Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions

Jimmy Lei Ba, Kevin Swersky, Sanja Fidler, Ruslan Salakhutdinov

ICCV 2015

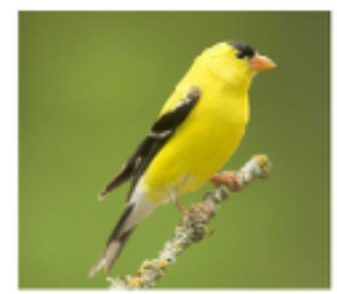
Presenter: Fartash Faghri

Zero-shot Learning

- Classify images of an unseen class given semantically or visually similar classes at training time.

- Shared knowledge between classes can be given in various forms, such as attributes or class descriptions.

American Goldfinch



Intuitive!

Attribute	Has?
Beak longer than head	✗
Solid yellow belly	✓
Black and white wings	✓
⋮	⋮

Antol et al. [1]

Contributions

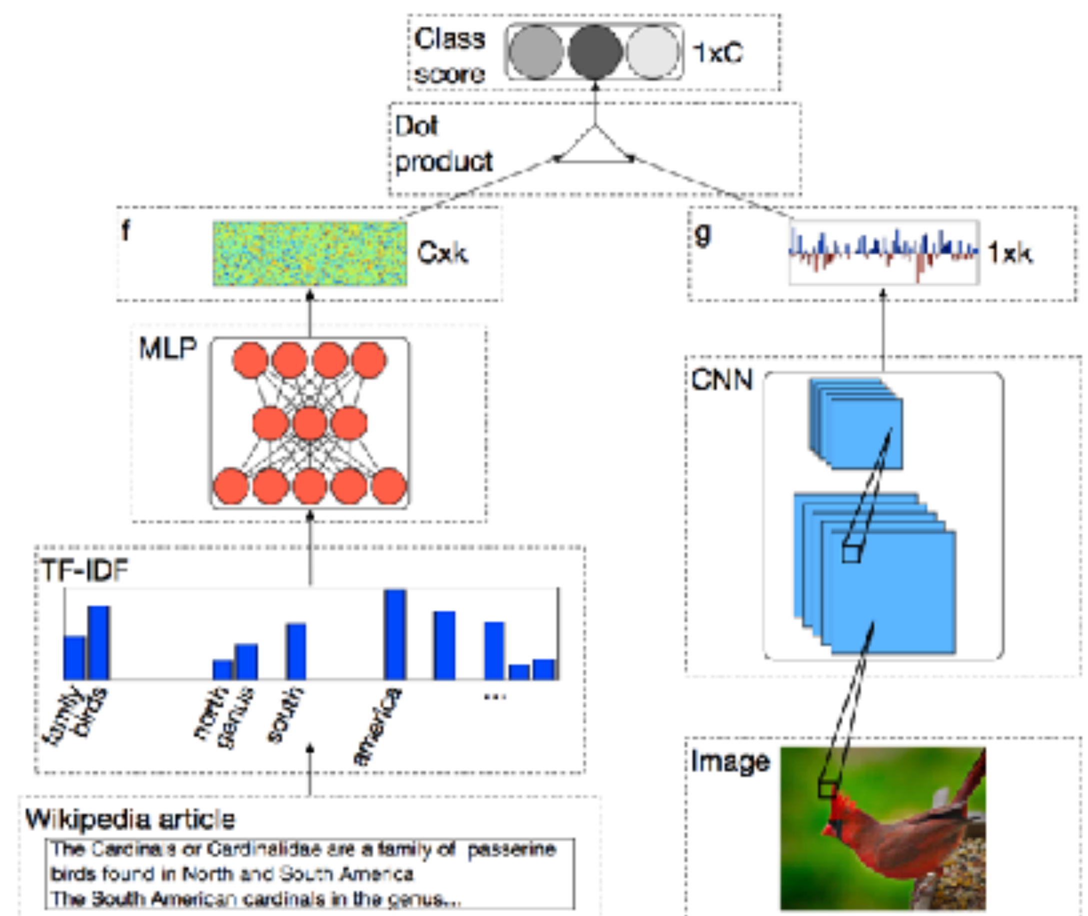
- The main contribution is the convolutional classifier. The rest of the contributions are shared with [2].

