ZSCRGAN: A GAN based Expectation Maximization Model for Zero-Shot Retrieval of Images from Textual Descriptions

Anurag Roy[1], Vinay Kumar Verma[2], Kripabandhu Ghosh[3], Saptarshi Ghosh[1]

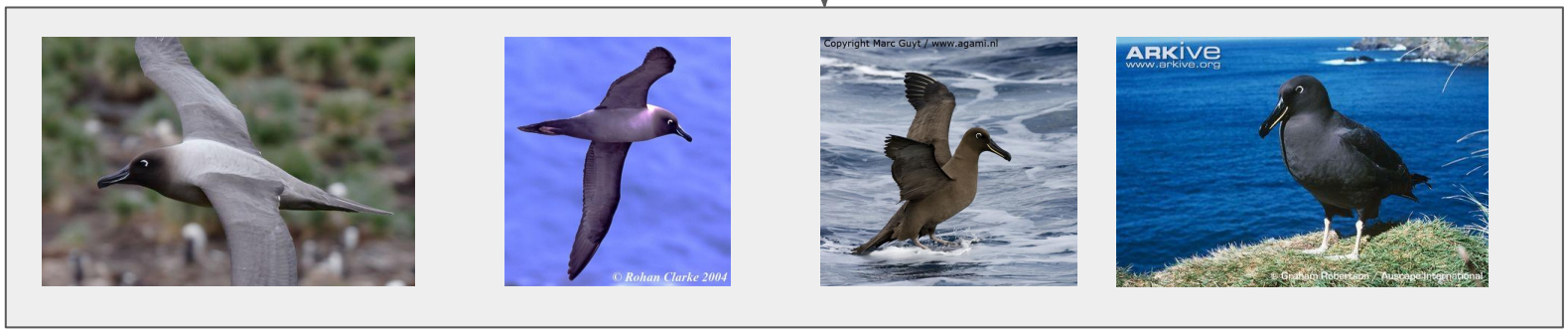
[1]Indian Institute of Technology Kharagpur, India,
[2]Duke University, Durham, North Carolina, USA
[3]Indian Institute of Science Education and Research (IISER) Kolkata, India

CIKM 2020, Full Paper

Background: Text-to-Image Retrieval

this is a grey bird with a black beak and a white eye.

query



retrieved images

Sooty Albatross

Problem Description:

Given a descriptive text as query, retrieve images that satisfy this description.

Text and Image source: CUB dataset

Background: Zero-Shot Learning Setup

A machine learning setup where the trained model has to predict **novel classes unseen** during training.

Tiger

Horse

Training Set

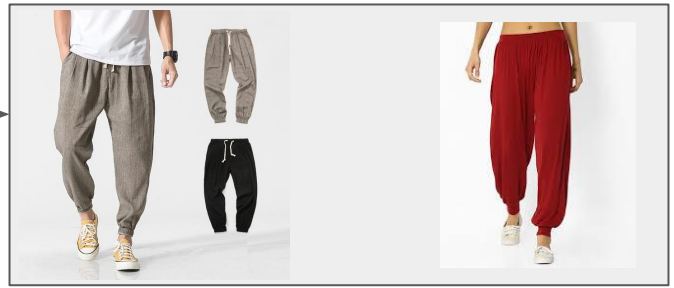
?

- Attributes of the classes are mapped with the features
- Attributes are shared across classes, both seen and unseen.

Image source: Wikipedia

Motivation for Text-to-Image Retrieval

baggy, long pants caught in at the ankle



trousers that become wider from the knees downward, forming a bell-like shape of the trouser leg.

Many a times, we forget the exact name of the product. Instead we remember the product description.

