Some plausible notes

Reporter: Yuren Cao

2018.11.30

The beginning of everything

- What's the difference between adversarial training between hard negative mining?
- Adversarial training: injects adversarial examples into training data to increase robustness[1]
- Hard negative mining: injects adversarial examples into training data to increase performance

[1] ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES ICLR 2018

Introduction

- Generation of adversarial examples
  - Methods: FGSM, C&W Attack, DeepFool, Zeroth Order Optimization…

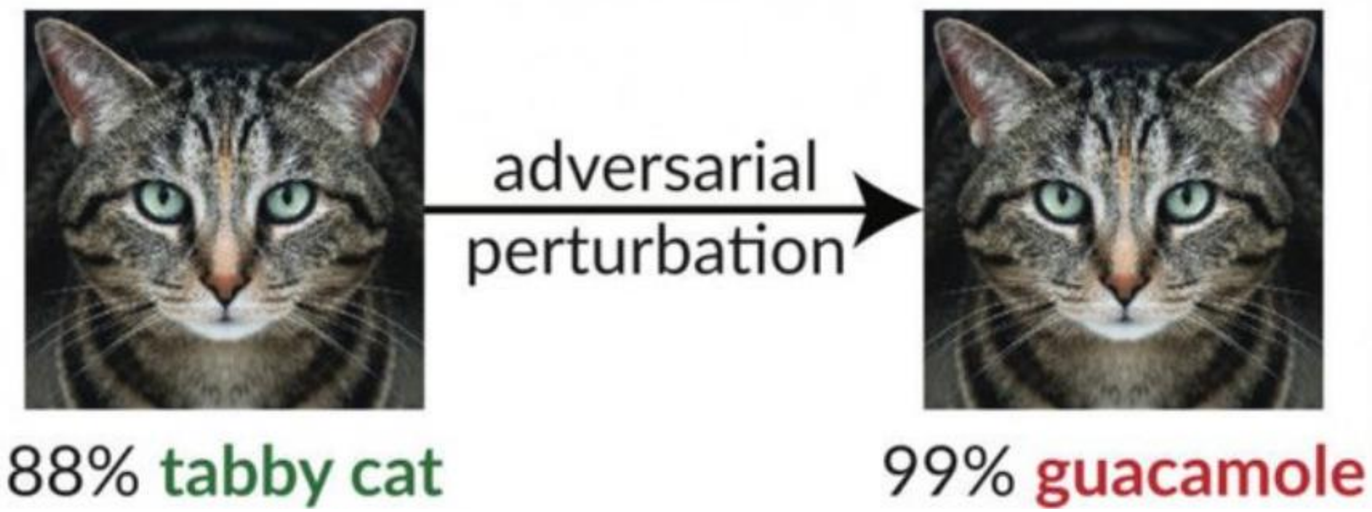


Table II: Taxonomy of Adversarial Examples

Attacks Methods	Adversarial Falsification	Adversary's Knowledge	Adversarial Specificity	Perturbation Scope	Perturbation Limitation	Attack Frequency	Perturbation Measurement	Datasets	Architectures
L-BFGS Attack [134]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, ImageNet, YoutubeDataset	AlexNet, QuocNet
Fast Gradient Sign Method (FGSM) [48]	False Negative	White-Box	Non-Targeted	Individual	N/A	One-time	element-wise	MNIST, ImageNet	GoogLeNet
Basic Iterative Method (BIM) and Iterative Least-Likely Class (ILLC) [75]	False Negative	White-Box	Non-Targeted	Individual	N/A	Iterative	element-wise	ImageNet	GoogLeNet
Jacobian-based Saliency Map Attack (JSMA) [101]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_2	MNIST	LeNet
DeepFool [97]	False Negative	White-Box	Non-Targeted	Individual	Optimized	Iterative	$\ell_p(p \in 1, \infty)$	MNIST, CIFAR10, ImageNet	LeNet, CaffeNet, GoogLeNet
CPPN EA Fool [99]	False Positive	White-Box	Non-Targeted	Individual	N/A	Iterative	N/A	MNIST, ImageNet	LeNet, AlexNet
C&W's Attack [27]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	$\ell_1, \ell_2, \ell_\infty$	MNIST, CIFAR10, ImageNet	GoogLeNet
Zeroth Order Optimization [30]	False Negative	Black-Box	Targeted & Non-Targeted	Individual	Optimized	Iterative	ℓ_2	CIFAR10, ImageNet	GoogLeNet
Universal Perturbation [96]	False Negative	White-Box	Non-Targeted	Universal	Optimized	Iterative	$\ell_p(p \in 1, \infty)$	ImageNet	CaffeNet, VGG, GoogLeNet, ResNet
Feature Adversary [115]	False Negative	White-Box	Targeted	Individual	Constraint	Iterative	ℓ_2	ImageNet	CaffeNet, VGG, AlexNet, GoogLeNet
Hot/Cold [113]	False Negative	White-Box	Targeted	Individual	Optimized & Constraint	One-time	PASS	MNIST, ImageNet	LeNet, GoogLeNet, ResNet
Natural GAN [147]	False Negative	Black-Box	Non-targeted	Individual	Optimized	Iterative	ℓ_2	MNIST, LSUN, SNLI	LeNet, LSTM, TreeLSTM
Model-based Ensembling Attack [86]	False Negative	White-Box	Targeted & Non-Targeted	Individual	Constraint	Iterative	ℓ_2	ImageNet	VGG, GoogLeNet, ResNet
Ground-Truth Attack [23]	False Negative	White-Box	Targeted	Individual	Optimized	Iterative	ℓ_1, ℓ_∞	MNIST	3-layer FC

