Computational Biomedicine Imaging & Modeling Center

Write a Classifier: Zero Shot Learning Using Purely Textual Descriptions

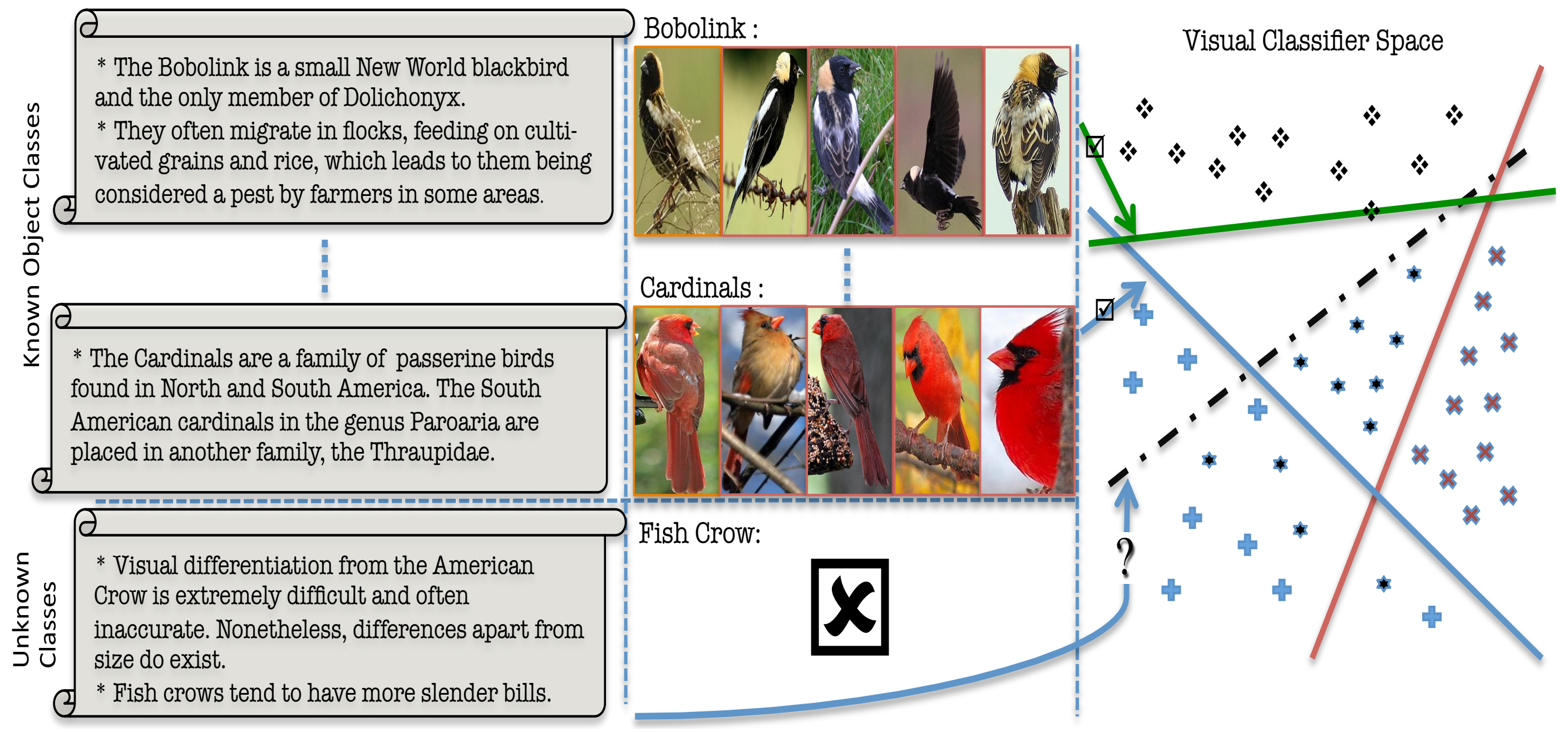
Mohamed Elhoseiny*, Babak Saleh, Ahmed Elgammal
Department of Computer Science, Rutgers University

Motivation

- One of the main challenges for scaling up object recognition systems is the lack of annotated images for real-world categories. Typically there are few images available for training classifiers for most of these categories; severe for fine-grained categorization.
- There are abundant of textual descriptions of these categories, which comes in the form of dictionary entries, encyclopedia articles, and various online resources. For example, it is possible to find several good descriptions of a "bobolink" in encyclopedias of birds, while there are only a few images available for that bird online.
- Attributes based approaches for zero shot learning deals with the dilemma of finding best set of visual attributes for object description.