Image source: Google

Motivation for Zero-Shot Text-to-Image Retrieval

- With frequent launch of new products (classes) along with variations of old ones, it is impractical to re-train the existing models every time a new product is launched

- Requires a lot of time and manual effort to collect and label data for training

Prior Works on Zero-shot Text-to-Image Retrieval

There have been a few prior works in this field:

- DS-SJE[1]: Jointly learns the image and text embeddings using discriminative models by optimizing a joint embedding loss function

- ZSL-GAN[2]: A GAN is trained to generate image embeddings from text embeddings similar to the respective class by using Wasserstein loss along with classification loss

- DADN[3]: Uses Dual GANs to project image & text embeddings to a joint embedding space; uses class label embeddings to learn overlap among classes

[1] A. Reed, Z. Akata, B. Schiele, and H. Lee. 2016. “Learning deep representations of fine-grained visual descriptions”. In Proc. IEEE CVPR.
[2] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. “A Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts”. In Proc. IEEE CVPR.
[3] Jingze Chi and Yuxin Peng. 2019. “Zero-shot Cross-media Embedding Learning with Dual Adversarial Distribution Network”. IEEE TCSVT.

Limitations of prior works

- DS-SJE[1]: Discriminative models used for text and image embeddings; lack visual imaginative capability
- ZSL-GAN[2]: The classification loss in the discriminator may lower the discriminative power of the discriminator
- DADN[3]: Performance is limited by the quality of class embeddings used

1] A. Reed, Z. Akata, B. Schiele, and H. Lee. 2016. “Learning deep representations of fine-grained visual descriptions”. In Proc. IEEE CVPR.
[2] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. “A Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts”. In Proc. IEEE CVPR.
[3] Jingze Chi and Yuxin Peng. 2019. “Zero-shot Cross-media Embedding Learning with Dual Adversarial Distribution Network”. IEEE TCSVT.

Our Contribution

Zero-Shot Text-to-Image Retrieval (ZSIR)

- Propose the use of **wrong classes** in ZSIR
- Use a **probabilistic approach** to project the images and text to a common space with the help of **Common Space Embedding Mapper (CSEM)**
- Train the model by maximizing the lower bound using **Expectation Maximization (EM)**

Details in section 3.3 of the paper

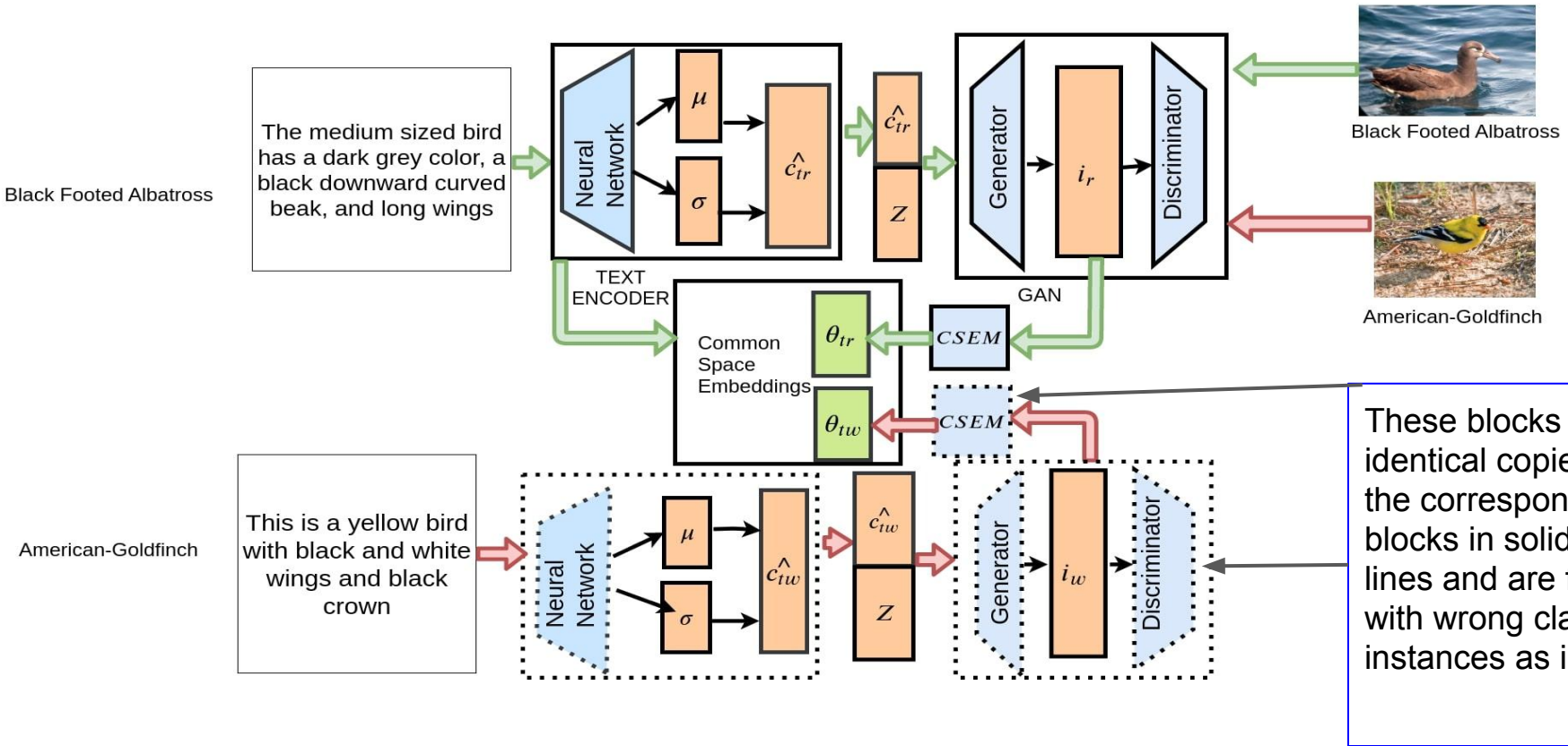
Our Approach: Maximizing joint probability