Introduction

- Focal loss:
  - a method of hard negative mining

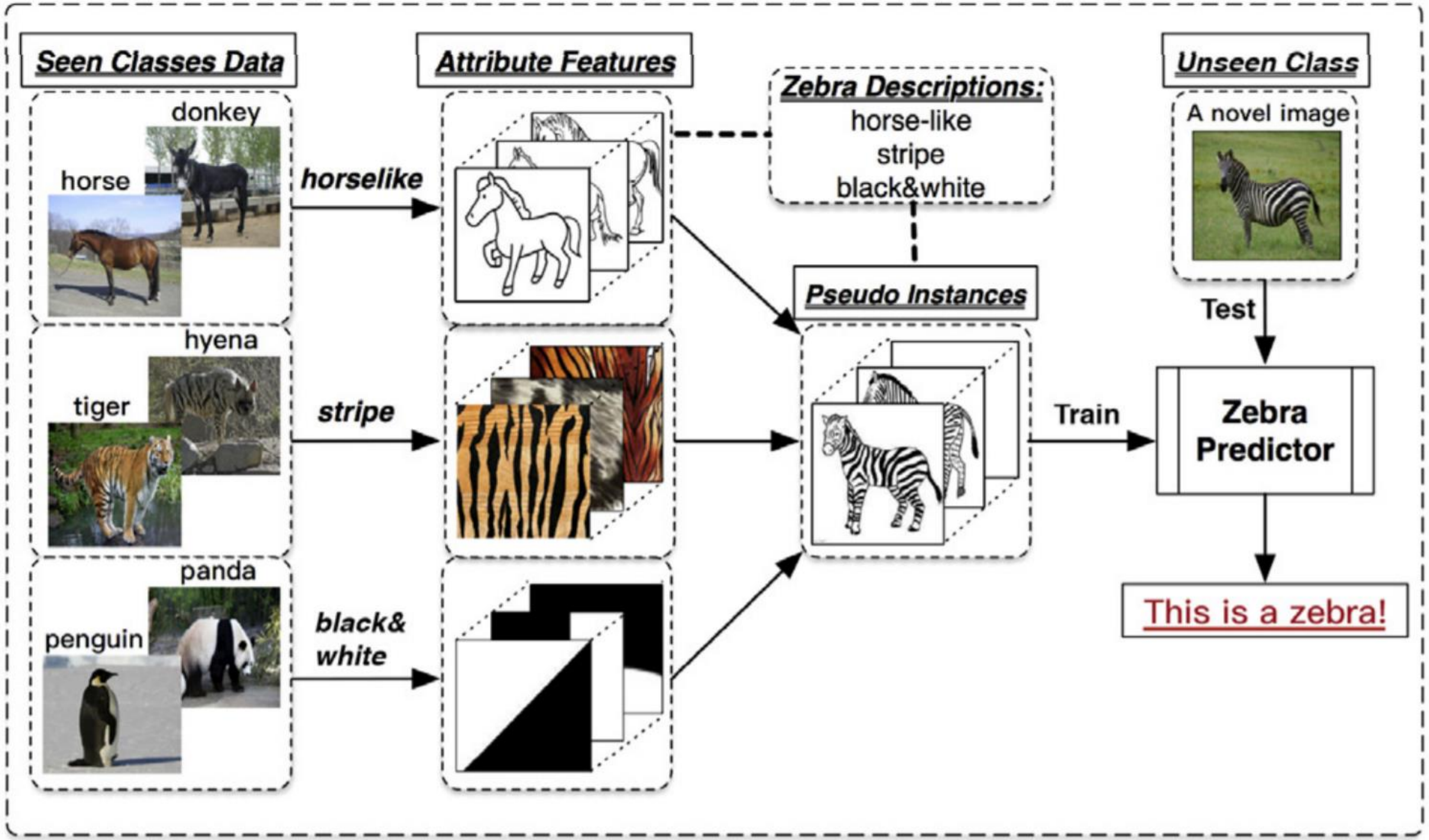
$$p_{\mathrm{t}} = \begin{cases} p & \text{if } y = 1 \\ 1 - p & \text{otherwise,} \end{cases}$$

$$\mathrm{CE}(p, y) = \mathrm{CE}(p_{\mathrm{t}}) = -\log(p_{\mathrm{t}}).$$

$$\mathrm{FL}(p_{\mathrm{t}}) = -(1 - p_{\mathrm{t}})^{\gamma} \log(p_{\mathrm{t}}).$$

