Reducing the need for large labeled dataset in the learning to learn framework

Ph.D. Project Proposal

Arman Afrasiyabi
Université Laval
Arman.afrasiyabi.1@ulaval.ca

Content

- Chapter I) Introduction
  - K-shot learning
  - Meta learning
  - Prototypical Neural Network
- Chapter II) Extending k-shot to k-plus shot
  - Preliminary work
  - Zero-Shot learning
  - Adding Decoder
- Chapter III) Interactive fast adaptation
  - Motivation
  - Active Meta-Learning
  - Memory-augmented model for quick adaptation
  - Proposal
- Chapter IV) Reducing Annotation Amount in Visual Learning
  - Motivation Active
  - Proposal

Content

- Chapter I) Introduction
    - K-shot learning
    - Meta learning
    - Prototypical Neural Network
- Chapter II) Extending k-shot to k-plus shot
    - Preliminary work
    - Zero-Shot learning
    - Adding Decoder
- Chapter III) Interactive fast adaptation
    - Motivation
    - Active Meta-Learning
    - Memory-augmented model for quick adaptation
    - Proposal
- Chapter IV) Reducing Annotation Amount in Visual Learning
    - Motivation Active
    - Proposal

Motivation

Deep learning is showing outstanding performances in different areas ...

Problem:

we need many examples for training

not aligned with the learning process in the human level

(Lake, Salakhutdinov, and Tenenbaum, 2015)