Problem Definition

- The main question we address in this paper is how to use purely textual description of categories with no training images to learn visual classifiers for these categories.
- We propose an approach for zero-shot learning of object categories where the description of unseen categories comes in the form of typical text such as an encyclopedia entry, without the need to explicitly defined attributes.



* The Bobolink is a small New World blackbird and the only member of Dolichonyx.
* They often migrate in flocks, feeding on cultivated grains and rice, which leads to them being considered a pest by farmers in some areas.

* The Cardinals are a family of passerine birds found in North and South America. The South American cardinals in the genus Paroaria are placed in another family, the Thraupidae.

* Visual differentiation from the American Crow is extremely difficult and often inaccurate. Nonetheless, differences apart from size do exist.
* Fish crows tend to have more slender bills.

Contributions

- First approach that predicts explicit visual classifier parameters of unseen classes from typical text, as encyclopedia entry.
- Two baselines were designed based on Regression and Domain Adaptation (DA) functions.
- We proposed a quadratic program that involves both Regression and DA functions.

Project Website: https://sites.google.com/site/mhelhoseiny/projects/computer-vision-projects/Write_a_Classifier
Includes the data. The code will be available shortly on it.

Formulations

We denote textual features and visual features as **t** ∈ $\mathcal{T}$ (textual domain) and **x** ∈ $\mathcal{V}$ (visual domain), respectively. In all of our experiments, **t** is tf-idf features and **x** is classeme features.

