ACM INTERNATIONAL CONFERENCE ON
MULTIMEDIA RETRIEVAL (ICMR 2021)

Text-Enhanced Attribute-Based Attention for Generalized Zero-Shot Fine-Grained Image Classification

Yan-He Chen and Mei-Chen Yeh

Department of Computer Science and Information Engineering

National Taiwan Normal University

Generalized zero-shot learning

Training set (seen)

fox

$x \in X$

elephant

Test set (seen & unseen)

fox

$z \in \{X, X'\}$

horse

elephant

lion

$X \cap X' = \emptyset$

Fine-grained image classification

Small inter-class variations

Artic Tern

Caspian Tern

Fine-grained image classification

Large intra-class variations

Laysan albatross: Adult

Laysan albatross: Juvenile

Attributes for zero-shot learning

Contributions of the paper

- For each class, a classification model should focus on the most relevant attributes.
- We propose to leverage auxiliary information in the form of ***text descriptions*** to achieve the goal.

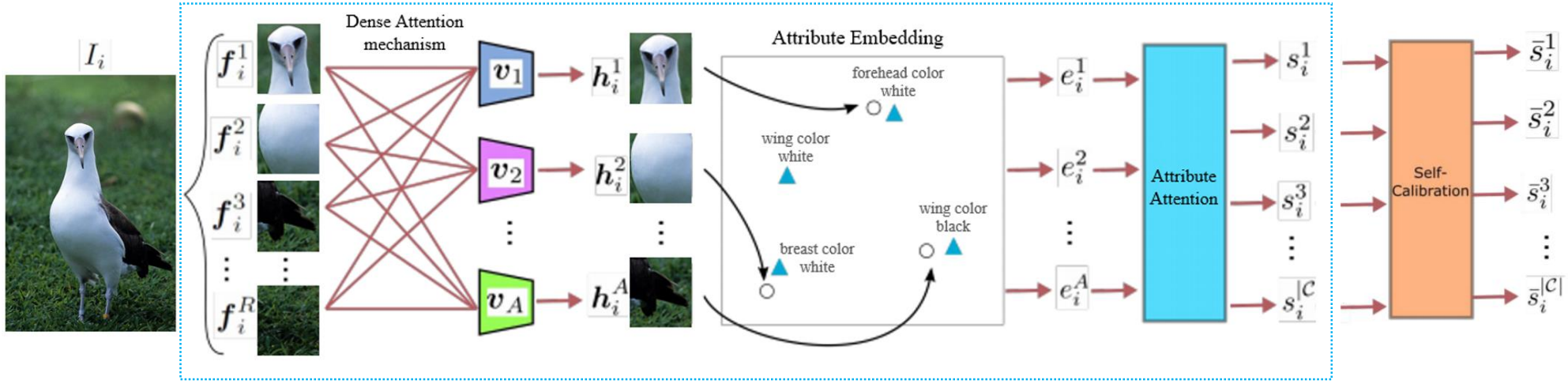
Adult

Very Large seabird with narrow, pointed wings and distinctive looping fight style; rarely flaps wings. Underwings are patchy white with variable amounts of dark. Grayish smudge on face.

Juvenile

Chicks are wispy brown with pale or blond highlights and a dark bill. Laysan Albatrosses nest among short vegetation, mainly in the remote Northwest Hawaiian Islands; also small colonies on Kauai and Oahu and on a few Mexican Islands.