Some provoking works

➢ Attribute based model



Some provoking works

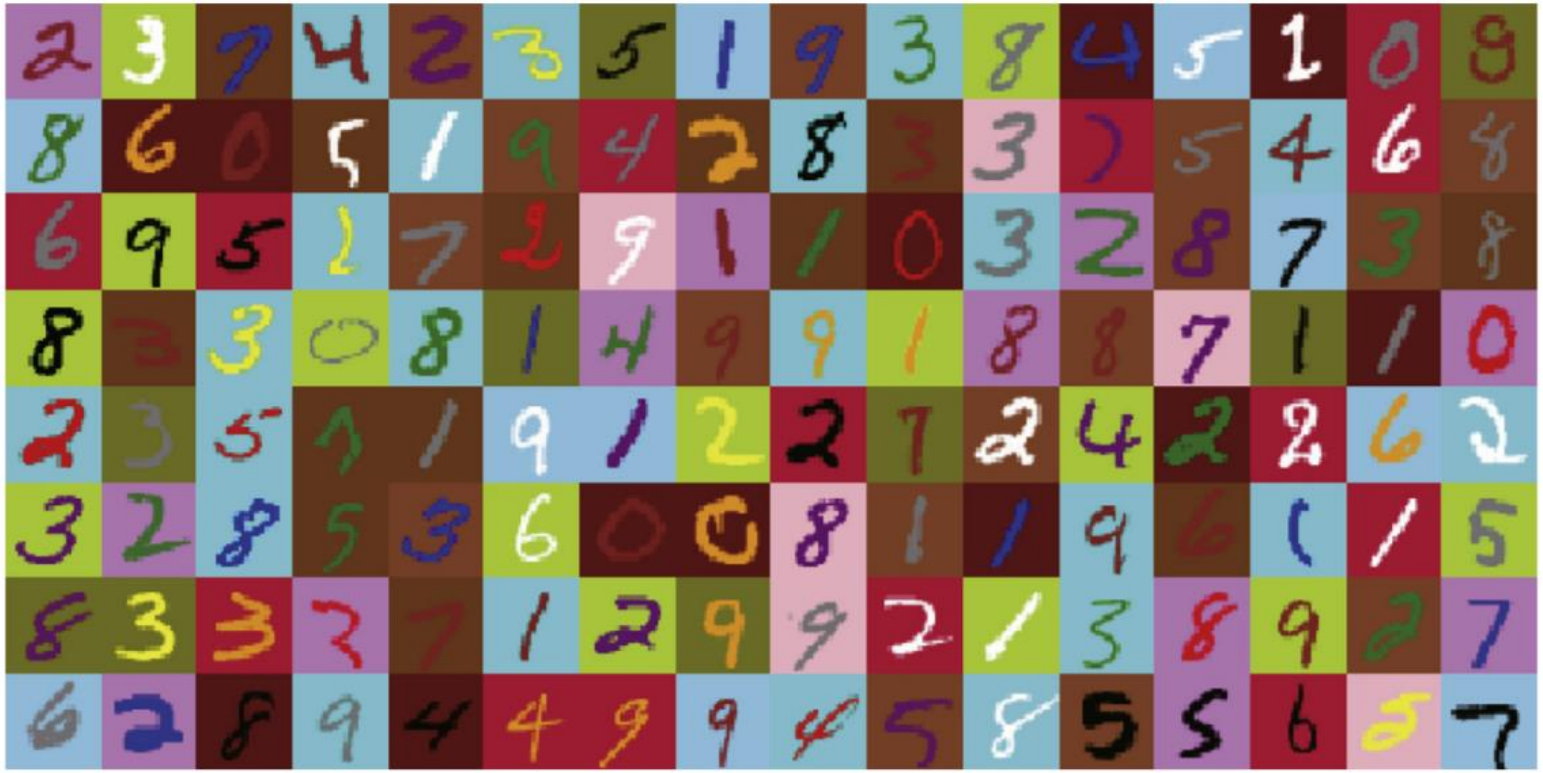
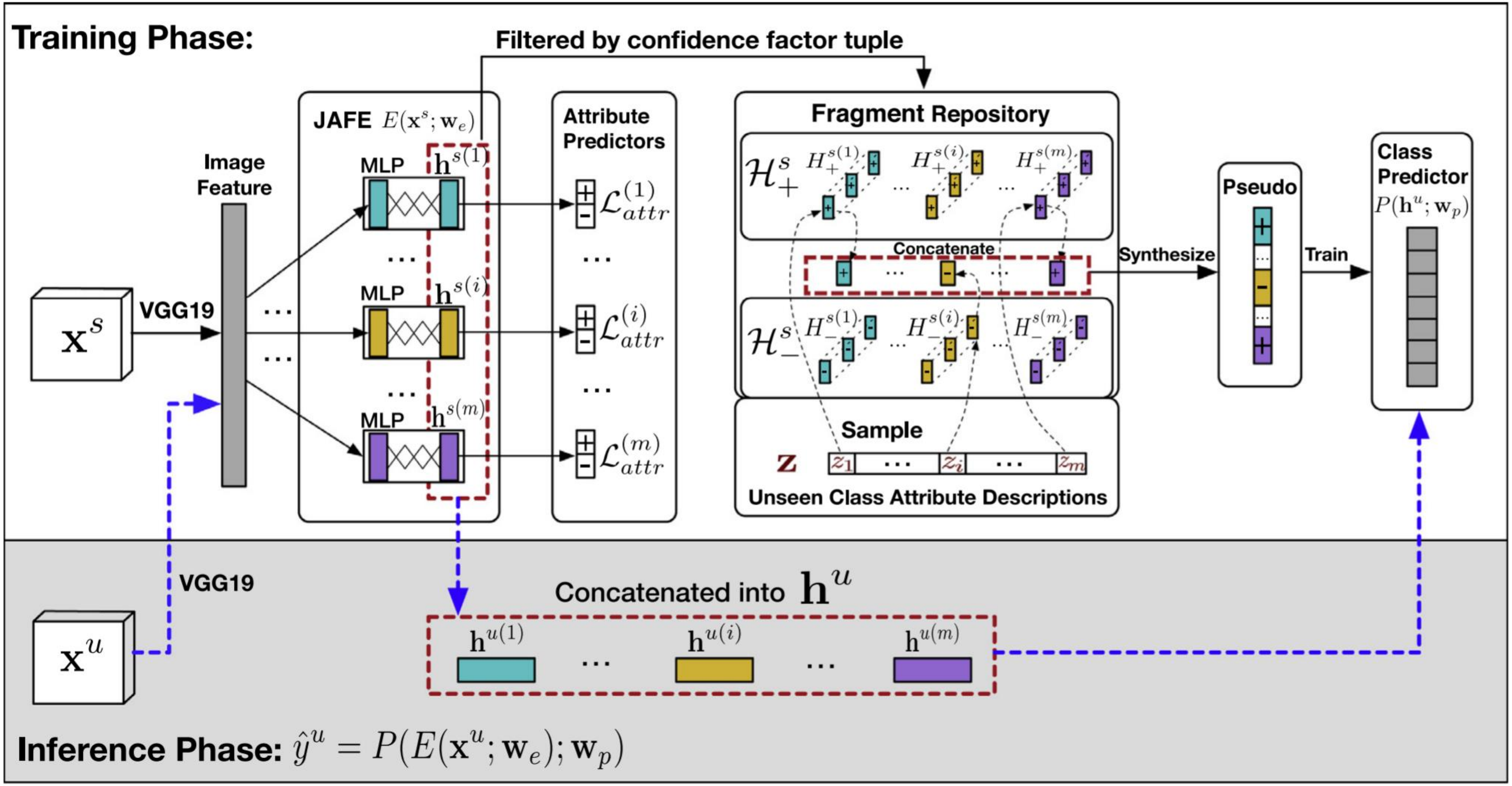


