Yeni Kavramları En Az Denetim ile Temsil Etme ve Açıklama

Zeynep Akata

Bilim Akademisi - Bilkent Üniversitesi
Yapay Öğrenme Yaz Okulu 2020
30 Haziran 2020

Outline

Outline

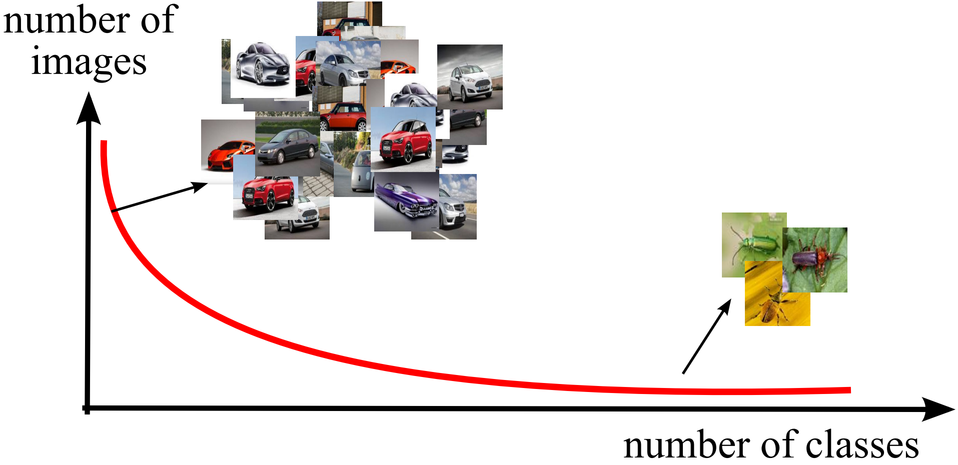
Generalized Low-Shot Learning with Side-Information

Generating Natural Language Explanations for Visual Decisions

Summary and Future Work

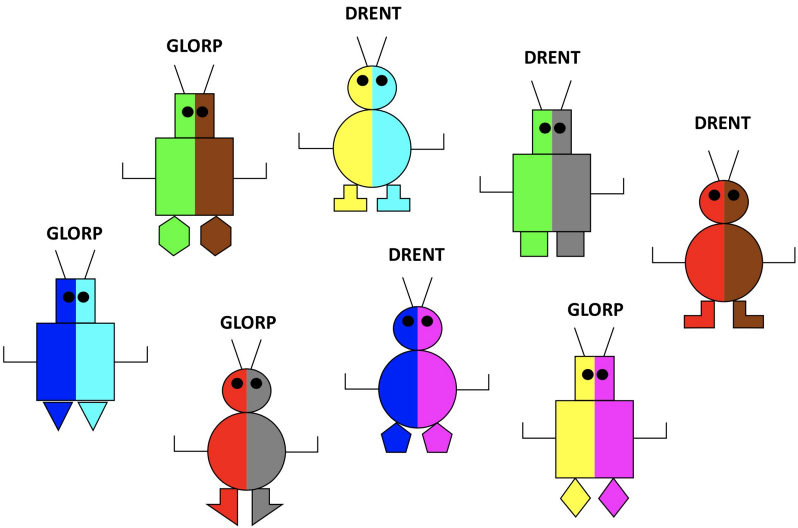
Data Distribution in Large-Scale Datasets

Akata et.al. TPAMI’14



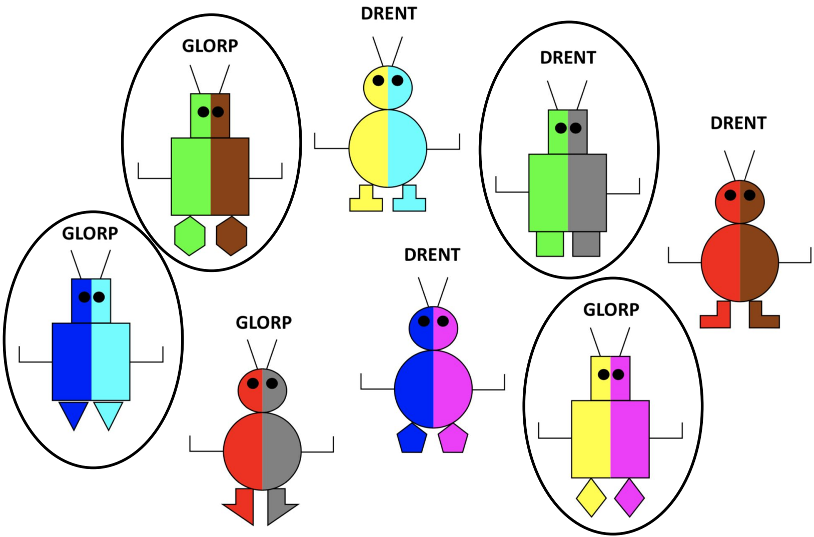
Learning via Explanation

Lombrozo TICS'16

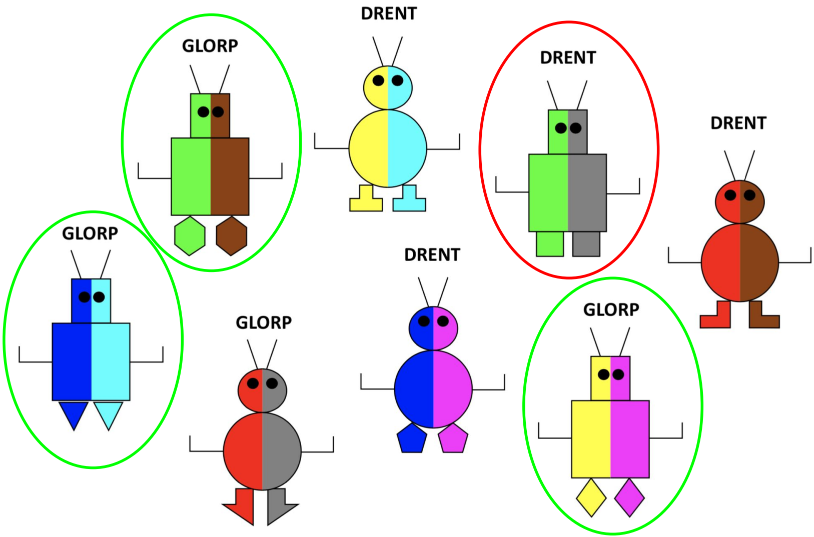


Learning via Explanation

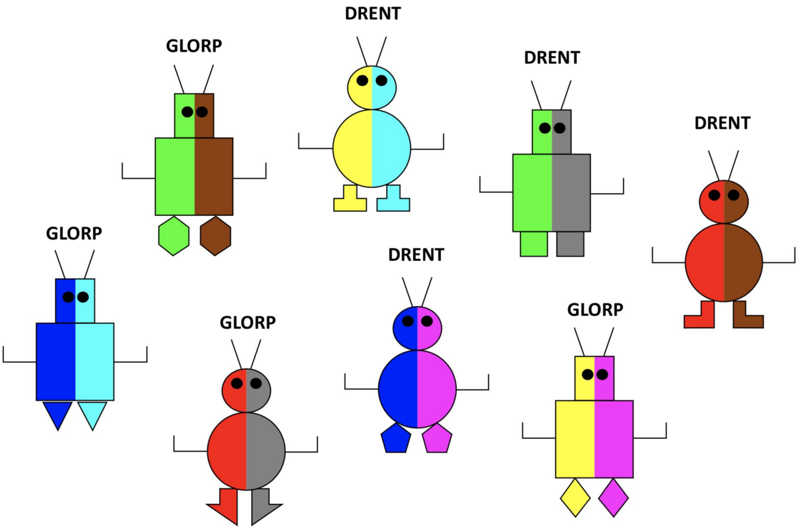
Lombrozo TICS'16



Learning via Explanation

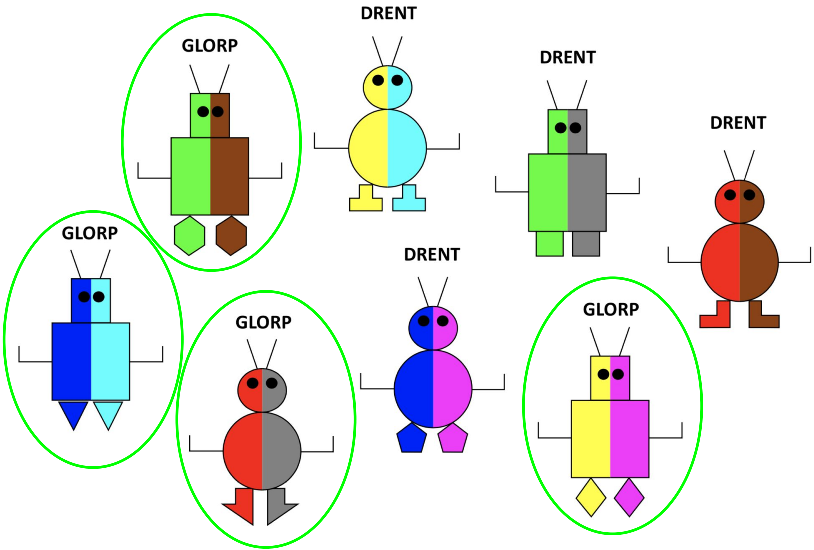


Lombrozo TICS'16



Learning via Explanation

Lombrozo TICS'16



Attributes as Explanations

Lampert et al. CVPR'09

images | **attributes** | **class**

black-white
has tail
lives on land
small

zebra
[1 0 1 1 0 1]