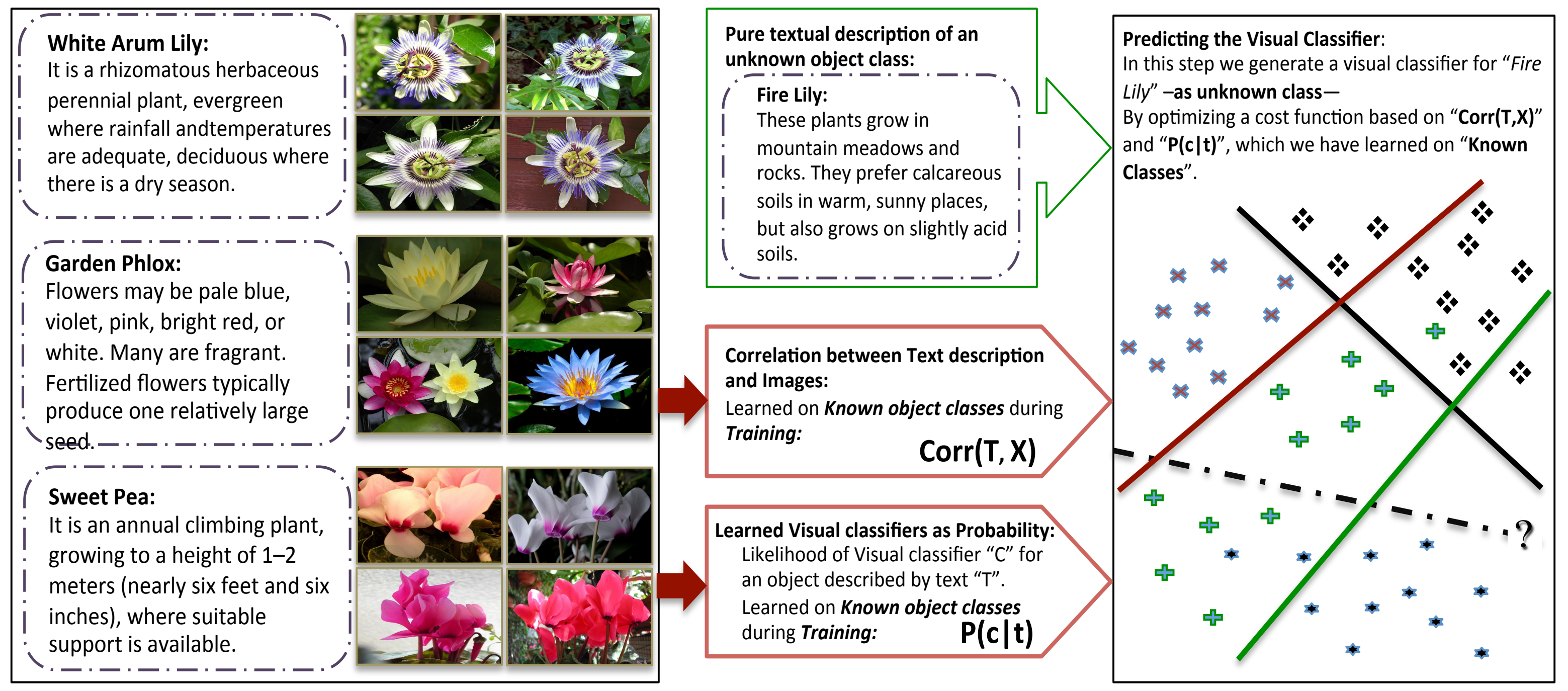
1) Regression (Reg) Baseline

- A set of one-vs-all classifiers $\{\mathbf{c}_k\}$ are learned, one for each seen class. Given $\{(\mathbf{t}_k;\ \mathbf{c}_k)\}$, a regressor is learned that can be used to give a prior estimate for $p_{reg}(\mathbf{c}|\mathbf{t})$. In our experiments, we used Gaussian Process Regression (GPR) and Twin Gaussian Processes (TGP).
- The predicted classifier for textual feature vector $\mathbf{t}_*$ is obtained as $c_{reg}(\mathbf{t}_*) = \arg\max_c\ [p_{reg}(\mathbf{c}|\mathbf{t}_*)]$.

2) Domain Adaptation (DA) Baseline

- A a linear (or nonlinear kernalized) transfer function **W** between $\mathcal{T}$ and $\mathcal{V}$.
- W can be learned by optimizing, with a suitable regularizer, over constraints of the form $\mathbf{t}^T \mathbf{W}\, \mathbf{x} > l$ if **t** and **x** belong to the same class, and $\mathbf{t}^T \mathbf{W}\, \mathbf{x} < u$ otherwise, where **x** is a visual feature vector amended by 1, l and u are model parameters.
- It is not hard to see that this transfer function can act as a classifier. Given a textual feature **t** and a test image, represented by **x** , a classification decision can be obtained by $\mathbf{t}^T \mathbf{W}\, \mathbf{x} > b$, we set b to (l + u)/2 .
- The predicted classifier is obtained as $c_{DA}(\mathbf{t}_*) = \mathbf{t}_*^T \mathbf{W}$.

3) DA-Reg Quadratic Program (Better than 1 and 2)



White Arum Lily:
It is a rhizomatous herbaceous perennial plant, evergreen where rainfall andtemperatures are adequate, deciduous where there is a dry season.

Garden Phlox:
Flowers may be pale blue, violet, pink, bright red, or white. Many are fragrant. Fertilized flowers typically produce one relatively large seed.

Sweet Pea:
It is an annual climbing plant, growing to a height of 1–2 meters (nearly six feet and six inches), where suitable support is available.

Pure textual description of an unknown object class:

Fire Lily:
These plants grow in mountain meadows and rocks. They prefer calcareous soils in warm, sunny places, but also grows on slightly acid soils.

Correlation between Text description and Images:
Learned on ***Known object classes*** during ***Training:*** Corr(T, X)

Learned Visual classifiers as Probability:
Likelihood of Visual classifier "C" for an object described by text "T".
Learned on ***Known object classes*** during ***Training:*** P(c|t)

Predicting the Visual Classifier:
In this step we generate a visual classifier for "*Fire Lily*" **–as unknown class–**
By optimizing a cost function based on "**Corr(T,X)**" and "**P(c|t)**", which we have learned on "**Known Classes**".

