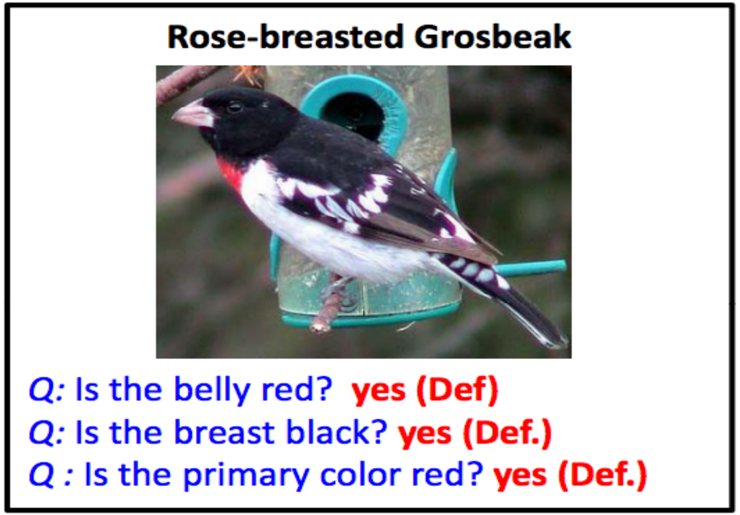
- Mechanical Turk Interface
  - Collect non-expert answers.
  - Use prototypical image.
  - Use random answer for eval.

Evaluation

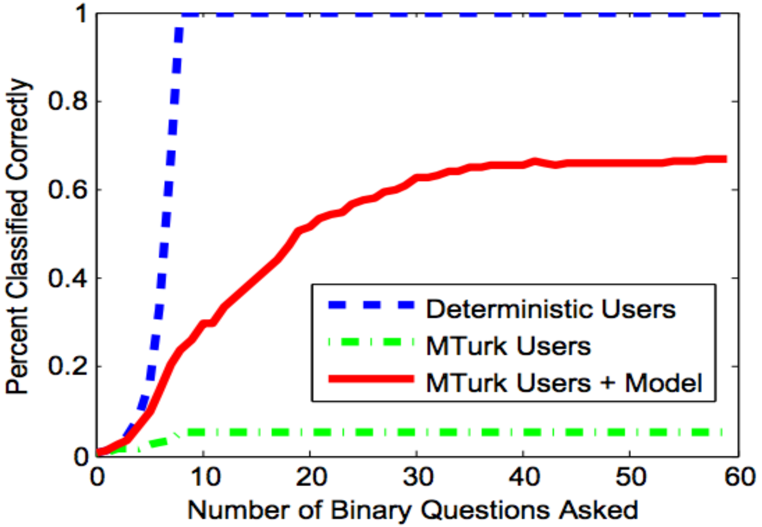
- Two cases:
  - Without CV
  - With CV (1-vs-all SVM, Attributes classifier)
- Two methods:
  - Ask exactly T questions, measure correct ratio (%)
  - Early termination, measure average # of questions

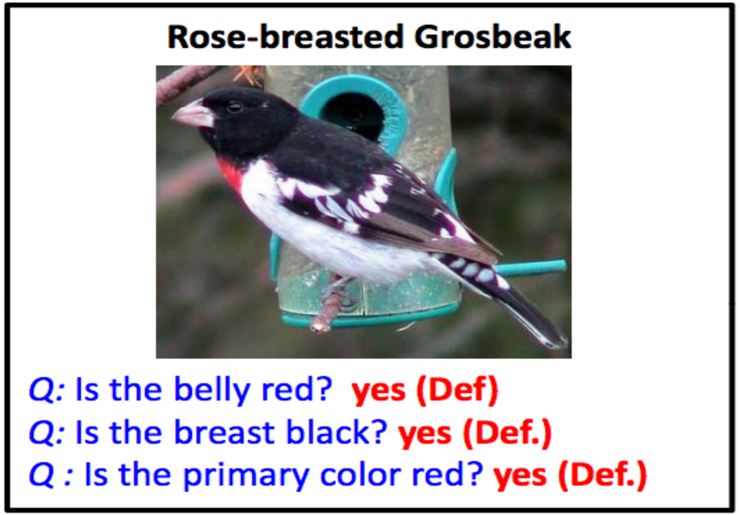
Modeling User Response (Method 1)





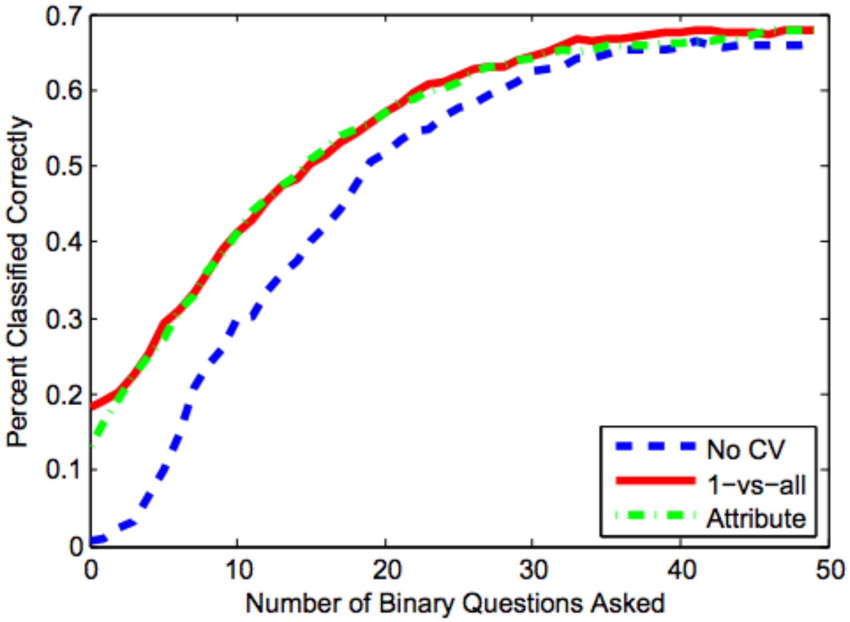
Modeling User Response (Method 1)