gray
has tail
lives in water
big

[0 1 1 0 1 0]
whale

Attributes as Explanations

Lampert et al. CVPR'09

images | **attributes** | **class**

black-white
has tail
lives on land
small

zebra
[1 0 1 1 0 1]

gray
has tail
lives in water
big

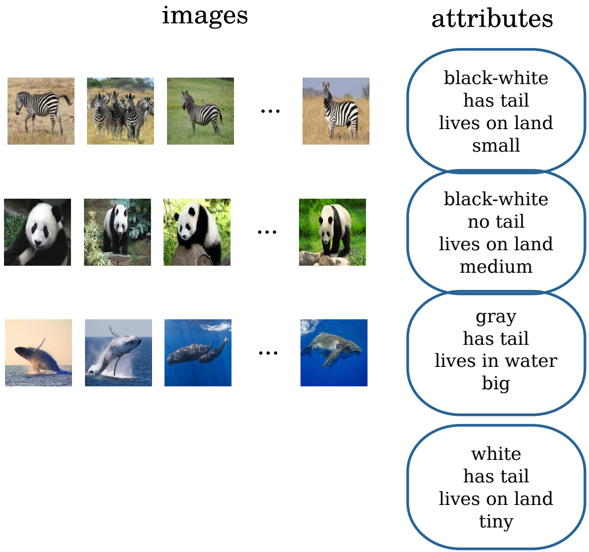
[0 1 1 0 1 0]
whale

Attributes as Explanations

Lampert et al. CVPR'09

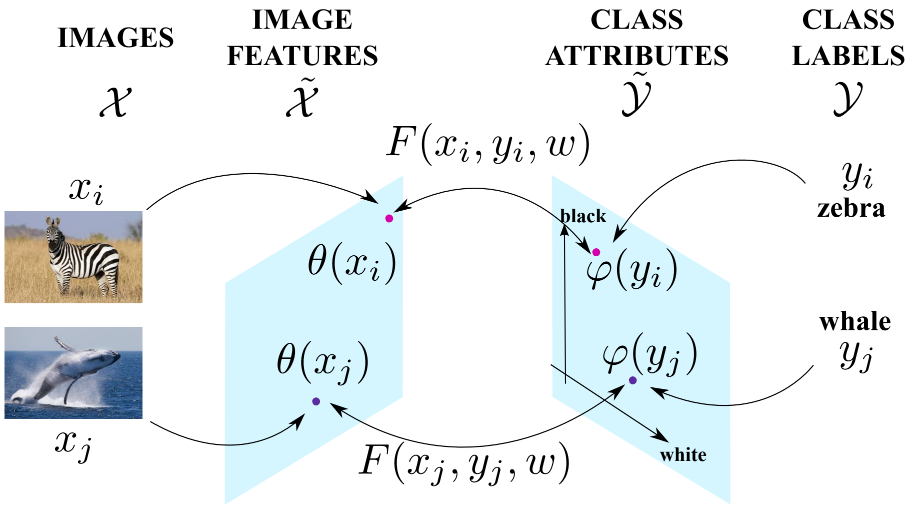
images	attributes	class
	black-white has tail lives on land small	zebra [1 0 1 1 0 1]
	gray has tail lives in water big	[0 1 1 0 1 0] whale

Generalized Zero-Shot Learning



Muldimodal Embeddings

Akata et al. CVPR'13 & TPAMI'16



$\mathcal{S} = \{(x, y, \varphi(y)) \mid x \in \mathcal{X}, y \in \mathcal{Y}^s, \varphi(y) \in \mathcal{C}\}$ and $\mathcal{U} = \{(y, \varphi(y)) \mid y \in \mathcal{Y}^u, \varphi(y) \in \mathcal{C}\}$

Multimodal Embeddings

$\mathcal{S} = \{(x, y, \varphi(y)) \mid x \in \mathcal{X}, y \in \mathcal{Y}^s, \varphi(y) \in \mathcal{C}\}$ and $\mathcal{U} = \{(y, \varphi(y)) \mid y \in \mathcal{Y}^u, \varphi(y) \in \mathcal{C}\}$

Learn $f : \mathcal{X} \to \mathcal{Y}$ by minimizing regularized empirical risk:

$$\frac{1}{N} \sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W)$$

$L(.)$ = loss function, $\Omega(.)$ = regularization term, using pairwise ranking loss:

Multimodal Embeddings

Akata et al.CVPR'13 & TPAMI'16

$\mathcal{S} = \{(x, y, \varphi(y)) \mid x \in \mathcal{X}, y \in \mathcal{Y}^s, \varphi(y) \in \mathcal{C}\}$ and $\mathcal{U} = \{(y, \varphi(y)) \mid y \in \mathcal{Y}^u, \varphi(y) \in \mathcal{C}\}$

Learn $f : \mathcal{X} \to \mathcal{Y}$ by minimizing regularized empirical risk:

$$\frac{1}{N}\sum_{n=1}^{N} L(y_n, f(x_n; W)) + \Omega(W)$$

$L(.)$ = loss function, $\Omega(.)$ = regularization term, using pairwise ranking loss:

$$L(x_n, y_n, y; W) = \sum_{y \in \mathcal{Y}^s} [\Delta(y_n, y) + F(x_n, y; W) - F(x_n, y_n; W)]_+$$

with the compatibility function: $F(x, y; W) = \theta(x)^T W \varphi(y)$

Benchmark Example Datasets

Animals with Attributes (AWA) [Lampert et.al. CVPR'09]	50 cls	85 att	
Caltech UCSD-Birds (CUB) [Wah et.al.'11]	200 cls	312 att	

Benchmark Results

	CUB			AWA		
Method	**u**	**s**	**H**	**u**	**s**	**H**
Supervised Learning	–	82.1	–	–	96.2	–
Multimodal Embeddings	23.7	62.8	34.4	16.8	76.1	27.5

u/**s**: $acc_{\mathcal{Y}^{u/s}} = \frac{1}{\|\mathcal{Y}^{u/s}\|} \sum_{c=1}^{\|\mathcal{Y}^{u/s}\|} \frac{\#\text{ correct in c}}{\#\text{ samples in c}}$ and **H** $= \frac{2 * acc_{\mathcal{Y}^s} * acc_{\mathcal{Y}^u}}{acc_{\mathcal{Y}^s} + acc_{\mathcal{Y}^u}}$

How to Tackle the Missing Data Problem?

Labels are difficult to obtain, attributes require expert knowledge

How to Tackle the Missing Data Problem?

Labels are difficult to obtain, attributes require expert knowledge

Proposed solution: Free text to image synthesis!

Detailed Visual Descriptions as Side Information

Reed et al. CVPR'16

The bird has a white underbelly, black feathers in the wings, a large wingspan, and a white beak.

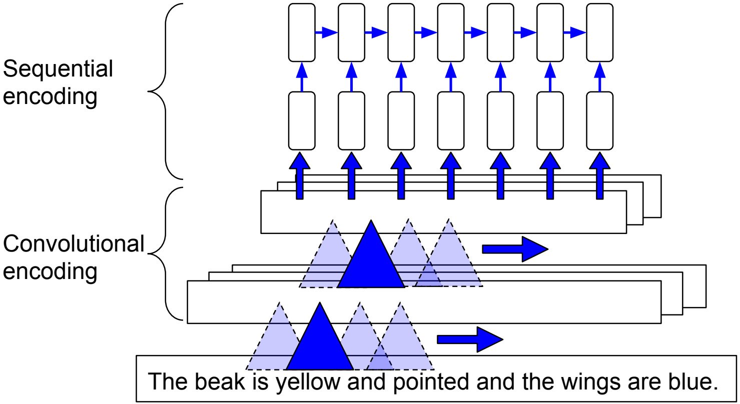
This bird has distinctive-looking brown and white stripes all over its body, and its brown tail sticks up.