Predicted Classifier

$$\hat{c}(\mathbf{t}_*) = \underset{\mathbf{c},\zeta_i}{\operatorname{argmin}} \Big[\mathbf{c}^\mathsf{T}\mathbf{c} - \alpha \mathbf{t}_*{}^\mathsf{T}\mathbf{W}\mathbf{c} - \beta \ln(p_{reg}(\mathbf{c}|\mathbf{t}_*)) + \gamma \sum \zeta_i\Big]$$

$$s.t.: -(\mathbf{c}^\mathsf{T}\mathbf{x}_i) \geq \zeta_i,\ \ \zeta_i \geq 0,\ \ i = 1 \cdots N$$

$$\mathbf{t}_*{}^\mathsf{T}\mathbf{W}\mathbf{c} \geq l$$

$$\alpha, \beta, \gamma, l : \text{hyperparameters}$$

- $\{\mathbf{x}_i$, i=1:N$\}$ are the visual features of the images of the seen classes.
- Given ln p_{reg} **(c|t)** from the TGP and **W** , this equation reduces to a quadratic program on **c** with linear constraints

Experimental Results

- We provide Text Augmentation of the CUB-Birds and Oxford-Flower datasets.
- We computed the ROC curve and report the area under that curve (AUC) as a comparative measure of different approaches.

Table 1: Comparative Evaluation on the Flowers and Birds Datasets

Approach	Flowers Avg AUC (+/- std)	Birds Avg AUC (+/- std)
GPR	0.54 (+/- 0.02)	0.52 (+/- 0.001)
TGP	0.58 (+/- 0.02)	0.61 (+/- 0.02)
DA	0.62(+/- 0.03)	0.59 (+/- 0.01)
Our Approach	**0.68** (+/- 0.01)	**0.62** (+/- 0.02)

1) Flower Dataset

Table 2: Percentage of classes that the proposed approach makes an improvement in predicting over the baselines (relative to the total number of classes in each dataset

baseline	Flowers (102) % improvement	Birds (200) % improvement
GPR	100 %	98.31 %
TGP	66 %	51.81 %
DA	54%	56.5%

Table 3: Top-5 classes with highest combined improvement in Flower dataset

class	TGP (AUC)	DA (AUC)	Our (AUC)	% Improv.
2	0.51	0.55	0.83	57%
28	0.52	0.54	0.76	43.5%
26	0.54	0.53	0.76	41.7%
81	0.52	0.82	0.87	37%
37	0.72	0.53	0.83	35.7 %