Terminology

Deep Learning

K-shot Learning

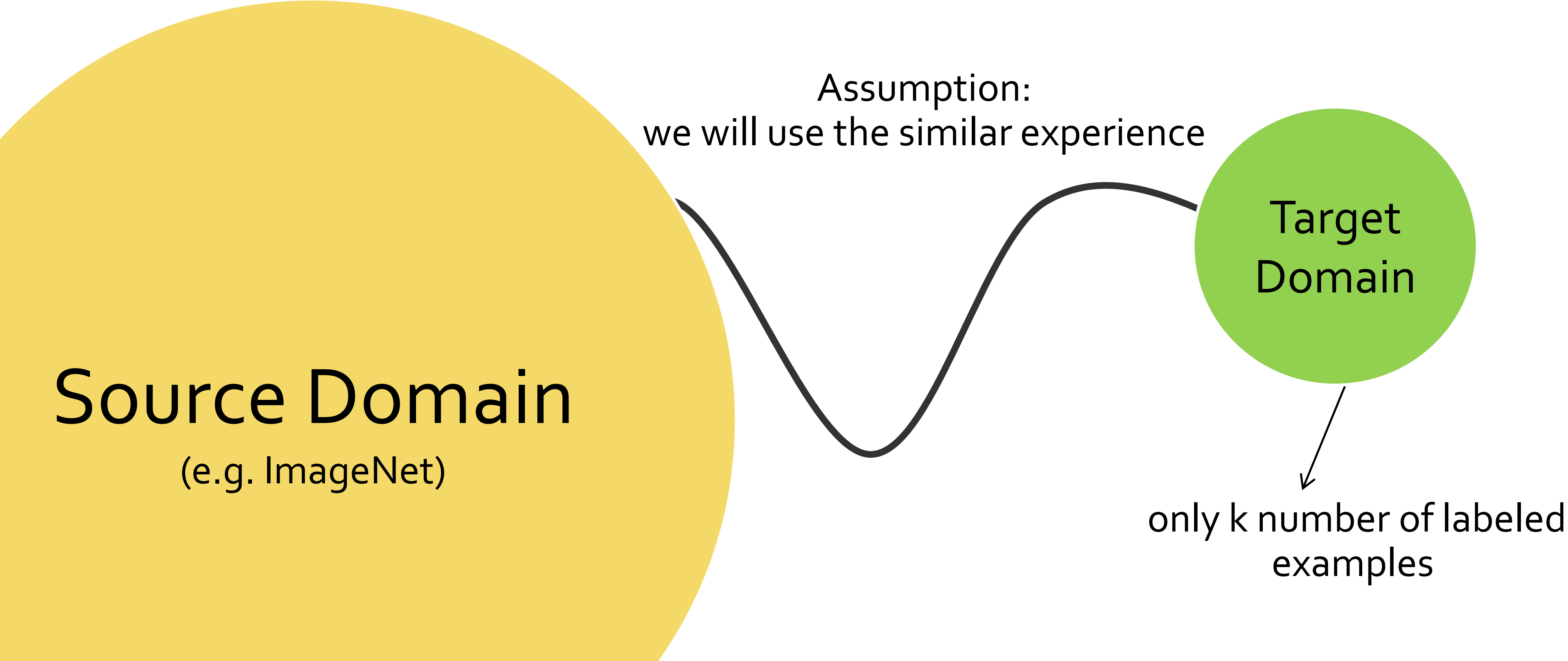
Meta Learning

Distance based Learning

Transferring the knowledge

K-shot learning: is supervised transferring knowledge

using previous experience



Transfer Learning by fine-tune????

Problem

CS231n Convolutional Neural Networks for Visual Recognition

(These notes are currently in draft form and under development)

Table of Contents:

- Transfer Learning
- Additional References

1. *New dataset is small and similar to original dataset.* Since the data is small, it is not a good idea to fine-tune the ConvNet due to overfitting concerns. Since the data is similar to the original data, we expect higher-level features in the ConvNet to be relevant to this dataset as well. Hence, the best idea might be to train a linear classifier on the CNN codes.
2. *New dataset is large and similar to the original dataset.* Since we have more data, we can have more confidence that we won't overfit if we were to try to fine-tune through the full network.
3. *New dataset is small but very different from the original dataset.* Since the data is small, it is likely best to only train a linear classifier. Since the dataset is very different, it might not be best to train the classifier form the top of the network, which contains more dataset-specific features. Instead, it might work better to train the SVM classifier from activations somewhere earlier in the network.
4. *New dataset is large and very different from the original dataset.* Since the dataset is very large, we may expect that we can afford to train a ConvNet from scratch. However, in practice it is very often still beneficial to initialize with weights from a pretrained model. In this case, we would have enough data and confidence to fine-tune through the entire network.