- We formulate the text to image retrieval problem as maximizing **log Q(I, φ | ψ_c)** which is the joint probability of text embeddings **φ** and relevant image embeddings **I**.
- We derive the lower bound of **log Q(I, φ | ψ_c)**
- We maximize the lower bound using Expectation Maximization.

Details in section 3.3 of the paper

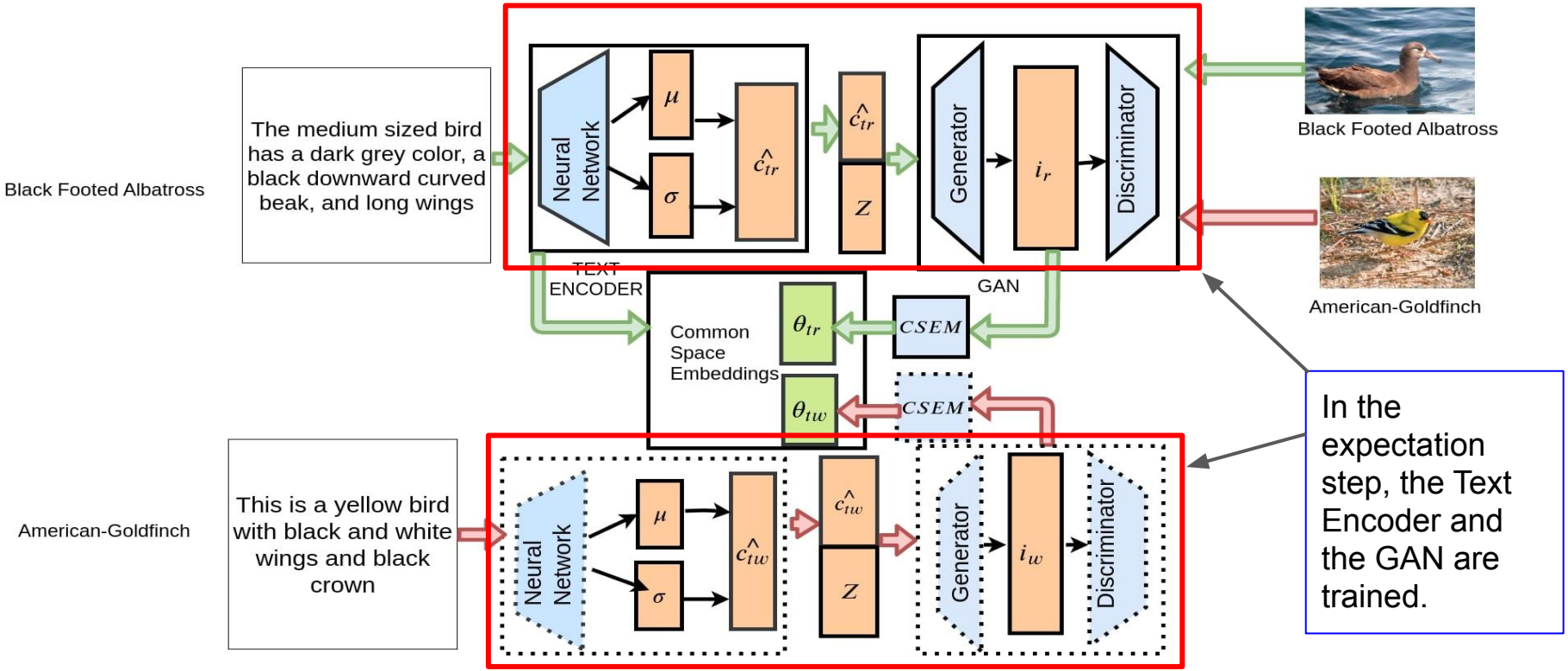
Proposed model: ZSCRGAN

ZSCRGAN



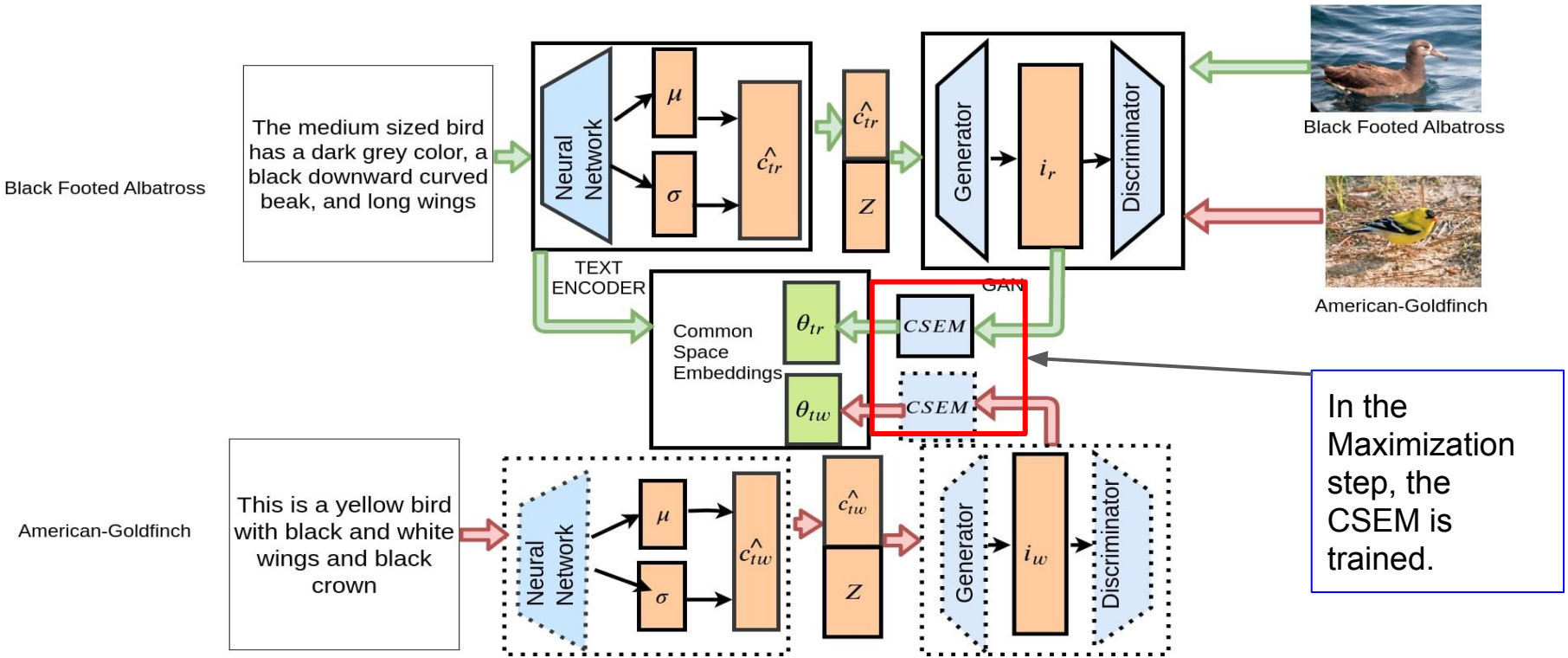
Training -- Expectation Step

ZSCRGAN



Training -- Maximization Step

ZSCRGAN