This flower has a central white blossom surrounded by large pointed red petals which are veined and leaflike.

Light purple petals with orange and black middle green leaves

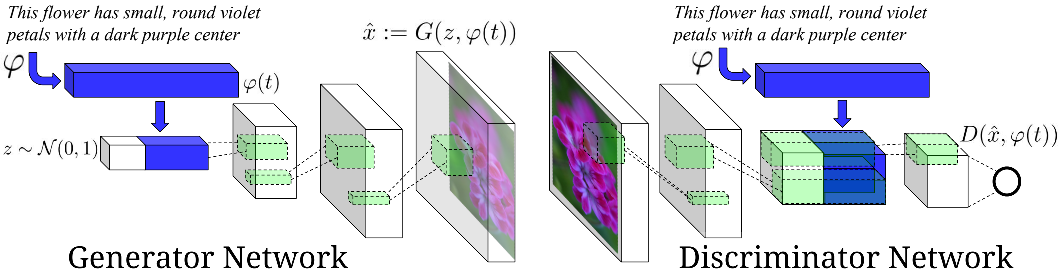
Deep Representations of Text

Reed et al. CVPR'16



GAN[1] Conditioned on Text

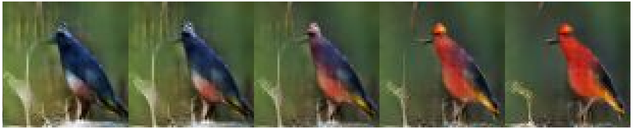
Reed et al. ICML'16 & NIPS'16



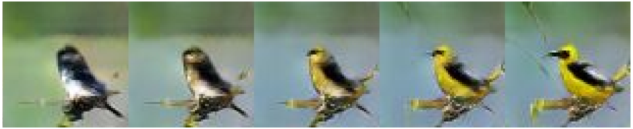
[1] Generative Adversarial Networks [Goodfellow et al. NIPS'14]

Text to Image Synthesis Results

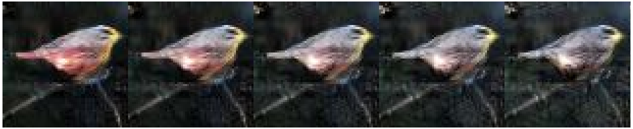
‘Blue bird with black beak’ $\rightarrow$
‘Red bird with black beak’



‘Small blue bird with black wings’ $\rightarrow$
‘Small yellow bird with black wings’



‘This bird is bright.’ $\rightarrow$ ‘This bird is dark.’



Generalized Zero-Shot Learning with Synthesized Images

	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4

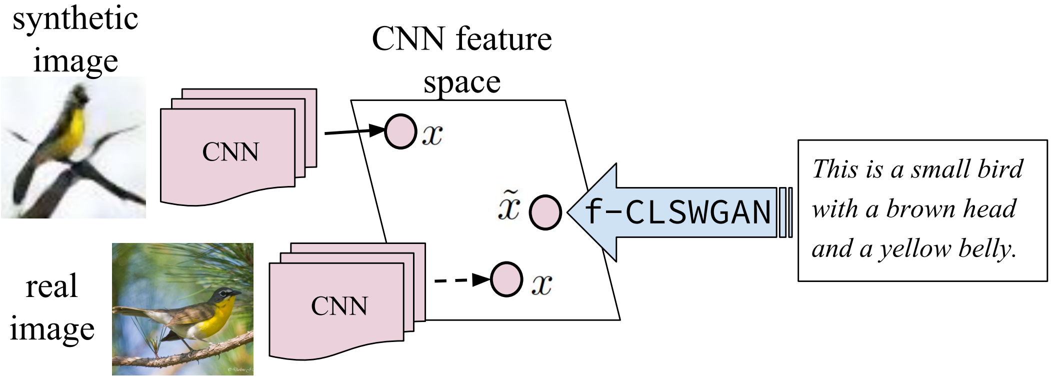
Generalized Zero-Shot Learning with Synthesized Images

	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4
With generated images	23.8	48.5	31.9

This is not better than having no images!

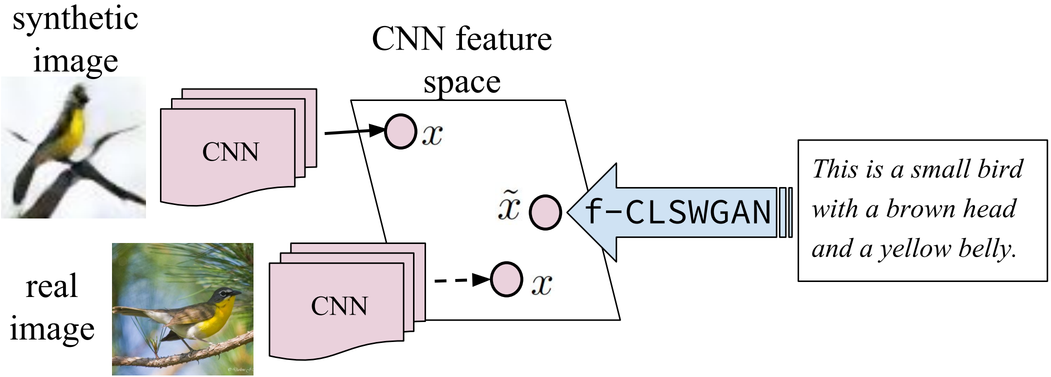
f-CLSWGAN for Text to Image Feature Synthesis

Xian et al. CVPR'18



f-CLSWGAN for Text to Image Feature Synthesis

Xian et al. CVPR'18



$\mathcal{S} = \{(x, y, \varphi(y)) \mid x \in \mathcal{X}, y \in \mathcal{Y}^s, \varphi(y) \in \mathcal{C}\}$ and
$\mathcal{U} = \{(\tilde{x}, y, \varphi(y)) \mid \tilde{x} = G(z, \varphi(y)), y \in \mathcal{Y}^u, \varphi(y) \in \mathcal{C}\}$: combine to train a classifier

Generalized Zero-Shot Learning with Synthesized Image Features

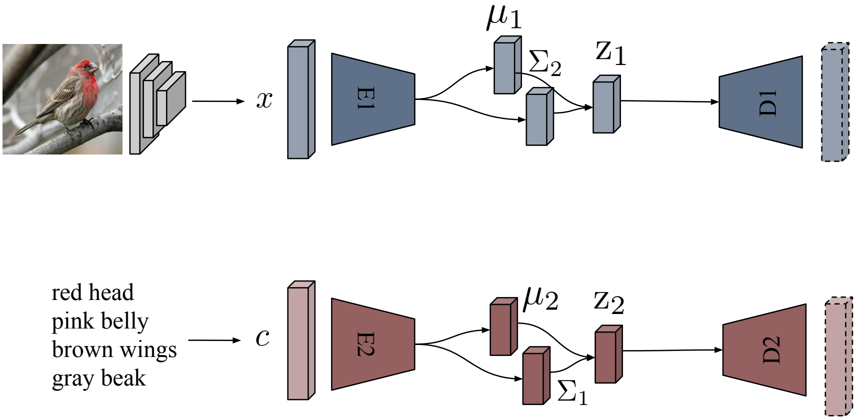
	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4
With generated images	23.8	48.5	31.9

Generalized Zero-Shot Learning with Synthesized Image Features

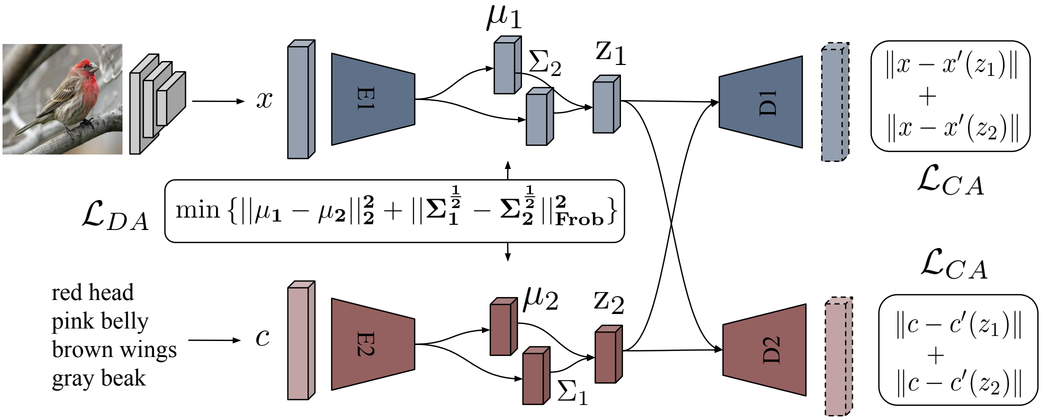
	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4
With generated images	23.8	48.5	31.9
With generated features (`f-CLSWGAN`)	43.7	57.7	**49.7**