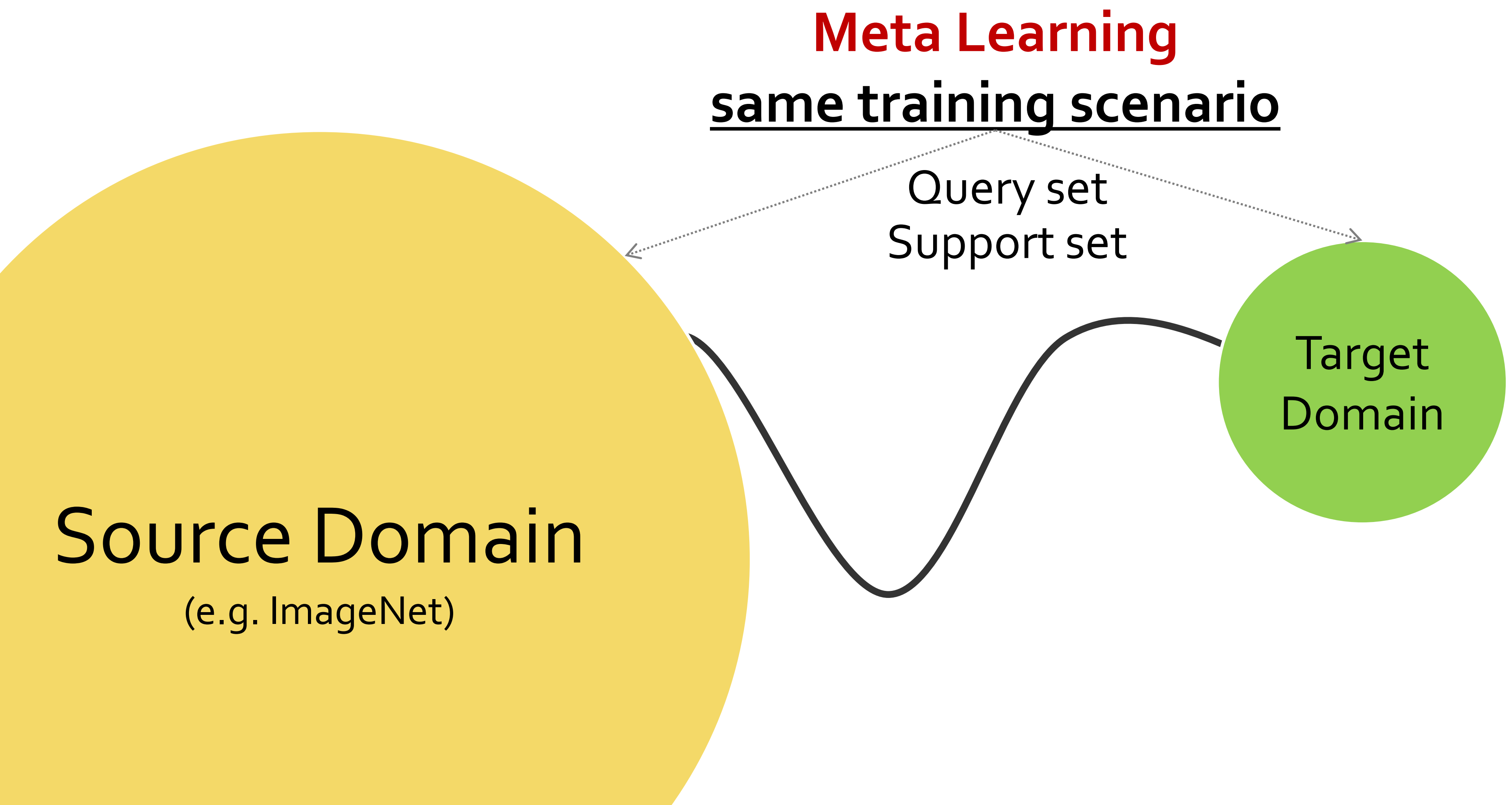
Gradient Descent Challenge

(k examples might not enough; overfitting problem

Three approaches are proposed:

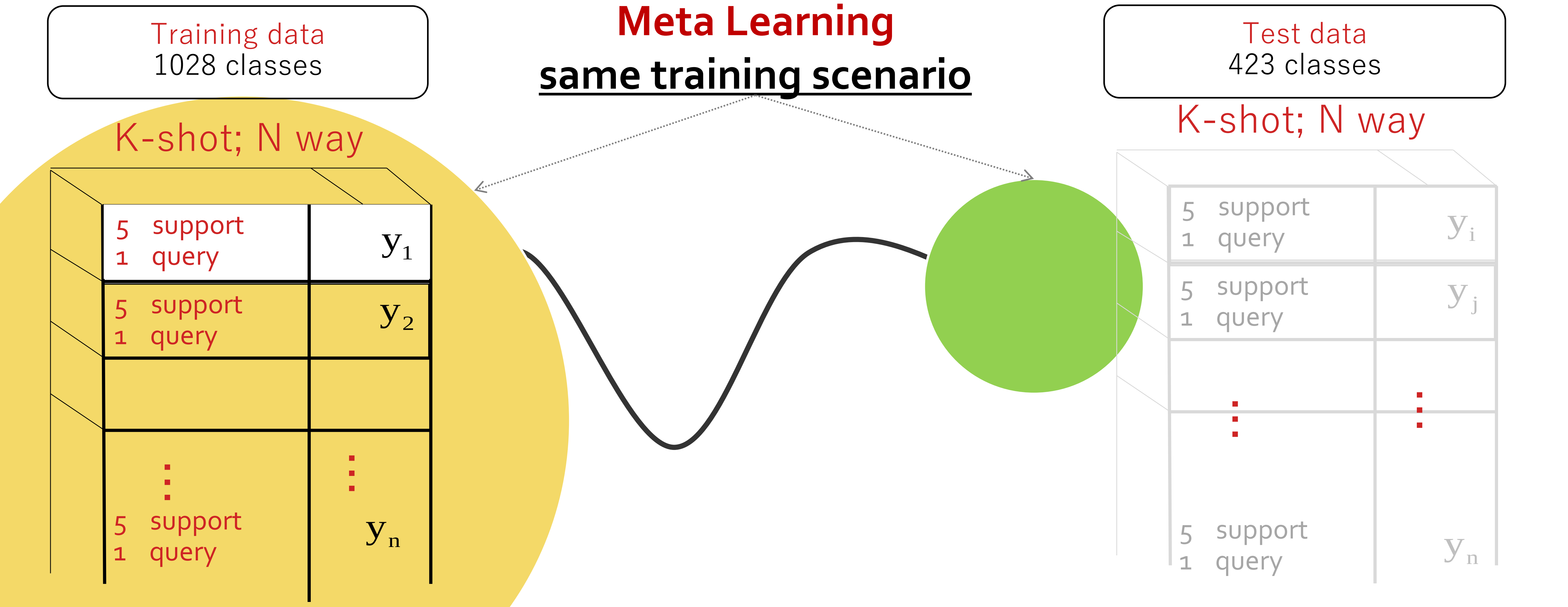
- Distance based:
  - Matching networks (Vinyals et al. 2016)
  - Prototypical Networks (Snell et al. 2017)
- Gradient descent based:
  - Meta-Learner LSTM (Ravi & Larochelle, 2017)
  - MAML (Finn et al. 2017)
- Black box neural network
  - MANN (Santoro et al. 2016)
  - SNAIL (Mishra et al. 2018)

Learning to learn or Meta-learning: a framework for k-shot learning



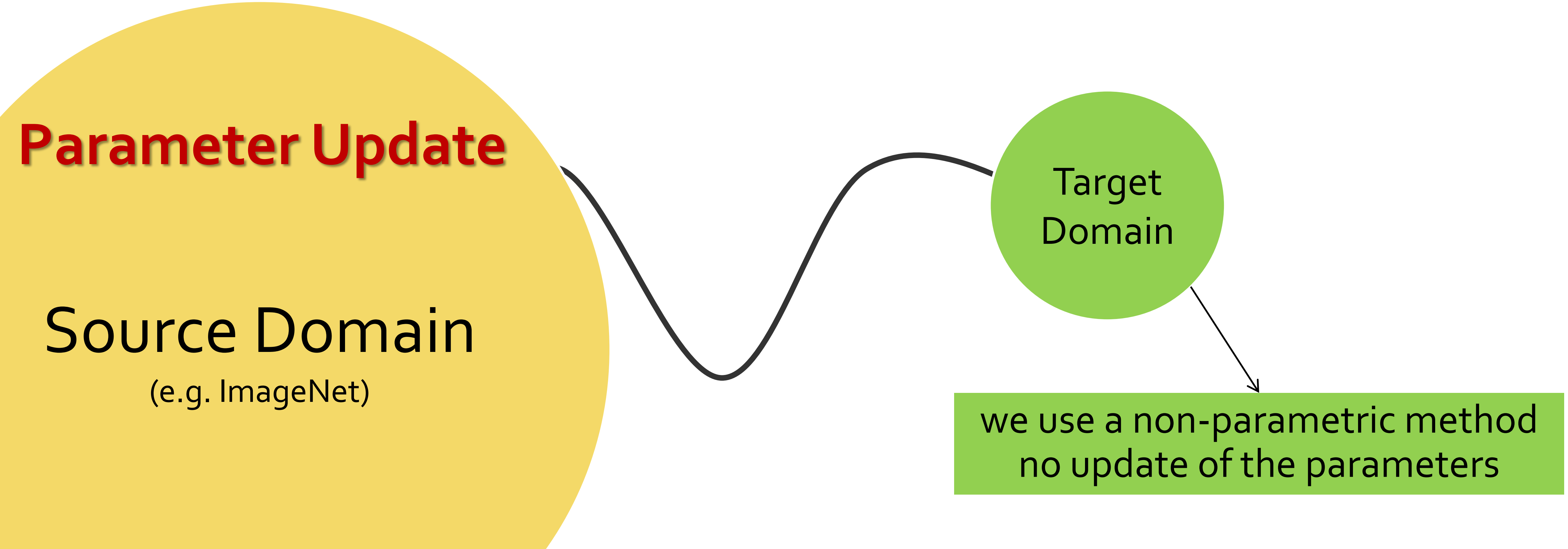
In source domain training, we pretend to have only "k" example in each iteration. (Like target domain)