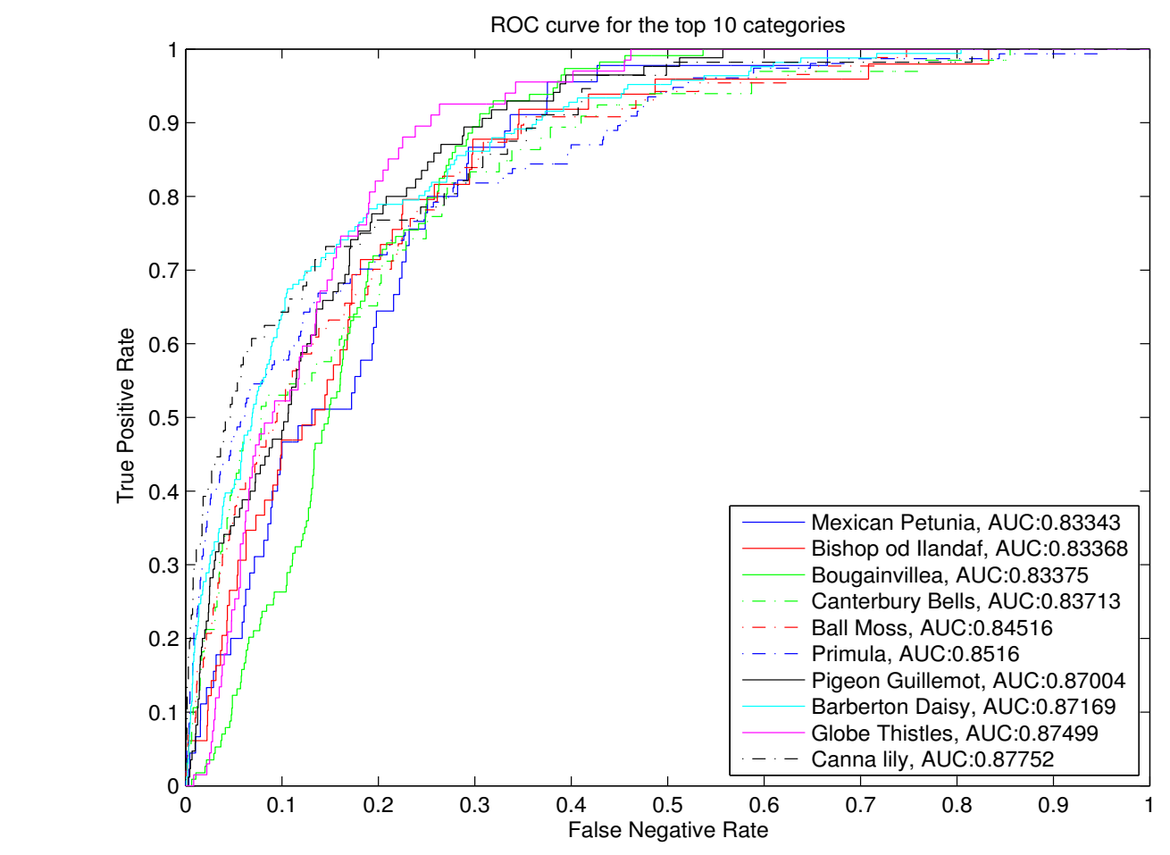


Fig 1: ROC curves of best 10 predicted classes (best seen in color) for Flower Dataset

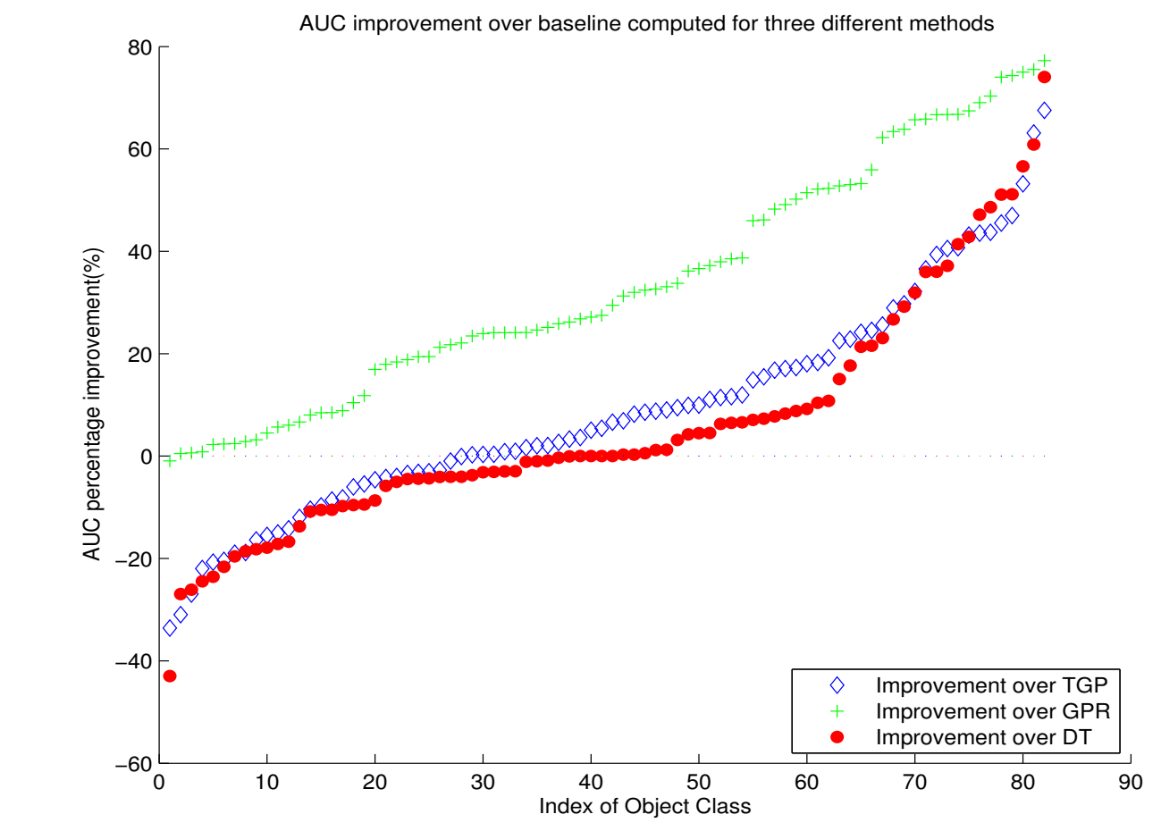


Fig 2: AUC improvement over the three baselines (GPR, TGP, DA) on Flower dataset.