CADA-VAE for Text to Latent Feature Synthesis

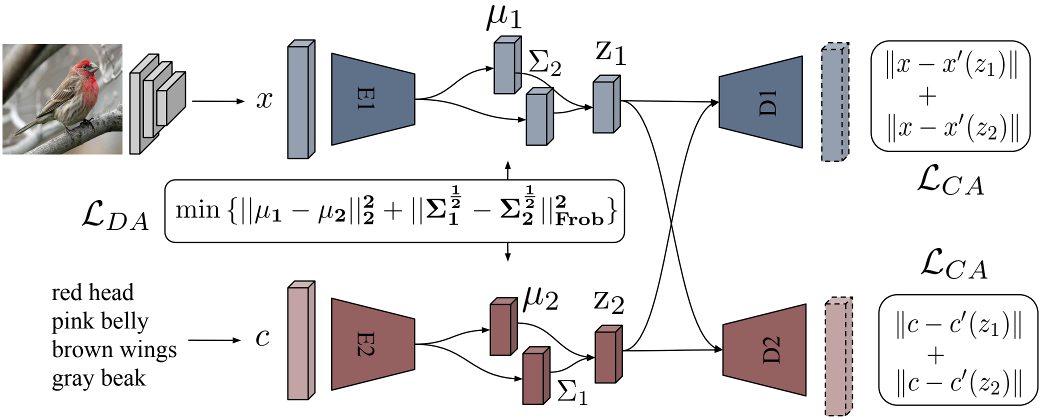
Schönfeld et al. CVPR'19



CADA-VAE for Text to Latent Feature Synthesis Schönfeld et al. CVPR'19



CADA-VAE for Text to Latent Feature Synthesis   Schönfeld et al. CVPR'19

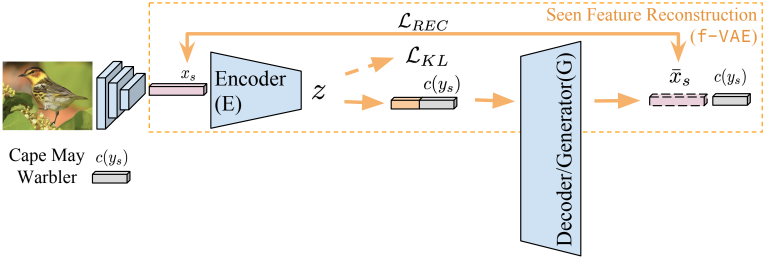


$\mathcal{S} = \{(z, y, c) \mid z \in z_1, y \in \mathcal{Y}^s, c \in \mathcal{C}\}$ and
$\mathcal{U} = \{(z, y, c) \mid z \in z_2, y \in \mathcal{Y}^u, c \in \mathcal{C}\}$: combine to train a classifier

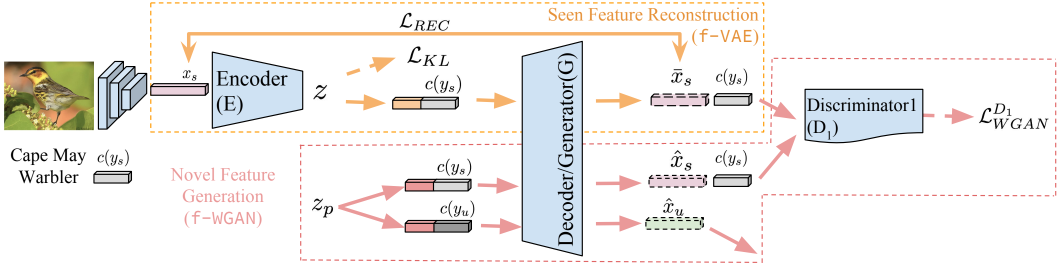
Generalized Zero-Shot Learning with Latent Features

	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4
With generated images	23.8	48.5	31.9
With generated features (`f-CLSWGAN`)	43.7	57.7	49.7
With generated features (`CADA-VAE`)	63.6	51.6	**52.4**

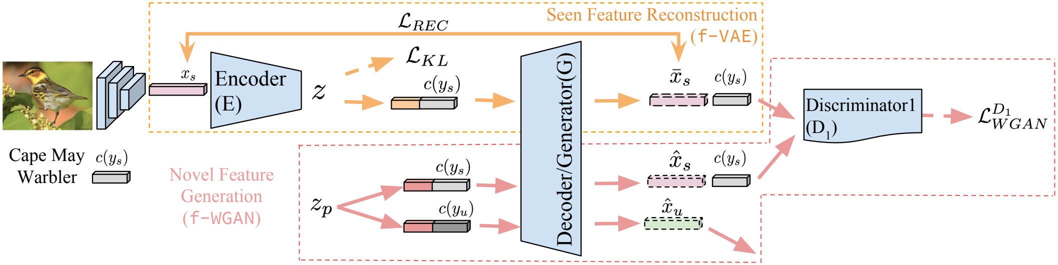
f-VAEGAN-D2 for Text to Image Feature Synthesis Xian et al. CVPR'19



f-VAEGAN-D2 for Text to Image Feature Synthesis Xian et al. CVPR'19



f-VAEGAN-D2 for Text to Image Feature Synthesis   Xian et al. CVPR'19

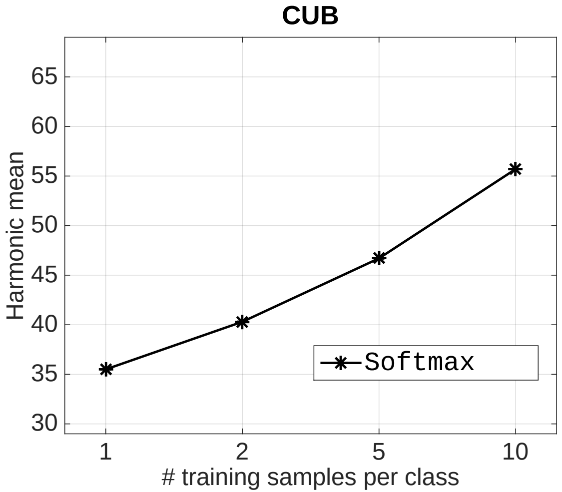


$\mathcal{S} = \{(x_s, y, c(y_s)) \mid x_s \in \mathcal{X}, y \in \mathcal{Y}^s, c(y_s) \in \mathcal{C}\}$ and
$\mathcal{U} = \{(\hat{x}_u, y, c(y_u)) \mid \hat{x}_u = G(z, \varphi(y)), y \in \mathcal{Y}^u, c(y_u) \in \mathcal{C}\}$: combine to train a classifier

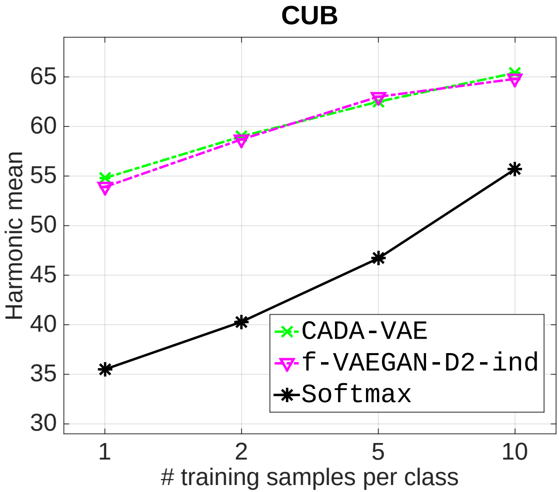
Generalized Zero-Shot Learning with Synthesized Image Features

	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4
With generated images	23.8	48.5	31.9
With generated features (`f-CLSWGAN`)	43.7	57.7	49.7
With generated features (`CADA-VAE`)	63.6	51.6	52.4
With generated features (`f-VAEGAN-D2`)	63.2	75.6	**68.9**