Fig. 6. Some examples of C-MNIST. The images from same class own the same digit, b-color and f-color. The size per class is almost 70k/1k=70. Best viewed in color.

Jiang Lu et al. Attribute-Based Synthetic Network (ABS-Net): Learning more from pseudo feature representations



Jiang Lu et al. Attribute-Based Synthetic Network (ABS-Net): Learning more from pseudo feature representations

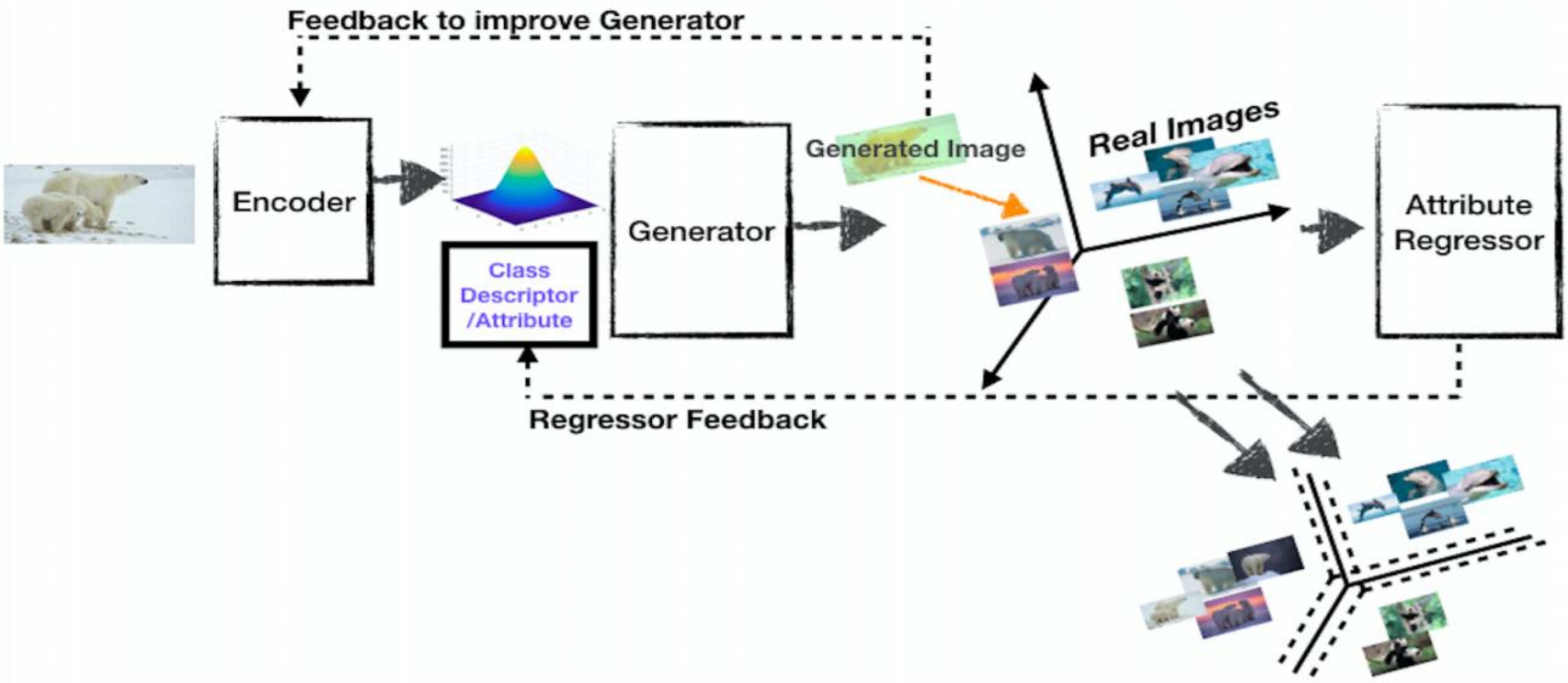
How & Why it works?