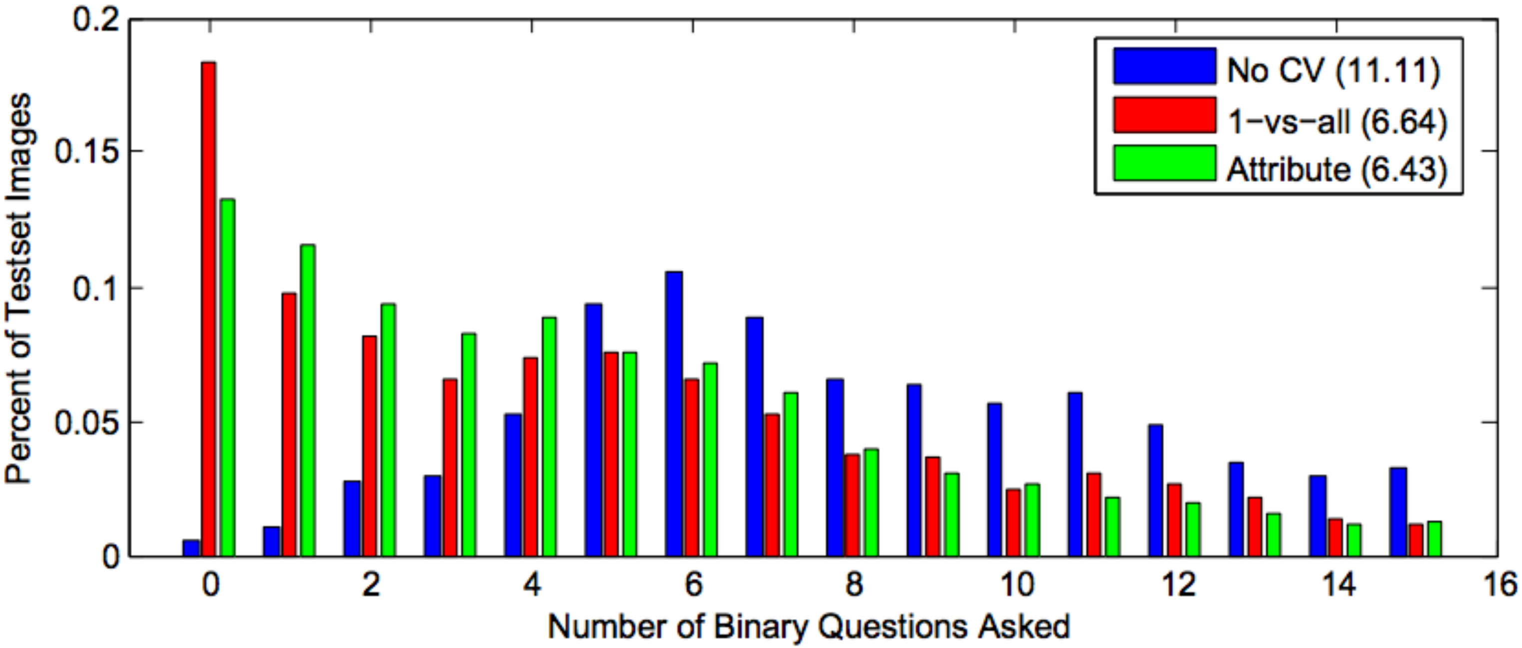
- Non-expert responses need modeling
- Much human effort is still needed for a usable service

Benefit of CV (Method 1)



CV achieves higher accuracy w/ less questions

Where does CV help (Method 2)



CV helps most for easy classification tasks

Where does CV help? (A case)

- W/o vision:

  HasShapePerchingLike

  has largest information gain

- W/ vision:

  HasThroatColorWhite

  likelihood of classes change

Western Grebe

w/ vision:

Q #1: Is the throat white? **yes (Def.)**

w/o vision:

Q #1: Is the shape perching-like? **no (Def.)**

Contribution

- A platform incorporates CV and human recognition
  - Flexible for a variety of CV algorithms
- [Hard question] human input drives up the performance
  - Stochastic model makes user response reliable
  - Needs further work on question picking
- [Easy question] CV can efficiently reduce human labor
  - Possibly human-aware CV can work better

Thank You!

The algorithm

Algorithm 1 Visual 20 Questions Game

1: $U^0 \leftarrow \emptyset$
2: **for** $t = 1$ to 20 **do**
3: $\quad j(t) = \max_k \mathrm{I}(c; u_k | x, U^{t-1})$
4: $\quad$ Ask user question $q_{j(t)}$, and $U^t \leftarrow U^{t-1} \cup u_{j(t)}$.
5: **end for**
6: Return class $c^* = \max_c p(c|x, U^t)$

Trade-off

- (+) Provide a practical service to collect data
  - Minimize human efforts -> exclude in the future
  - Pluggable platform to test CV algorithms

- (-) Selection of questions are tricky to the performance
  - Rely on already gained experts’ knowledge