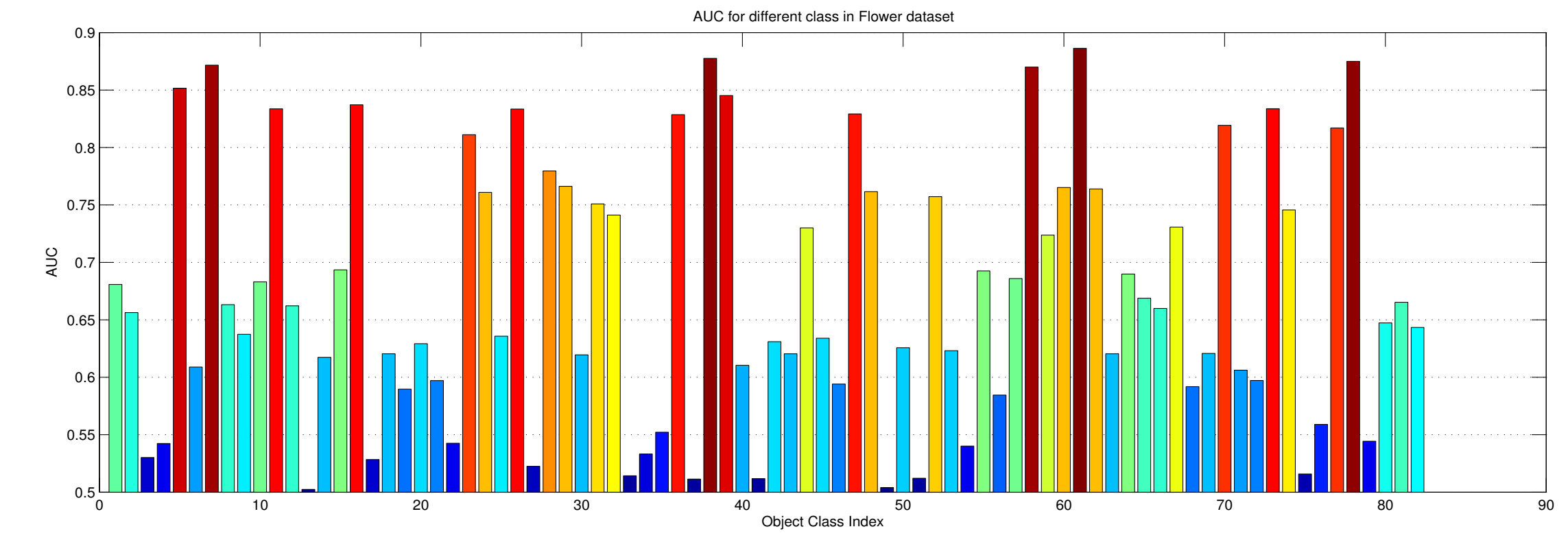


Fig 3: AUC of the predicated classifiers for all classes of Flower dataset

2) Birds Dataset

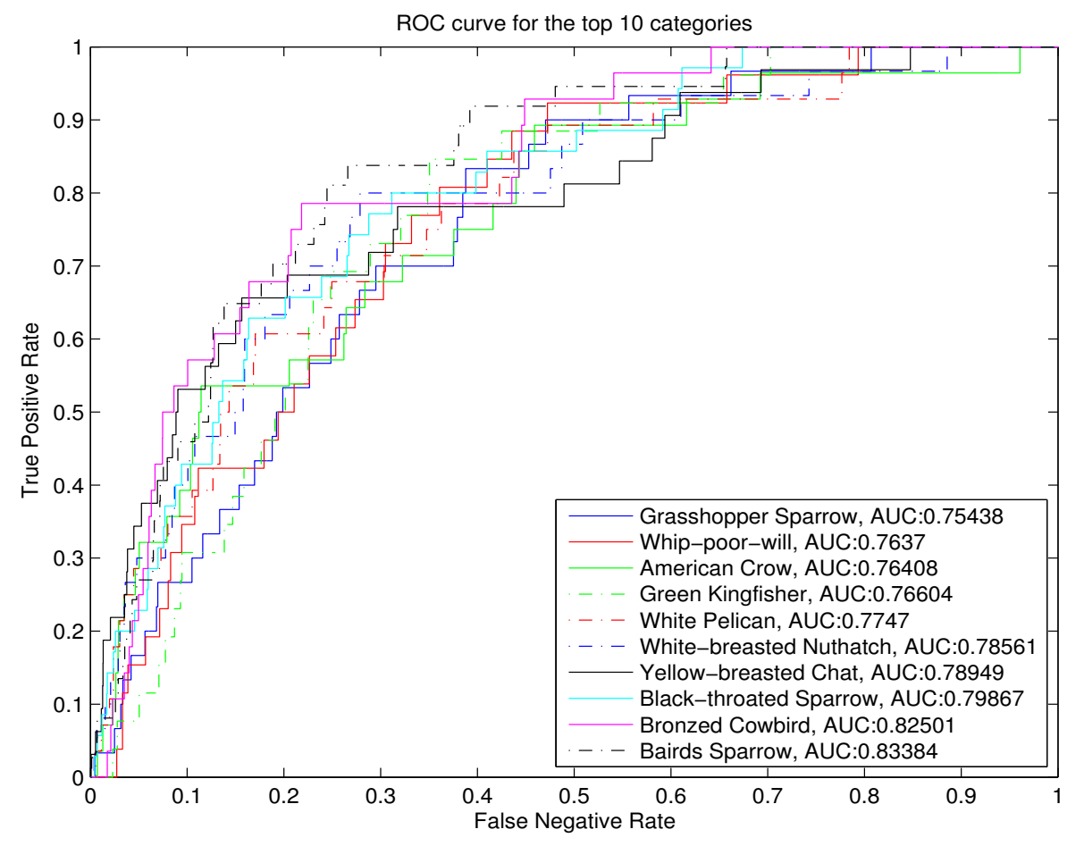


Fig 4: ROC curves of best 10 predicted classes (best seen in color) for Bird dataset