- Get a scope for whole dataset
  - Include seen & unseen data
  - Set attributes for data representation
  - Generate unseen data
  - Reconstruct training dataset

Means that the we know and let model know the distribution of whole dataset(seen + synthetic)

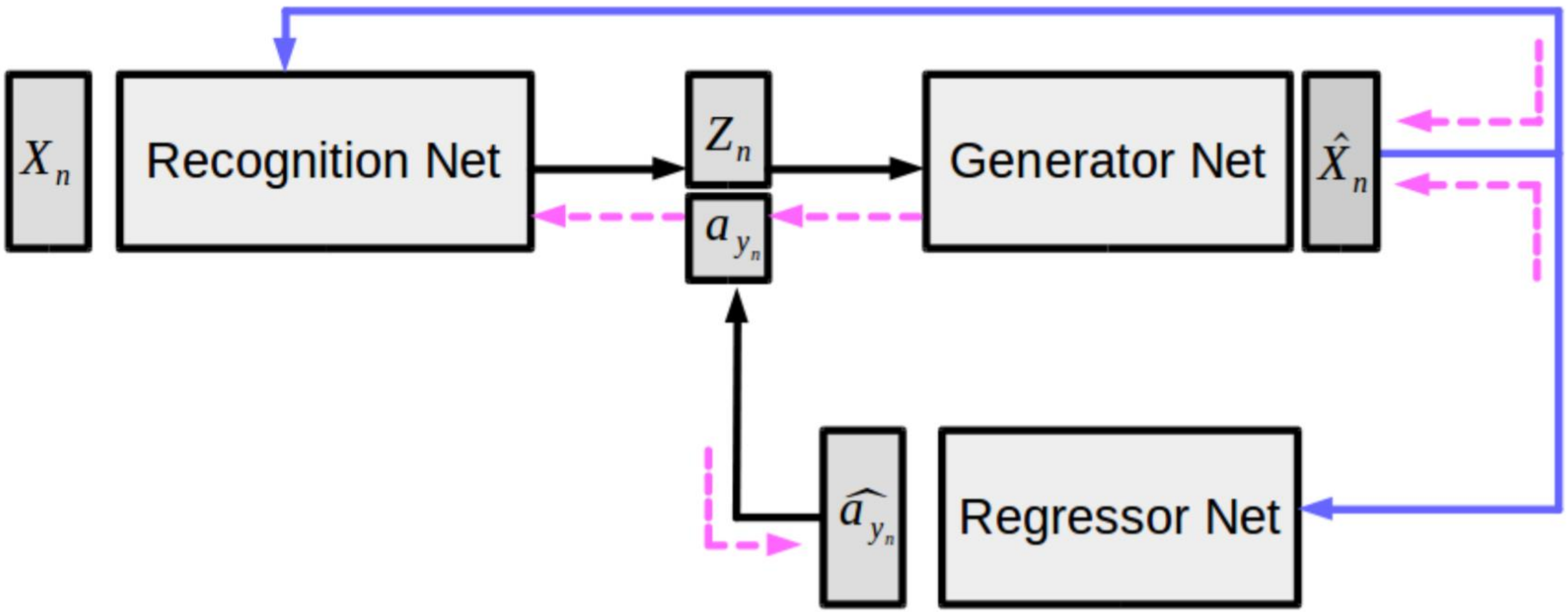
Jiang Lu et al. Attribute-Based Synthetic Network (ABS-Net): Learning more from pseudo feature representations

Some provoking works



Vinay Kumar Verma et al. Generalized Zero-Shot Learning via Synthesized Examples CVPR 2018