Training

- The loss function used to train the GAN is Wasserstein loss
- Text-Encoder is trained along with the Generator during the Expectation step
- The CSEM is trained using the Triplet Loss.

Details in section 3.3 of the paper

Experiments

- Benchmark datasets used
  - CUB - images of 200 categories of birds with descriptions
  - Flower - images of 102 categories of flowers with descriptions
  - North American Birds (NAB) - images of 404 categories of birds with descriptions
  - Animals with Attributes(AwA) - images of 50 different categories of animals with 85 dimensional attributes
  - Wikipedia (Wiki) articles - 2,866 image-text pairs taken from Wikipedia documents categorized into 10 classes.

- Evaluation Metrics used
  - Mean Average Precision(mAP)@50
  - Precision(Prec)@50
  - Top-1 Accuracy

Results on CUB dataset

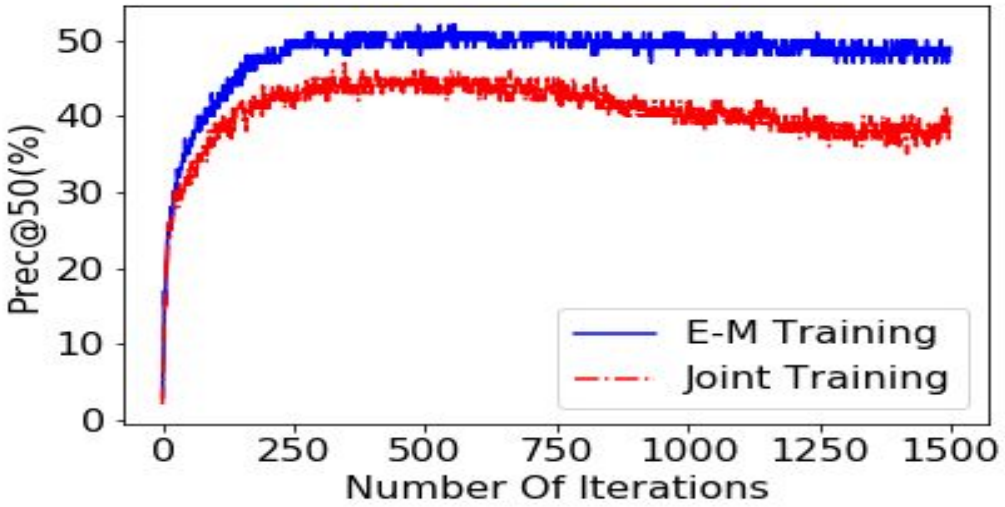
Retrieval Model	Prec@50(%)	mAP@50(%)	Top-1 Acc(%)
SE-ZSL	29.3	45.6	59.6
fCLSWGAN	36.1	52.3	64.0
DS-SJE	45.6	58.8	54.0
ZSL-GAN	42.2	59.2	60.0
DADN	48.9	62.7	68.0
ZSCRGAN (proposed)	52.0	65.4	72.0