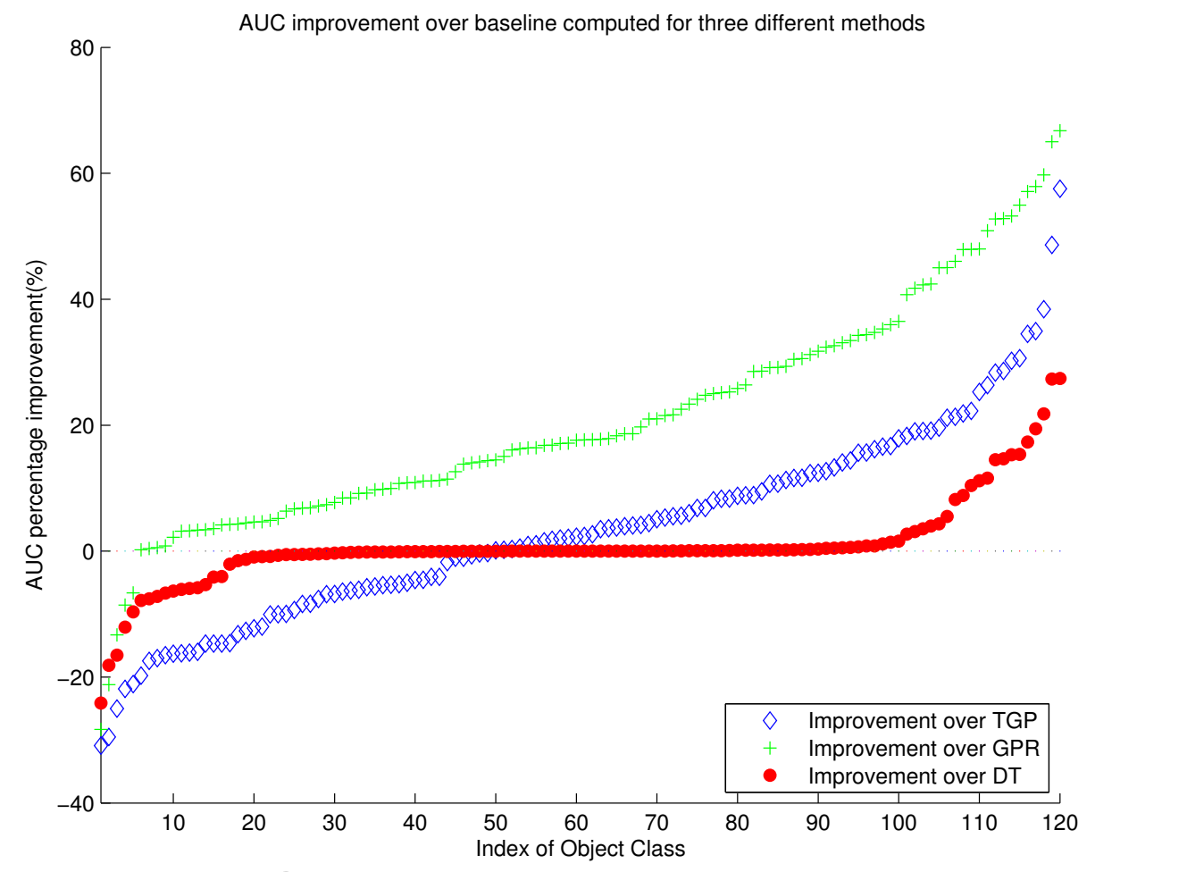


Fig 5: AUC improvement over the three baselines (GPR, TGP, DA) on Birds dataset.