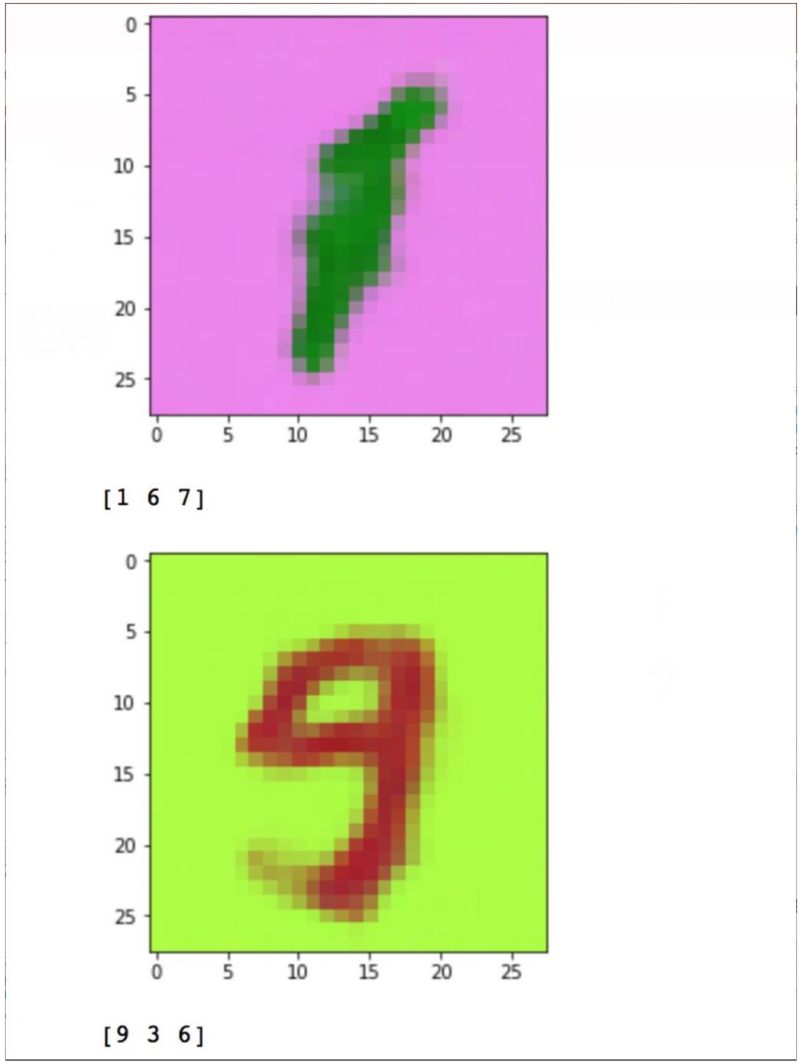
Some provoking works



Vinay Kumar Verma et al. Generalized Zero-Shot Learning via Synthesized Examples CVPR 2018

Perfect, in a despairing way

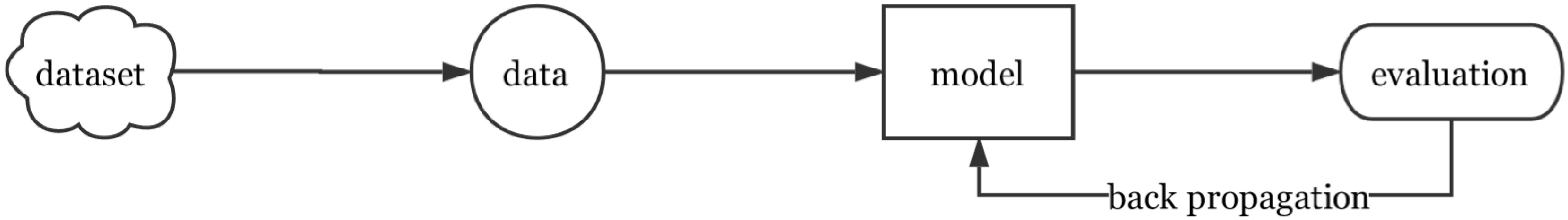
- Experiment Setting:
  - Training data: 0-4 red & 5-9 green
  - Target data: 0-4 green & 5-9 red
  - Extra: binary attributes vector
  - Model: ONLY CVAE (2-3 layer Convs)
  - Epochs: 5000
  - AUC : 0.997