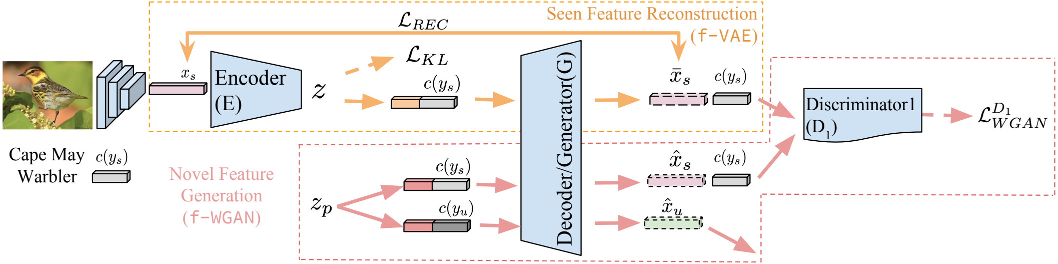
Generalized Few-Shot Learning Results



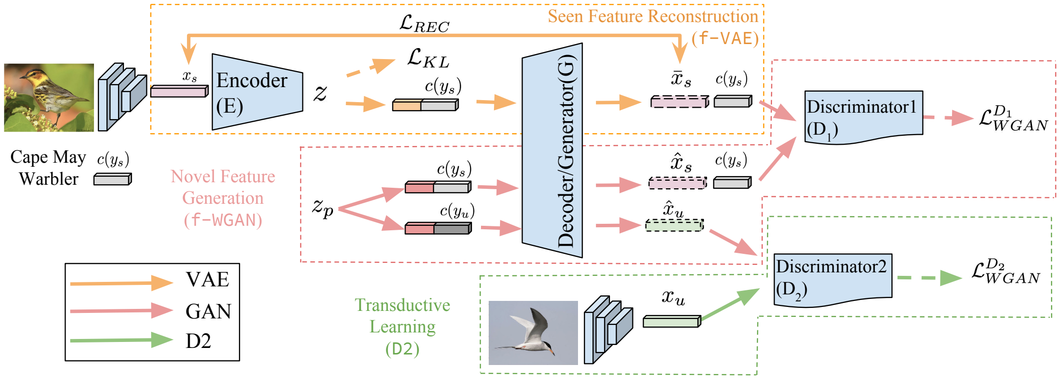
Generalized Few-Shot Learning Results



f-VAEGAN-D2 for Text to Image Feature Synthesis Xian et al. CVPR'19



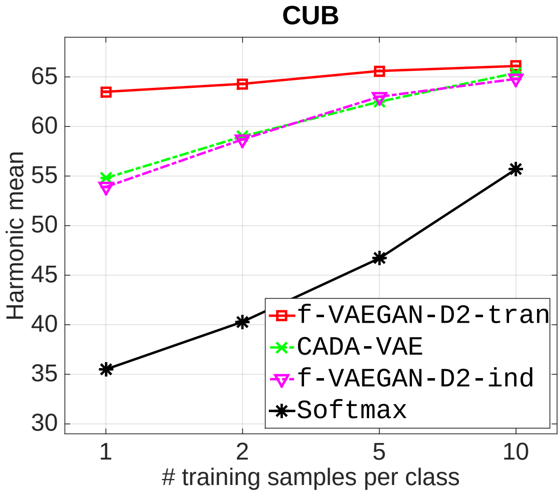
f-VAEGAN-D2 for Text to Image Feature Synthesis   Xian et al. CVPR'19



Generalized Zero-Shot Learning with Synthesized Image Features

	CUB		
Data	**u**	**s**	**H**
Only real data	23.7	62.8	34.4
With generated images	23.8	48.5	31.9
With generated features (`f-CLSWGAN`)	43.7	57.7	49.7
With generated features (`CADA-VAE`)	63.6	51.6	52.4
With generated features (`f-VAEGAN-D2`)	63.2	75.6	**68.9**
With generated features (`f-VAEGAN-D2` tran)	73.8	81.4	**77.3**

Generalized Few-Shot Learning Results



Conclusions

Language complements visual information

1. Provides an intuitive interface for the model
2. Strong and generalizable: any-shot image classification
3. Guides generative models for learning representations

Akata et al. IEEE CVPR 2013, 2015, 2016, TPAMI 2014, 2016
Reed et al. IEEE CVPR 2016 & ICML 2016 & NIPS 2016
Xian et al. IEEE CVPR 2016, 2017, 2018, 2019a, 2019b
Schönfeld et al. IEEE CVPR 2019; Dutta and Akata IEEE CVPR 2019

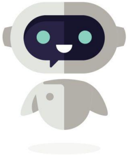
Outline

Generalized Low-Shot Learning with Side-Information

Generating Natural Language Explanations for Visual Decisions

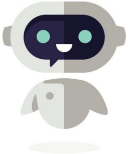
Summary and Future Work

Human Machine Communication: Visual Question Answering

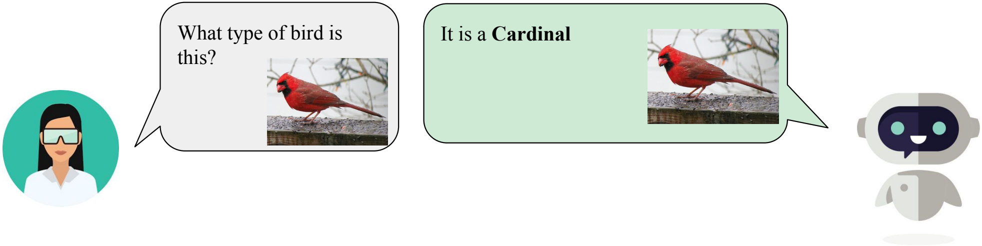


Human Machine Communication: Visual Question Answering

What type of bird is this?



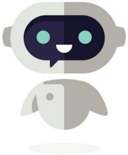
Human Machine Communication: Visual Question Answering



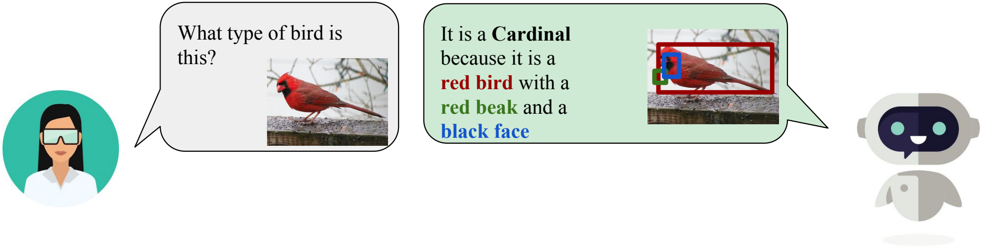
Human Machine Communication: Visual Question Answering

What type of bird is this?

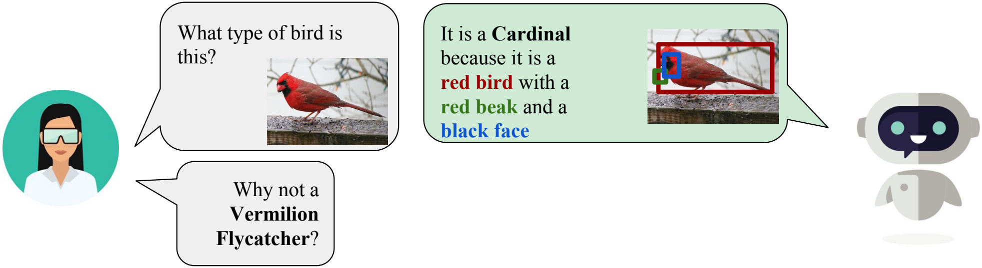
It is a **Cardinal** because it is a red bird with a red beak and a black face



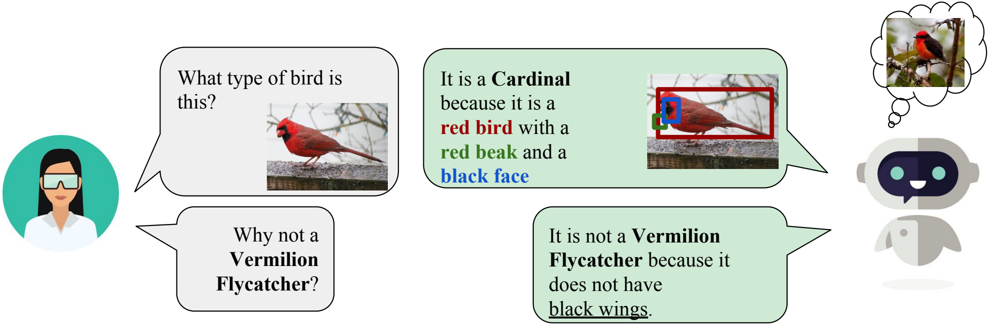
Human Machine Communication: Visual Question Answering