- Predicts visual classes using <u>text corpus</u>, in particular, the <u>encyclopedia corpus</u>. This overcomes the difficulty of hand-crafted attributes.

- The key difference with the most related work is that image and text features are transformed into a joint embedding space.

Classifier

- Image feature vectors: $x \in R^d$
- Text feature vectors: $t_c \in \mathbb{R}^p$
- A linear classifier: $\hat{y}_c = w_c^\top g_v(x),$
- Image transformation: $g_v : \mathbb{R}^d \mapsto \mathbb{R}^k$
- Text transformation: $f_t : \mathbb{R}^p \mapsto \mathbb{R}^k$



Convolutional Classifier

- Text can describe attributes (low) or objects (high).
- Classifier on fully connected features: $\hat{y}_c = w_c^\top g_v(x),$
- Classifier on convolutional features: $\hat{y'}_c = o\left(\sum_{i=1}^{K'} w'_{c,i} \check{*} a'_i\right),$
- Joint classifier: $\hat{y}_c = w_c^T g_v(x) + o\left(\sum_{i=1}^{K'} w'_{c,i} \check{*} g'_v(a)_i\right).$
- $o(\cdot)$ is a global pooling function.

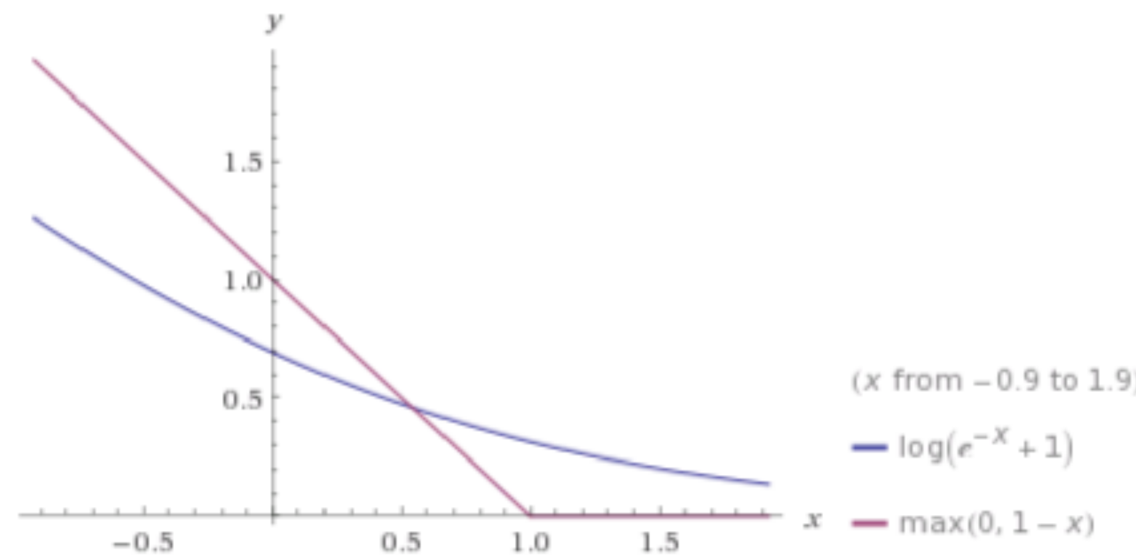
Learning

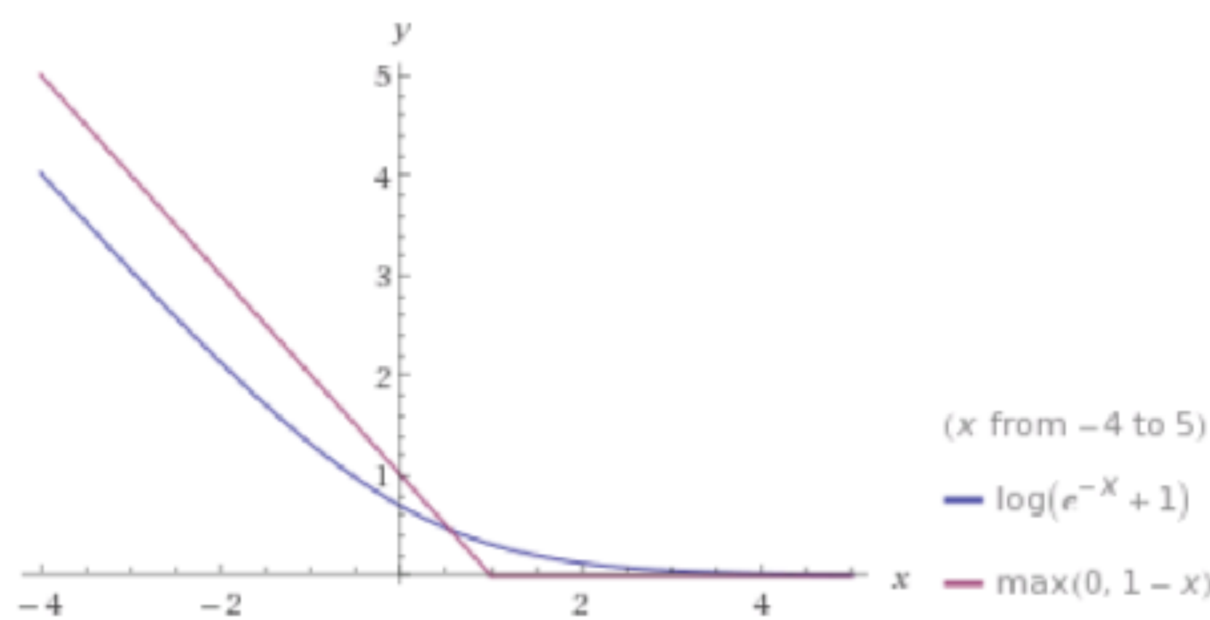
- Binary Cross Entropy: $$\mathcal{L}(W) = \sum_{i=1}^{N}\sum_{j=1}^{C}\Big[I_{i,j}\log\sigma(\hat{y}_j(x_i,t_j)) + (1-I_{i,j})\log(1-\sigma(\hat{y}_j(x_i,t_j)))\Big],$$

  where σ is the sigmoid function $y = 1/(1+e^{-x})$.

- Hinge Loss: $$\mathcal{L}(W) = \sum_{i=1}^{N}\sum_{j=1}^{C}\max(0, \epsilon - I_{i,j}\hat{y}_j(x_i,t_j)).$$
- Euclidean Distance between $g_v(x)$ and $f_t(t_c)$

Loss Comparison





Produced by WolframAlpha

Experiments

- DA: the model is similar to the hinge loss form
- DA+GP: in that model multiple text descriptions can be given for a class, GP part gives $p(c|t)$, a prior.
- fc baseline feat.: features from [2], HOG, GIST, etc
- ROC: true positive rate vs false positive rate