Results on Flower dataset

Retrieval Model	Prec@50(%)	mAP@50(%)	Top-1 Acc(%)
SE-ZSL	41.7	63.1	66.4
fCLSWGAN	44.1	67.2	71.2
DS-SJE	55.1	65.7	63.7
ZSL-GAN	38.7	46.6	45.0
DADN	20.8	28.6	25.0
ZSCRGAN (proposed)	59.5	69.4	80.0

Details in tables 3-6 of the paper

Analysis: E-M vs Joint Training



The E-M Training yields better results than the Joint Training approach.
Possible reason -- improper training of the generator in the Joint Training approach.

Details in section 4.5.1 of the paper

Analysis: Choice of wrong class during training

We chose the wrong class randomly, as instances of any class not belonging to the concerned target class.

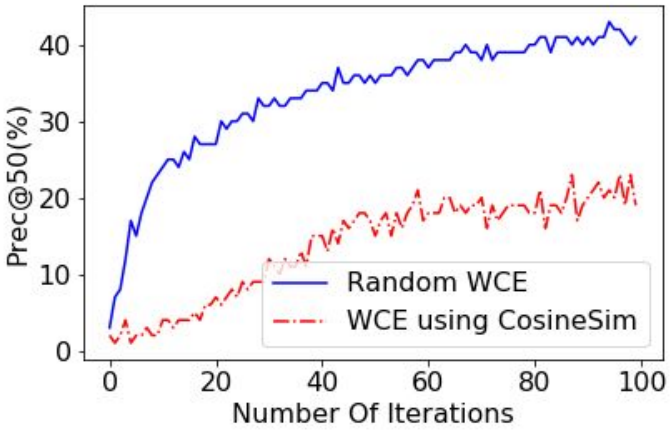
We also try the following ways of choosing the wrong class:

- Using K-Means: Images of a class co-occurring with maximum frequency with that of the target class in a cluster after k-Means clustering of the images.

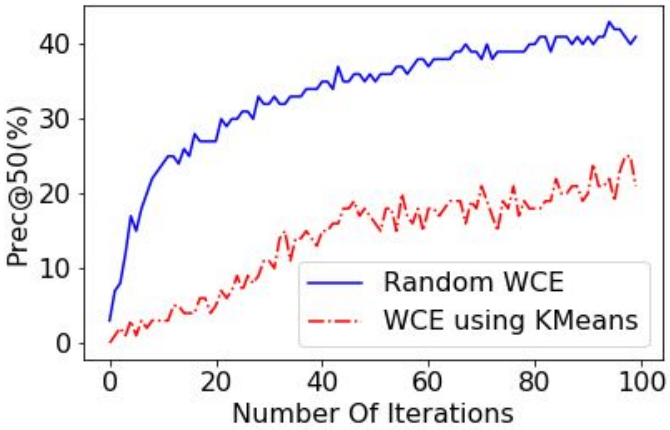
- Using Cosine Similarity : Class whose text-embedding has the maximum cosine similarity with that of the target class