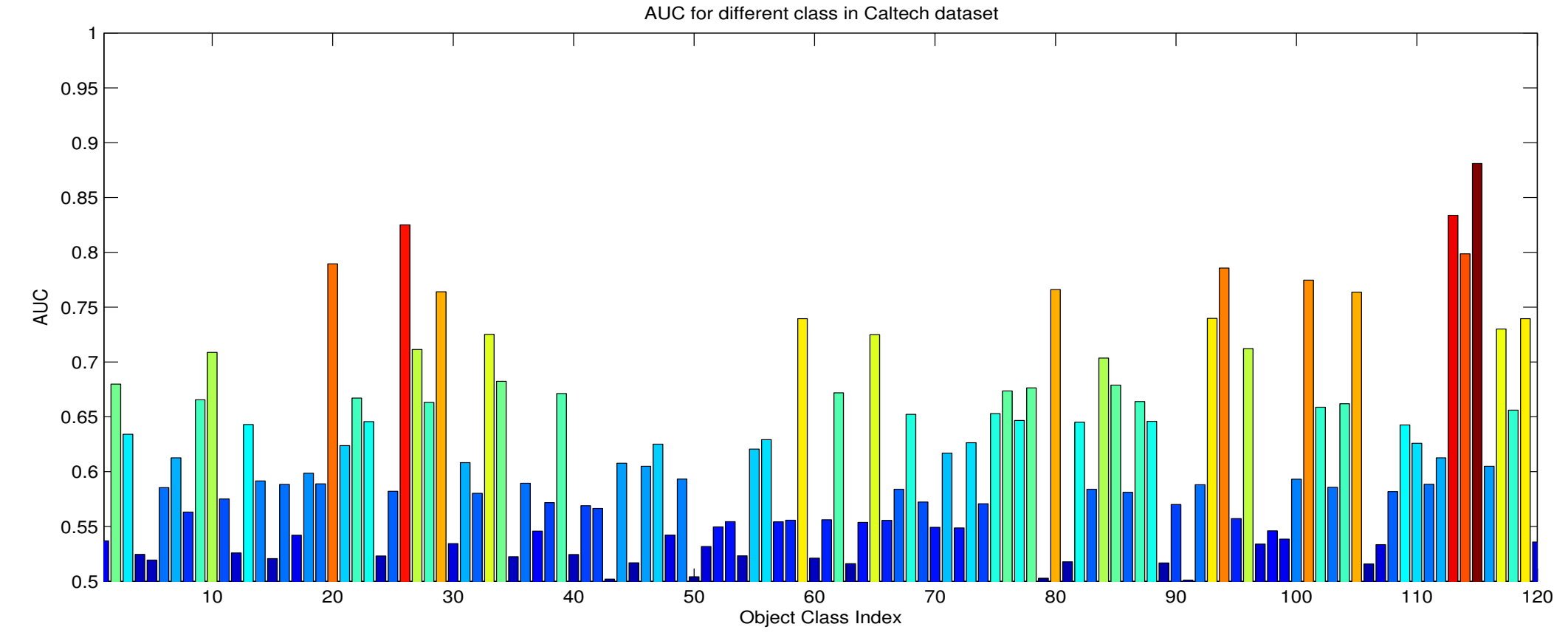


Fig 6: AUC of the predicated classifiers for all classes of Birds dataset