Human Machine Communication: Visual Question Answering

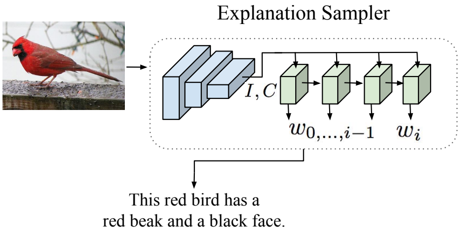


Human Machine Communication: Visual Question Answering



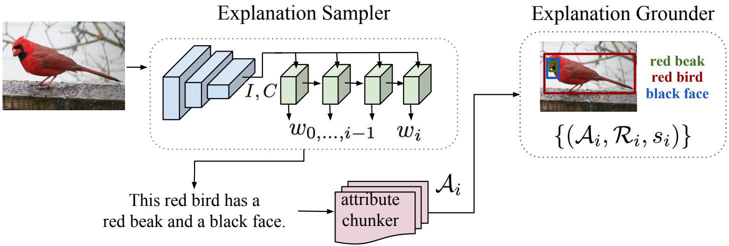
Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



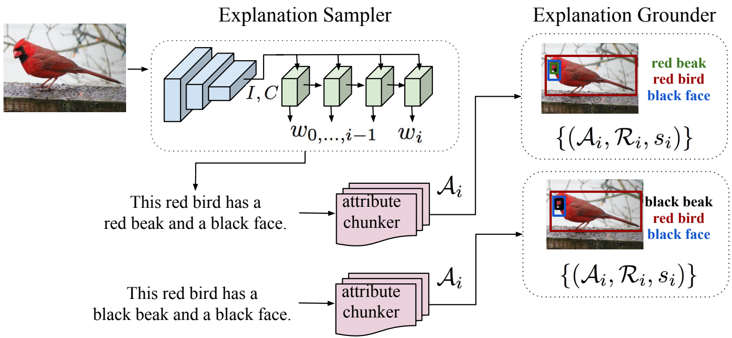
Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



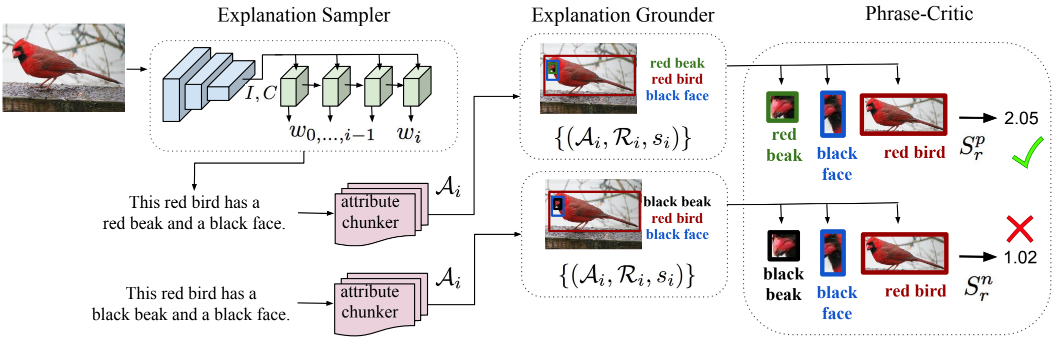
Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



Grounding Visual Explanations

Hendricks et al. ECCV'16 & ECCV'18



Generating Visual Explanations Results

This is a ***Downy Woodpecker*** *because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a black and white bird with a **red spot** on its crown.

This is a ***Downy Woodpecker*** *because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a white bird with a black wing and a black and white striped head.

Generating Visual Explanations Results

*This is a **Downy Woodpecker** because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a black and white bird with a **red spot** on its crown.

*This is a **Downy Woodpecker** because...*

D: this bird has a white breast black wings and a **red spot** on its head.
E: this is a white bird with a black wing and a black and white striped head.

Correct: Laysan Albatross, **Predicted:** Cactus Wren

Explanation: ...this is a brown and white spotted bird with a long pointed beak.

Correct & Predicted: Laysan Albatross

Explanation: ...this bird has a white head and breast with a long hooked bill.

***Cactus Wren* Definition:** ...this bird has a long thin beak with a brown body and black spotted feathers.
***Laysan Albatross* Definition:** ...this bird has a white head and breast a grey back and wing feathers and an orange beak.

Grounding Visual Explanations and Counterfactuals

This is a **Red Winged Blackbird** because ….

this is a **black bird** with a **red spot on its wingbars**.

Score: -11.29

this is a black bird with a red wing and a pointy black beak.

Grounding Visual Explanations and Counterfactuals

This is a **Red Winged Blackbird** because ….

this is a **black bird** with a **red spot on its wingbars**.

Score: -11.29

this is a black bird with a red wing and a pointy black beak.

Counterfactuals: Contrasting explanations are intuitive and informative

This bird is a **Crested Auklet** because this is a black bird with a small orange beak and it is not a **Red Faced Cormorant** because it does not have a long flat bill.

Textual Explanations for Self Driving Vehicles

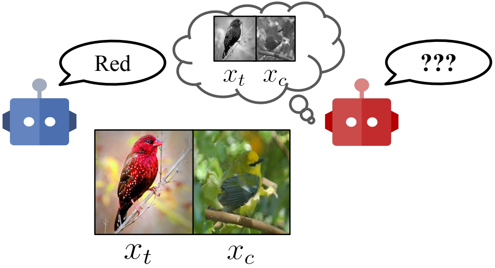
Kim et al. ECCV'18

The car heads down the road because traffic is moving at a steady pace.

The car is slowing because it is approaching a stop sign.

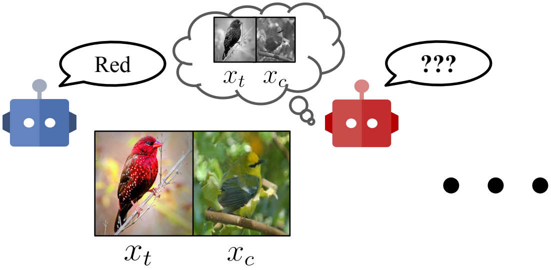
The car is stopped because the car in front of it is stopped.

Image reference game between agents with variations in the understanding of the world



Modeling Conceptual Understanding

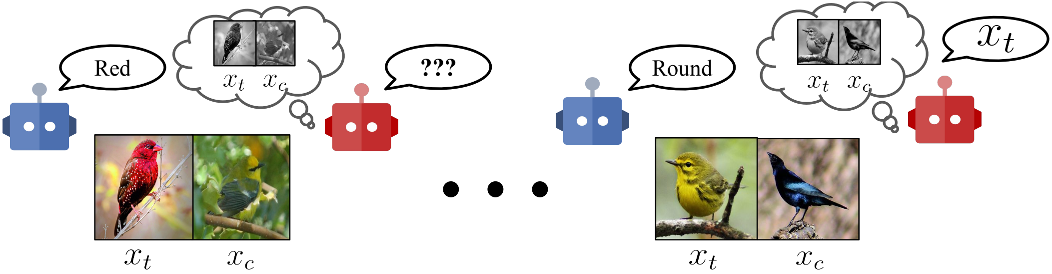
Image reference game between agents with variations in the understanding of the world



Modeling Conceptual Understanding

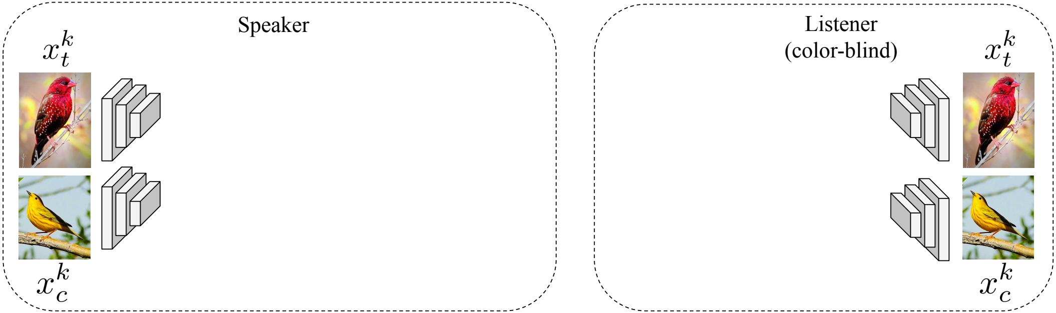
Rodriguez et al. NeurIPS'19

Image reference game between agents with variations in the understanding of the world