Results

Dataset	Model	ROC-AUC unseen	ROC-AUC seen	ROC-AUC mean	PR-AUC unseen	PR-AUC seen	PR-AUC mean
CU-Bird200-2010	DA (baseline feat.) [5]	0.59	—	—	—	—	—
	DA+GP [5] (baseline feat.)	0.62	—	—	—	—	—
	DA [15] (VGG feat.)	0.66	0.69	0.68	0.037	0.11	0.094
	Ours (fc baseline feat.)	0.69	0.93	0.85	0.09	0.20	0.19
	Ours (fc)	**0.82**	0.96	0.934	**0.10**	0.41	0.35
	Ours (conv)	0.73	0.96	0.91	0.043	0.34	0.28
	Ours (fc+conv)	0.80	**0.987**	**0.95**	0.08	**0.53**	**0.43**
CU-Bird200-2011	Ours (fc)	0.82	0.974	0.943	0.11	0.33	0.286
	Ours (conv)	0.80	0.96	0.925	0.085	0.15	0.14
	Ours (fc+conv)	**0.85**	**0.98**	**0.953**	**0.13**	**0.37**	**0.31**
Oxford Flower	DA (baseline feat.) [5]	0.62	—	—	—	—	—
	GPR+DA (baseline feat.) [5]	0.68	—	—	—	—	—
	Ours (fc baseline feat.)	0.63	0.96	0.86	0.055	0.60	0.45
	Ours (fc)	0.70	0.987	0.90	**0.07**	0.65	0.52
	Ours (conv)	0.65	0.97	0.85	0.054	0.61	0.46
	Ours (fc+conv)	**0.71**	**0.989**	**0.93**	0.067	**0.69**	**0.56**

Table 1. ROC-AUC and PR-AUC(AP) performance compared to other methods. The performance is shown for both the zero-shot unseen classes and test data of the seen training classes. The class averaged mean AUCs are also included. For both ROC-AUC and PR-AUC, we report the best numbers obtained among the models trained on different objective functions.

Results (cont.)

Metrics	BCE	Hinge	Euclidean
unseen ROC-AUC	**0.82**	0.795	0.70
seen ROC-AUC	**0.973**	0.97	0.95
mean ROC-AUC	**0.937**	0.934	0.90
unseen PR-AUC	**0.103**	0.10	0.076
seen PR-AUC	0.33	**0.41**	0.37
mean PR-AUC	0.287	**0.35**	0.31
unseen class acc.	0.01	0.006	**0.12**
seen class acc.	0.35	**0.43**	0.263
mean class acc.	0.17	**0.205**	0.19
unseen top-5 acc.	0.176	0.182	**0.428**
seen top-5 acc.	0.58	**0.668**	0.45
mean top-5 acc.	0.38	0.41	**0.44**

Table 2. Model performance using various objective functions on CUB-200-2010 dataset. The numbers are reported by training the fully-connected models.

Metrics	Conv5_3	Conv4_3	Pool5
mean ROC-AUC	**0.91**	0.6	0.82
mean PR-AUC	**0.28**	0.09	0.173
mean top-5 acc.	**0.25**	0.153	0.02

Table 3. Performance comparison using different intermediate ConvLayers from VGG net on CUB-200-2010 dataset. The numbers are reported by training the joint fc+conv models.

Model / Dataset	CUB-2010	CUB-2011	OxFlower
Ours (fc)	0.60	0.64	0.73
Ours(fc+conv)	**0.62**	**0.66**	**0.77**

Table 4. Performance of our model trained on the full dataset, a 50/50 split is used for each class.