Analysis: Choice of wrong class

Based on Cosine Similarity



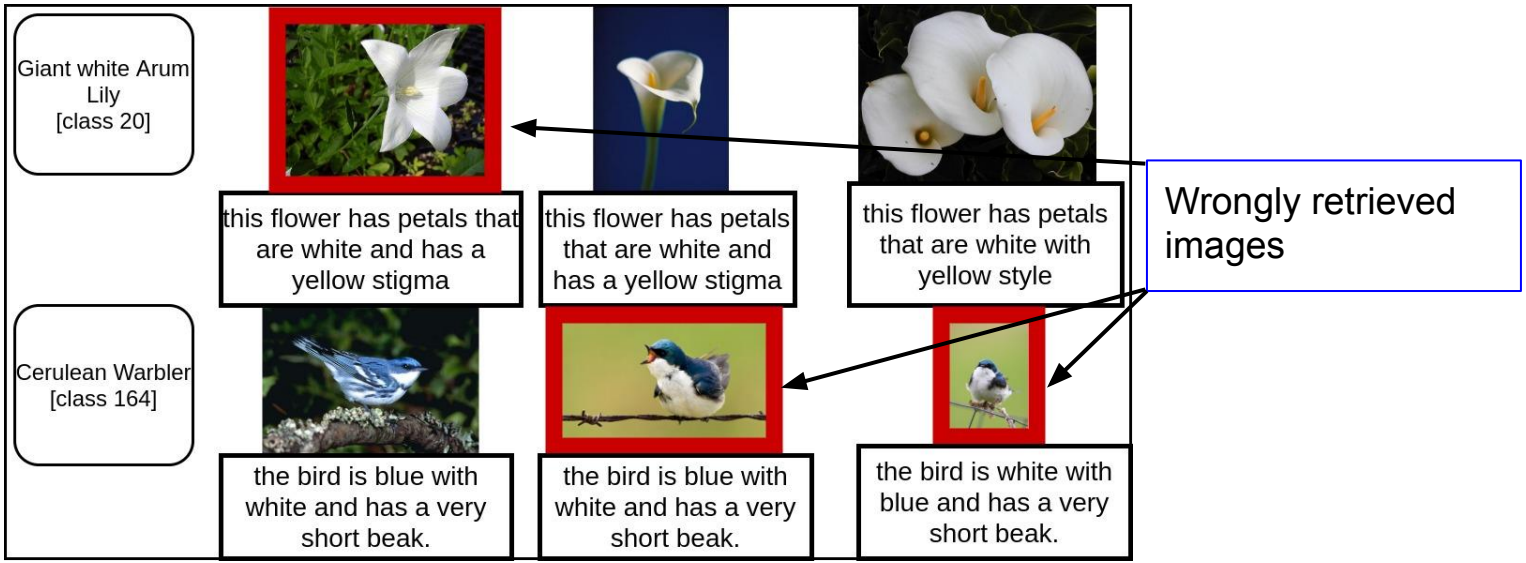
Based on kMeans



Restricting the choice of the wrong classes distorts the space of the wrong embeddings, and hence runs the risk of the model identifying embeddings from classes outside the space of distorted wrong embeddings as relevant.

Details in section 4.5.2 of the paper

Analysis: Errors of the proposed model



Even when the proposed model retrieves a wrong image, it is actually very similar to the description in the query

Details in section 4.6 of the paper

Ablation study on ZSCRGAN

Ablation study to understand the importance of various components

Retrieval Model	Prec@50(%) CUB dataset	Prec@50(%) Flowers dataset
ZSCRGAN (with all components)	**52.0**	**59.5**
w/o use of wrong class embedding	24.7	27.2
w/o R (regularizer) and Triplet Loss (CSEM)	23.8	33.7
w/o Triplet Loss	36.2	41.4
w/o R (regularizer)	48.4	35.2
w/o GAN	25.9	32.0

Details in section 4.7 of the paper

Conclusion

- Proposed a novel Zero-Shot text to image retrieval model
- Formulated an E-M setup for training the model
- Used Wrong Class embeddings for training
- Evaluated our model on a number of benchmark datasets.
- Detailed Analysis of our model
- We have made our code publicly available[1]

The link to our implementation: https://github.com/ranarag/ZSCRGAN

Future Work

In future we would like to extend our work for

- Other types of zero-shot cross-modal retrieval setups
- Various Types of zero-shot multi-view retrieval setups

Acknowledgement

- SIGIR Student Travel Grants Program for sponsoring the conference registration
- Nvidia Corporation for a TitanXp GPU which has been used for this research

Thank You !

Questions, Suggestions ....