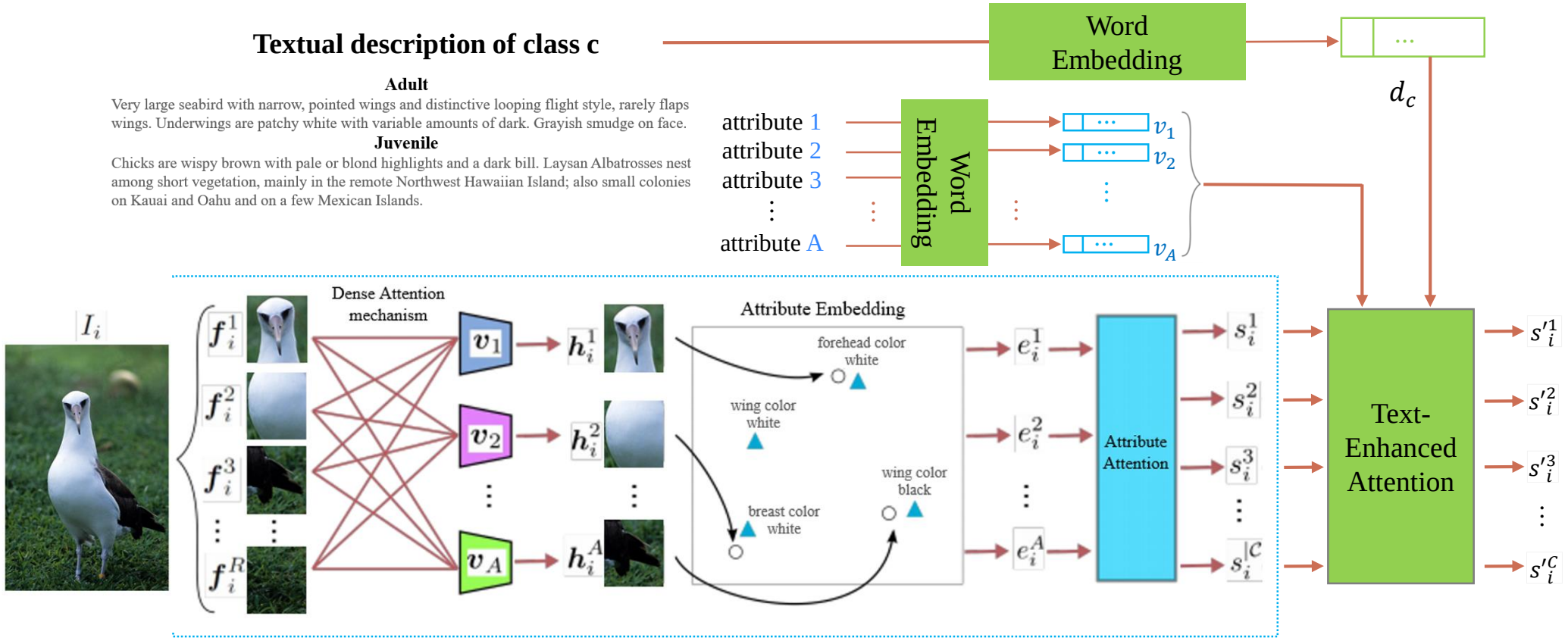
Text-enhanced attribute-based attention



DAZLE

Dat Huynh and Ehsan Elhamifar, “Fine-grained generalized zero-shot learning via dense attribute-based attention,” *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2020.

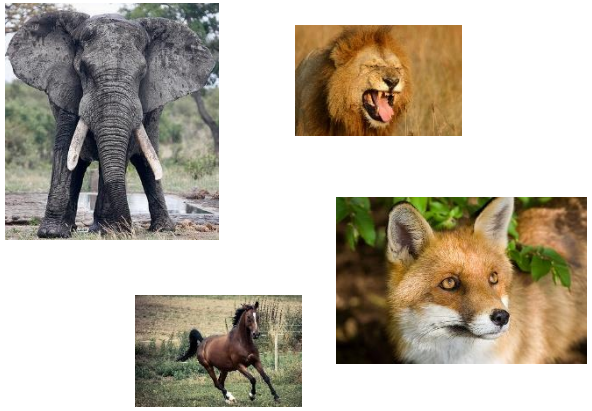
Text-enhanced attribute-based attention



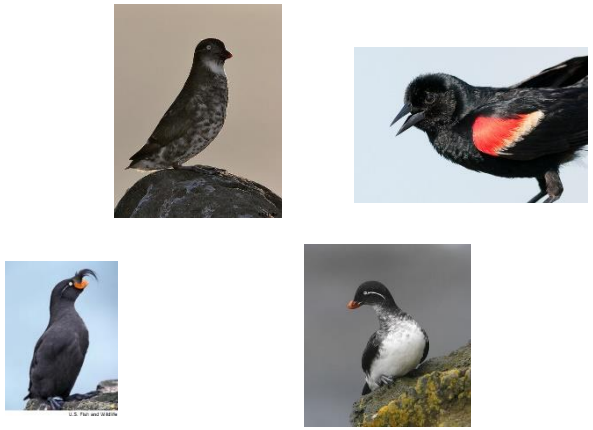
Dat Huynh and Ehsan Elhamifar, "Fine-grained generalized zero-shot learning via dense attribute-based attention," *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2020.

Datasets

Dataset	attributes	seen (val) / unseen classes	training / testing samples
CUB	312	100 (50) / 50	7,057 / 4,731
AWA2	85	27 (13) / 10	23,527 / 13,795



AWA2
(Coarse-grained)



CUB
(Fine-grained)

Results

Method	CUB			AWA2		
	acc_s	acc_u	H	acc_s	acc_u	H
DAZLE (dense attention)	57.6	42.4	48.8	72.1	46.8	56.8
Ours (text-enhanced attention)	**60.9**	**46.5**	**52.6**	**72.8**	**59.9**	**65.8**

- Our approach improves DAZLE in both datasets.
  - On CUB, the accuracy rates of recognizing seen and unseen classes are boosted by 2.4% and 4.1%, respectively. The harmonic mean is increased by 3.8%.
  - A performance gap for the unseen classes on AWA2 is also observed. The harmonic mean is increased by 9%.

Conclusions

- We improve DAZLE by devising a ***text-enhanced*** attribute-based attention mechanism that guides the feature extraction from most relevant image regions for important attributes.
- By leveraging ***textual descriptions***, the proposed method adapts well to unseen classes at inference.

More information:
http://www.csie.ntnu.edu.tw/~myeh