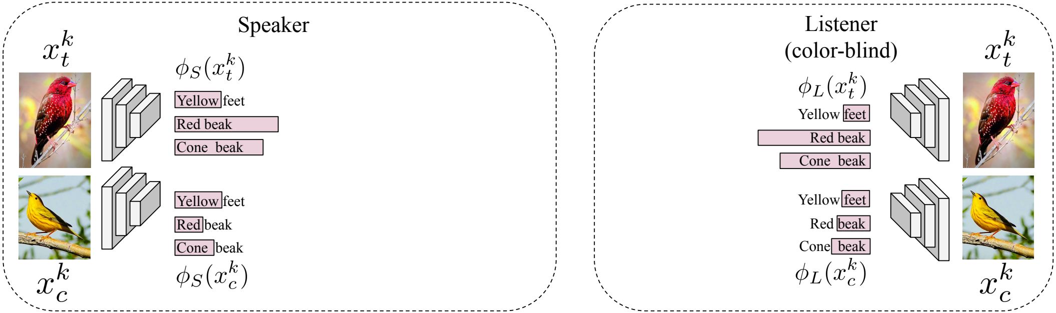
Modeling Conceptual Understanding

Rodriguez et al. NeurIPS'19



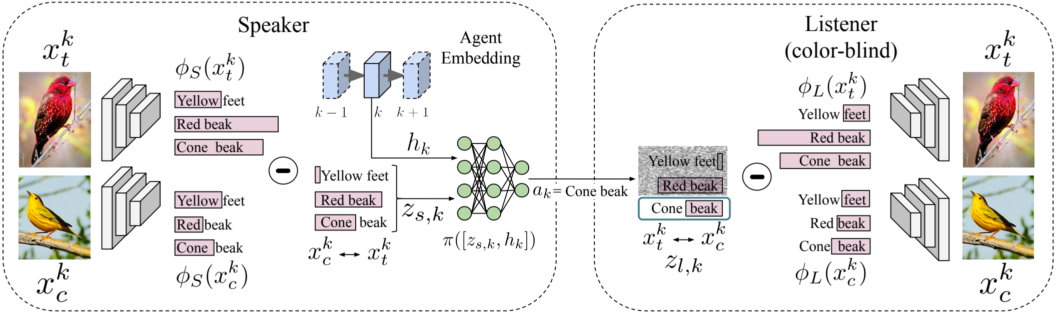
Modeling Conceptual Understanding

Rodriguez et al. NeurIPS'19



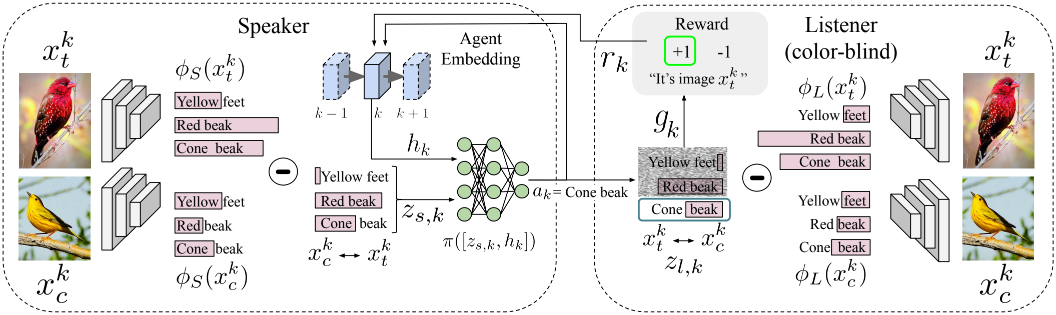
Modeling Conceptual Understanding

Rodriguez et al. NeurIPS'19



Modeling Conceptual Understanding

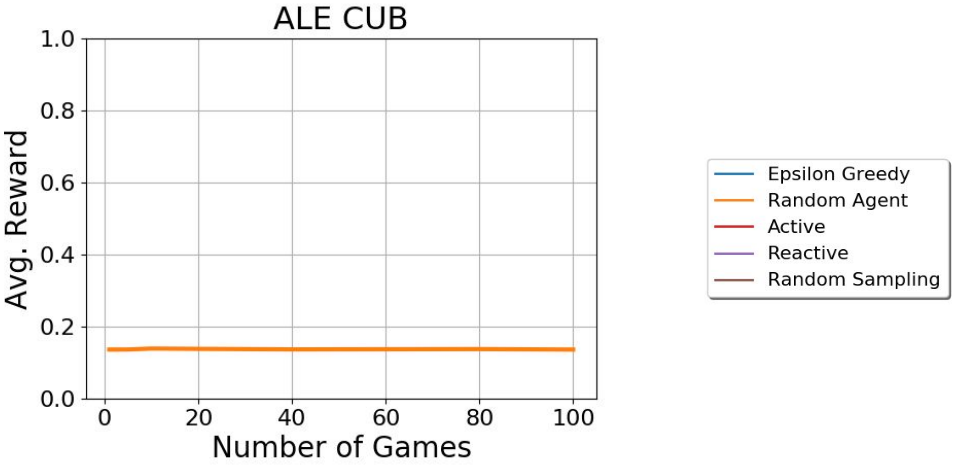
Rodriguez et al. NeurIPS'19



- Speaker adapts to the listener by incorporating information after each game

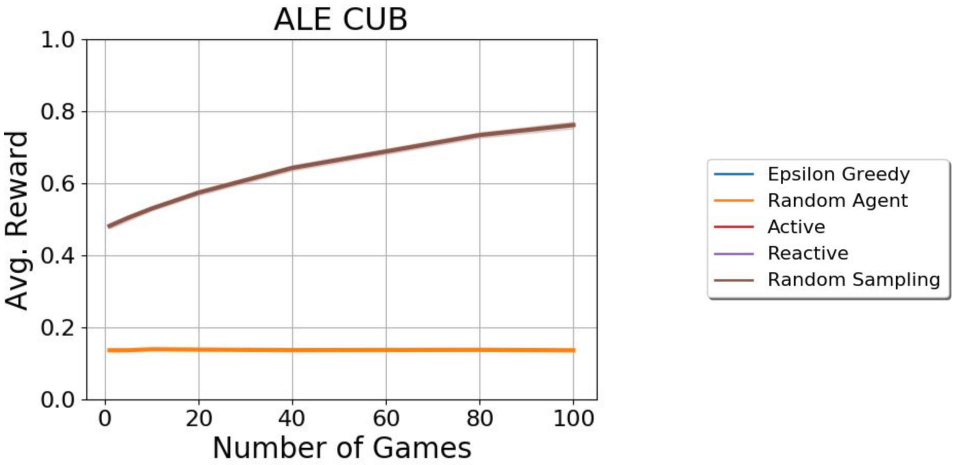
Modeling Conceptual Understanding Results

Rodriguez et al. NeurIPS'19



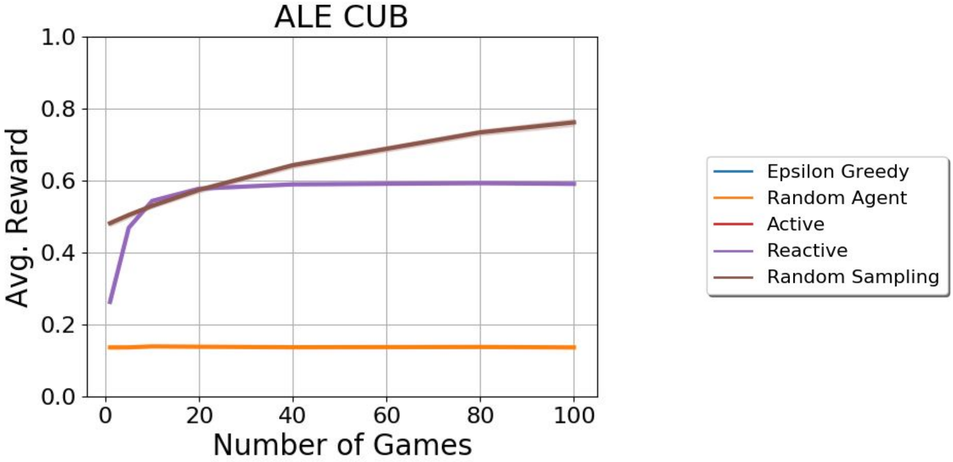
Modeling Conceptual Understanding Results

Rodriguez et al. NeurIPS'19



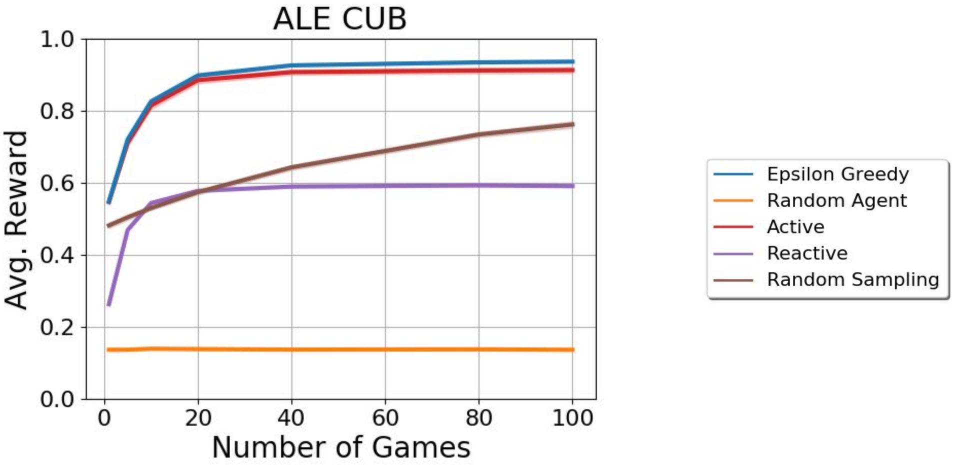
Modeling Conceptual Understanding Results