Experiment Joshua L

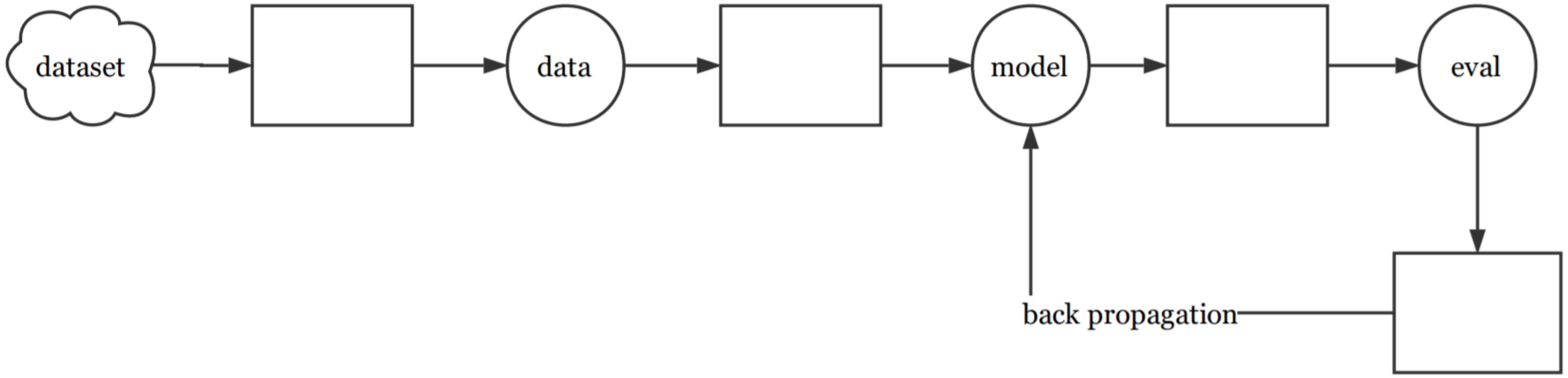
Some bricks for jade

- There is a general training phase in supervised learning

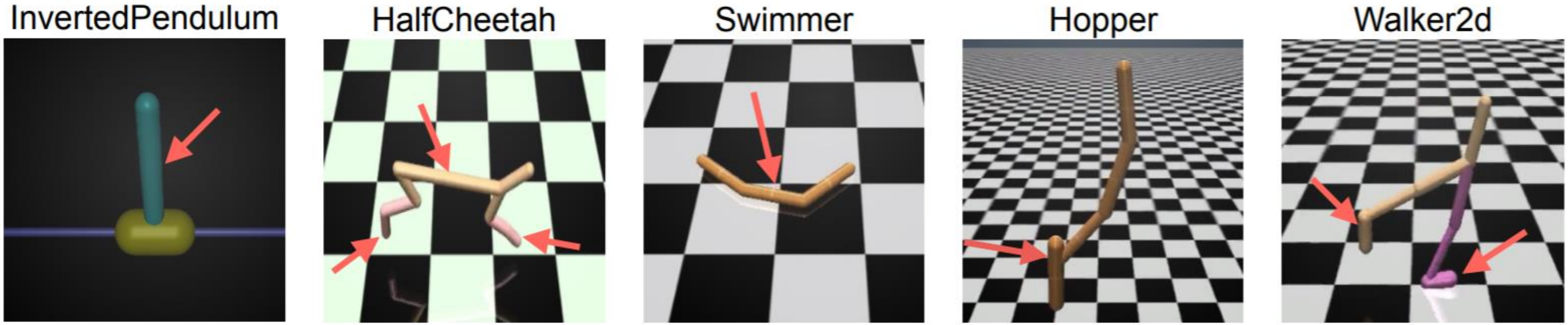


Some bricks for jade

- But if we do whatever we want, it can be:



A twice-told story:
Robust Adversarial RL（RARL）



Lerrel Pinto et al. (CMU Google Brain). ICML 2017

Two Routes

- Real-World Policy Learning

Easy to design Better Matching Data scarcity(Expensive Dangerous Time-intensive) Hard to generalize

- Learning in simulation

Easy to transfer Sufficient data Additional gap(Between environment and simulator) Hard to design

BOTH: Influenced by uncertainty Data-intensity

Design

- Goal

① Model the gap between simulations and real-world

② Learn a policy robust to all uncertainties

- Eureka !

Modeling errors can be viewed as extra forces(->disturbances) in the system(e.g. friction)

-> Representation: Adversarial agent

A smart way to extend information without understanding itself

Design

- Agent

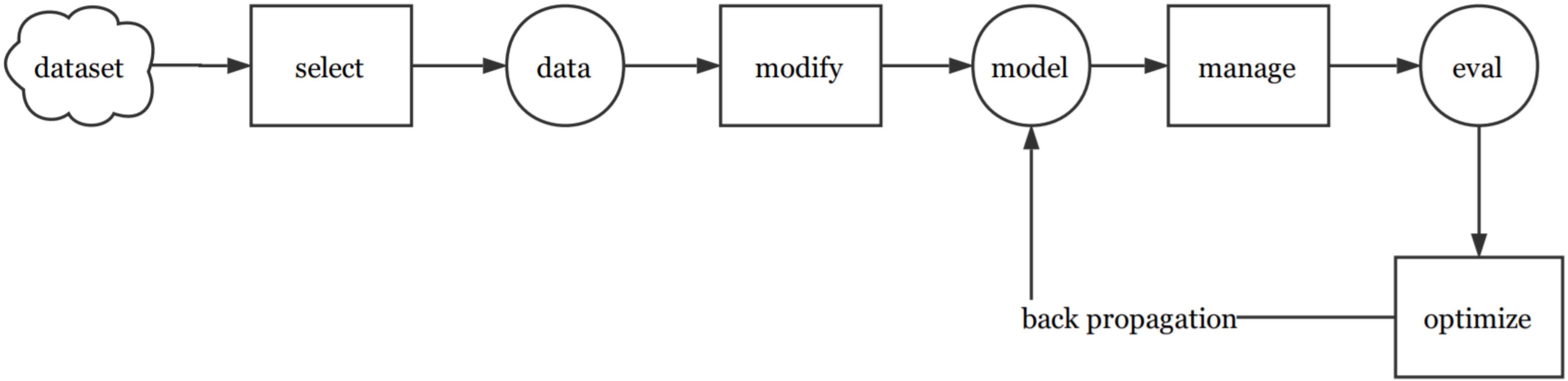
① Protagonist Agent

② Adversarial Agent

- Reward
  - The Protagonist Agent is rewarded for fulfil the original task goals
  - The Adversarial Agent is rewarded for the failure of the protagonist