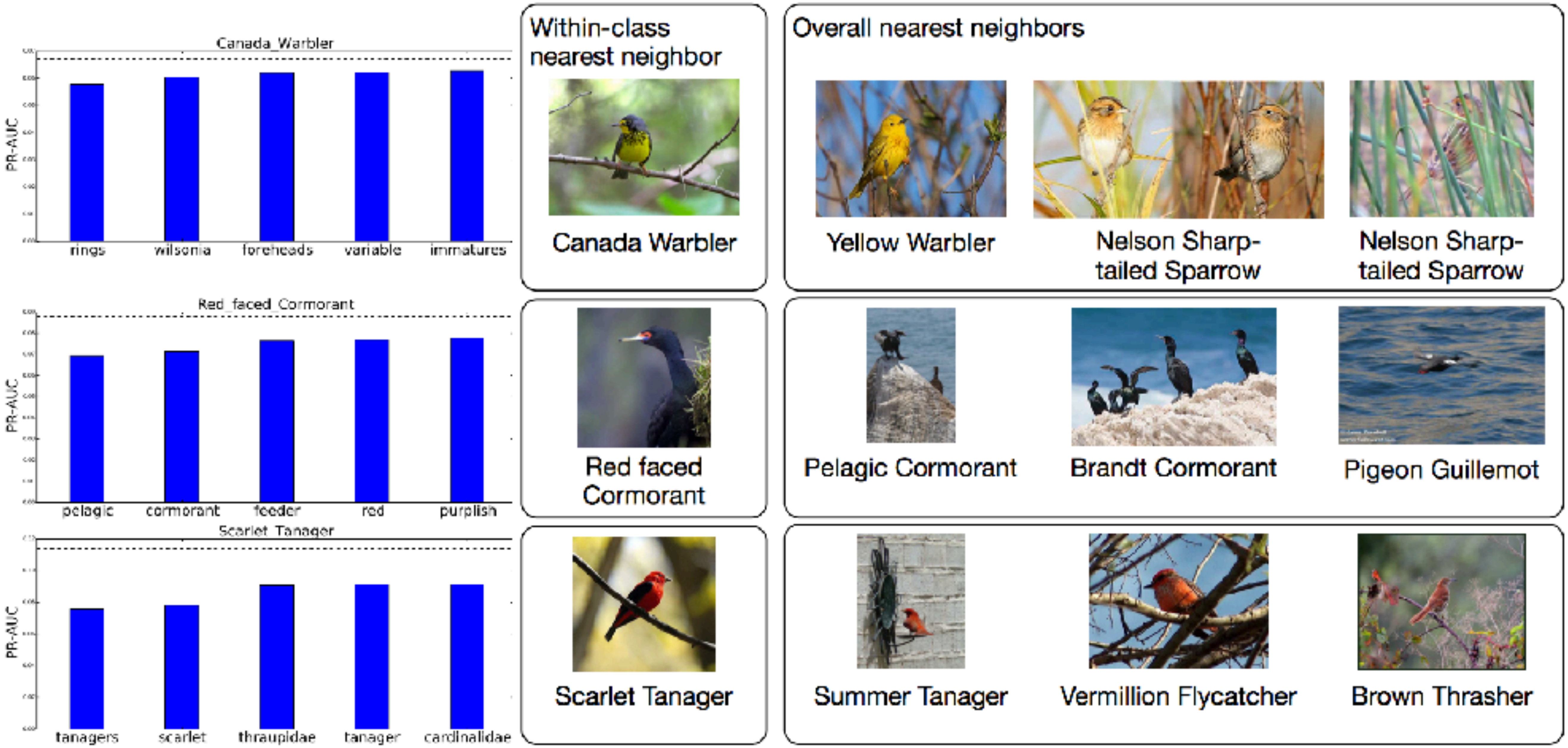


Figure 2. [LEFT]: Word sensitivities of unseen classes using the fc model on CUB200-2010. The dashed lines correspond to the test-set PR-AUC for each class. TF-IDF entries are then independently set to 0 and the five words that most reduce the PR-AUC are shown in each bar chart. Approximately speaking, these words can be considered to be important attributes for these classes. [RIGHT]: The Wikipedia article for each class is projected onto its feature vector w and the nearest image neighbors from the test-set (in terms of maximal dot product) are shown. The within-class nearest neighbors only consider images of the same class, while the overall nearest neighbors considers all test-set images.

References

- [1] Antol, Stanislaw, C. Lawrence Zitnick, and Devi Parikh. "Zero-shot learning via visual abstraction." European Conference on Computer Vision. Springer International Publishing, 2014.

- [2] Elhoseiny, Mohamed, Babak Saleh, and Ahmed Elgammal. "Write a classifier: Zero-shot learning using purely textual descriptions." Proceedings of the IEEE International Conference on Computer Vision. 2013.