Rodriguez et al. NeurIPS'19



Modeling Conceptual Understanding Results

Rodriguez et al. NeurIPS'19

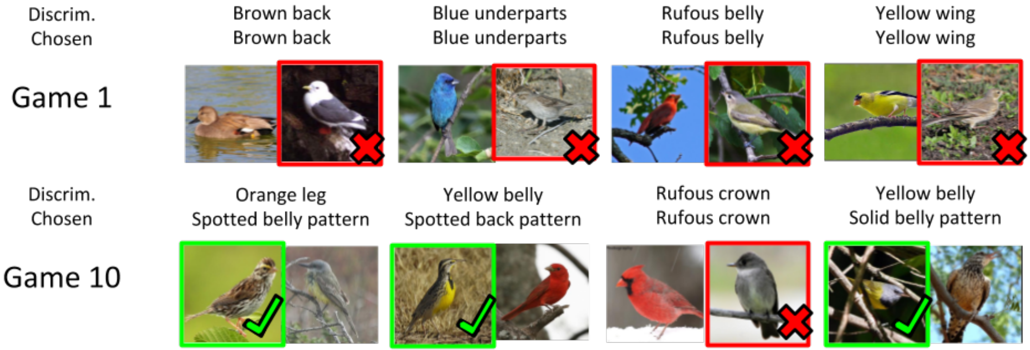


Modeling Conceptual Understanding Qualitative Results

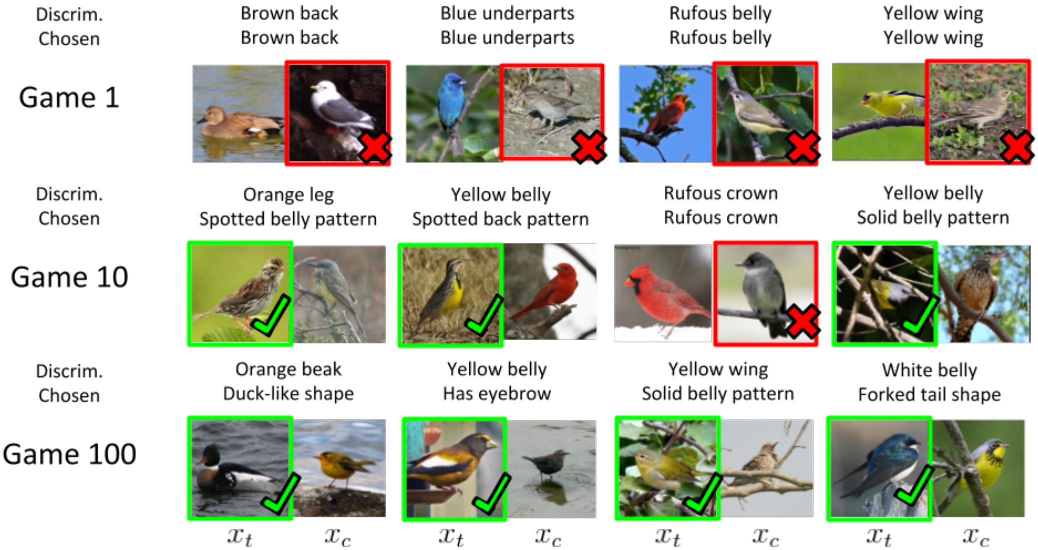
Discrim.	Brown back	Blue underparts	Rufous belly	Yellow wing
Chosen	Brown back	Blue underparts	Rufous belly	Yellow wing

Game 1

Modeling Conceptual Understanding Qualitative Results



Modeling Conceptual Understanding Qualitative Results



Conclusions

Generating visual/textual explanations

1. A means for model interpretation: necessary to improve deep models
2. Important criteria to trust deep models: through explanations
3. A step towards effective human-machine communication

Hendricks et al. ECCV 2016 & ECCV 2018,
Park et al. IEEE CVPR 2018, Kim et al. ECCV 2018
Rodriguez et.al. NeurIPS 2019

Outline

Generalized Low-Shot Learning with Side-Information

Generating Natural Language Explanations for Visual Decisions

Summary and Future Work

Summary

1. Multi-modal Joint Embeddings tackle lack of visual data
   [Akata et al. CVPR'13, CVPR'15, CVPR'16 & TPAMI'14, TPAMI'16]

Summary

1. Multi-modal Joint Embeddings tackle lack of visual data
   [Akata et al. CVPR'13, CVPR'15, CVPR'16 & TPAMI'14, TPAMI'16]
2. Vision and Language complement each other for generating novel concepts
   [Reed et al. CVPR'16 & ICML'16 & NIPS'16, Xian et al. CVPR'16, CVPR'17, CVPR'18, CVPR'19a & CVPR'19b, Schönfeld et al. CVPR'19, Dutta and Akata CVPR'19]

Summary

1. Multi-modal Joint Embeddings tackle lack of visual data
   [Akata et al. CVPR'13, CVPR'15, CVPR'16 & TPAMI'14, TPAMI'16]
2. Vision and Language complement each other for generating novel concepts
   [Reed et al. CVPR'16 & ICML'16 & NIPS'16, Xian et al. CVPR'16, CVPR'17, CVPR'18, CVPR'19a & CVPR'19b, Schönfeld et al. CVPR'19, Dutta and Akata CVPR'19]
3. Developing explainable deep models is important for user acceptance
   [Hendricks et al. ECCV'16 & ECCV'18, Park et al. CVPR'18, Kim et al. ECCV'18, Rodriguez et al. NeurIPS'19]

Future of Deeply Explainable Artificial Intelligence

Future of Deeply Explainable Artificial Intelligence

User: What happened?

Future of Deeply Explainable Artificial Intelligence

User: What happened?

AI: I was driving down an empty road. I decided to slow down as a ball appeared on the right. I saw a child running towards the ball, so I decided to stop.

Future of Deeply Explainable Artificial Intelligence

User: What happened?

AI: I was driving down an empty road. I decided to slow down as a ball appeared on the right. I saw a child running towards the ball, so I decided to stop.

User: What would have happened if you did not stop ?

Future of Deeply Explainable Artificial Intelligence

User: What happened?

AI: I was driving down an empty road. I decided to slow down as a ball appeared on the right. I saw a child running towards the ball, so I decided to stop.

User: What would have happened if you did not stop ?

AI: If there was an impact, the child would have gotten hurt.

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2014).
Good practice in large-scale learning for image classification.
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Akata, Z., Perronnin, F., Harchaoui, Z., and Schmid, C. (2016).
Label-embedding for image classification.
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Akata, Z., Reed, S., Walter, D., Lee, H., and Schiele, B. (2015).
Evaluation of output embeddings for fine-grained image classification.
In *IEEE Computer Vision and Pattern Recognition (CVPR).*

Corona, R., Alaniz, S., and Akata, Z. (2019).
Modeling conceptual understanding in image reference games.
In *Neural Information Processing Systems (NeurIPS).*

Hendricks, L.-A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016).
Generating visual explanations.
In *European Conference of Computer Vision (ECCV).*

Hendricks, L. A., Hu, R., Darrell, T., and Akata, Z. (2018).
Grounding visual explanations.
In *European Conference of Computer Vision (ECCV).*

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. (2018).
Textual explanations for self driving vehicles.
In *European Conference of Computer Vision (ECCV).*

Reed, S., Akata, Z., Lee, H., and Schiele, B. (2016a).
Learning deep representations of fine-grained visual descriptions.
In *IEEE Computer Vision and Pattern Recognition (CVPR).*

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., and Lee, H. (2016b).
Generative adversarial text to image synthesis.
In *International Conference on Machine Learning (ICML)*.

Schoenfeld, E., Ebrahimi, S., Sinha, S., Darrell, T., and Akata, Z. (2019).
Generalized zero- and few-shot learning via aligned variational autoencoders.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xian, Y., Lampert, C., Schiele, B., and Akata, Z. (2018a).
Zero-shot learning- a comprehensive evaluation of the good, the bad and the ugly.
IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).

Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. (2018b).
Feature generating networks for zero-shot learning.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xian, Y., Sharma, S., Schiele, B., and Akata, Z. (2019).
F-vaegan-d2: A feature generating framework for any-shot learning.
In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Thank you!