Continuing…

➢ To name the processes in the diagram:



A game: Gods homing

➢ Many Adversarial tricks:

For a given target model:

- Use a method(e.g. FGSM) to get adversarial examples
- Put these adversarial examples into training set

✓ Target model get better performance

Step: Modify

A game: Gods homing

- Many GANs:
- Set a generator & a discriminator
- To get lots of new, fake data
- Put them(with/without originals) in target model(possibly, D)
- ✓ For G: get higher quality generated samples
- ✓ For D: target model get better performance

Step: Modify

A game: Gods homing

- Some AEs:

- Set an encoder & a decoder
- Let encoder learn the transformation: from data to latent representation
- Let decoder learn the transformation: from latent feature to required representation

✓ Do: conditional generation, data augmentation…

Step: Modify

Others

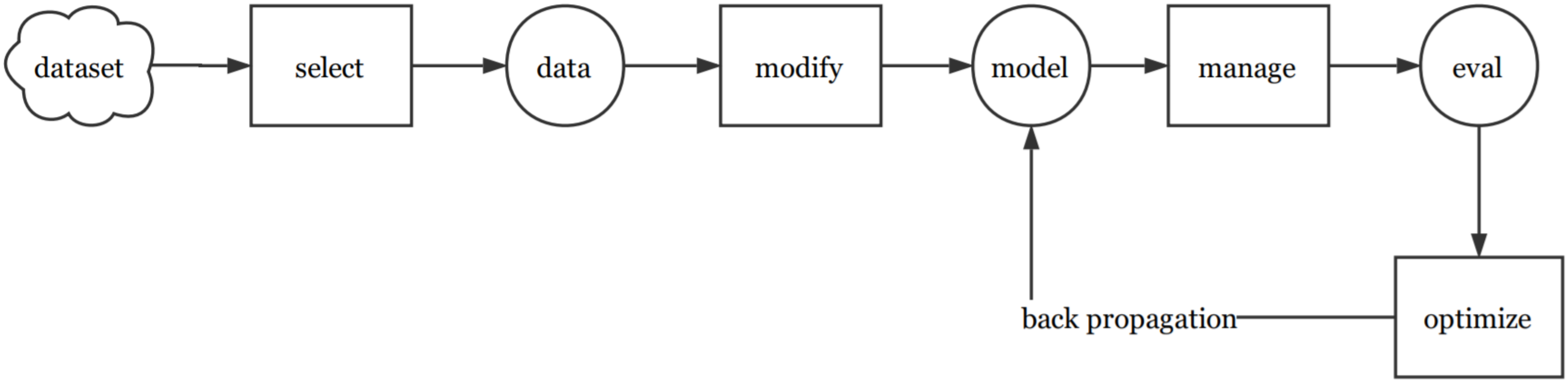
- Many loss tricks: e.g. Regularization methods
  - Step: Evaluation
- Many Active learning methods: e.g. Gradient based methods
  - Step: Optimization(Evaluation)
- Some novel models: e.g. Residual based structure
  - Step: Model
- Boosting , Bootstrap …

  -> Semi-Supervised, Unsupervised, Reinforcement learning …

Another perspective

Imprecisely: they are in different processes

- Hard negative mining: in manage/eval step
- Adversarial training: in modify step



Questions last year (Model-Example-Method)

- Evaluate: Data

Quality – Quantity: For a dataset (+Coverage)

High quality/quantity -> better performance? (respond to gaps)

- Another Vision:

Real – Fake:

Real data is better than fake data? (No direct relationship)

- Same criterion for majority tasks? (Specific better, usually)

Know more about data

- Information of dataset
  - All classes (seen in dataset + unseen in real)
    - e.g. attributes based
  - Distributions
    - e.g. imbalance(quality/quantity), “mixup”
  - Gaps
    - e.g. Train – Test, Test – Real, Train itself

Doing more about data to promote

- Provide more information about task
  - Attribute based model
  - Additional "unrelated" information
  - …
- Treat former steps as a "generalized model"
  - Use generative model to "make" data
  - Cut or modify data if we believe it may be benefit for target model
- Generalized Meta Learning?

Reference

- Attribute-Based Synthetic Network (ABS-Net): Learning more from pseudo feature representations. Jiang Lu et al. Pattern Recognition. Volume 80
- Generalized Zero-Shot Learning via Synthesized Examples. Vinay Kumar Verma et al. CVPR 2018
- ENSEMBLE ADVERSARIAL TRAINING: ATTACKS AND DEFENSES. Florian Tramer et al. ICLR 2018
- Focal Loss for Dense Object Detection. Tsung-Yi Lin et al. ICCV 2017
- Meta learning: a survey of trends and technologies. Christiane Lemke et al. 2013

Reference

- Robust Adversarial Reinforcement Learning Lerrel Pinto et al. ICML 2017
- mixup: Beyond Empirical Risk Minimization Hongyi Zhang et al. ICLR 2018

Thank you!