Learning to learn or Meta-learning: a framework for k-shot learning



Distance based approaches

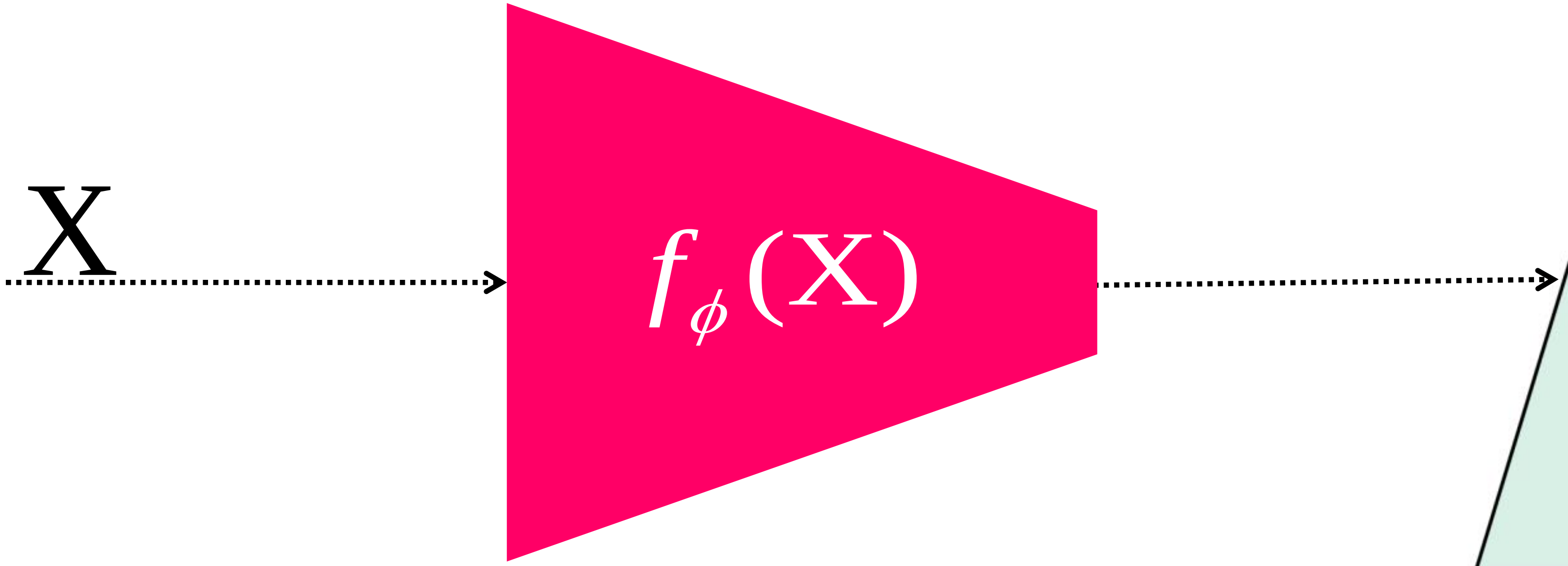
- Matching networks (Vinyals et al. 2016)
- Prototypical Networks (Snell et al. 2017)



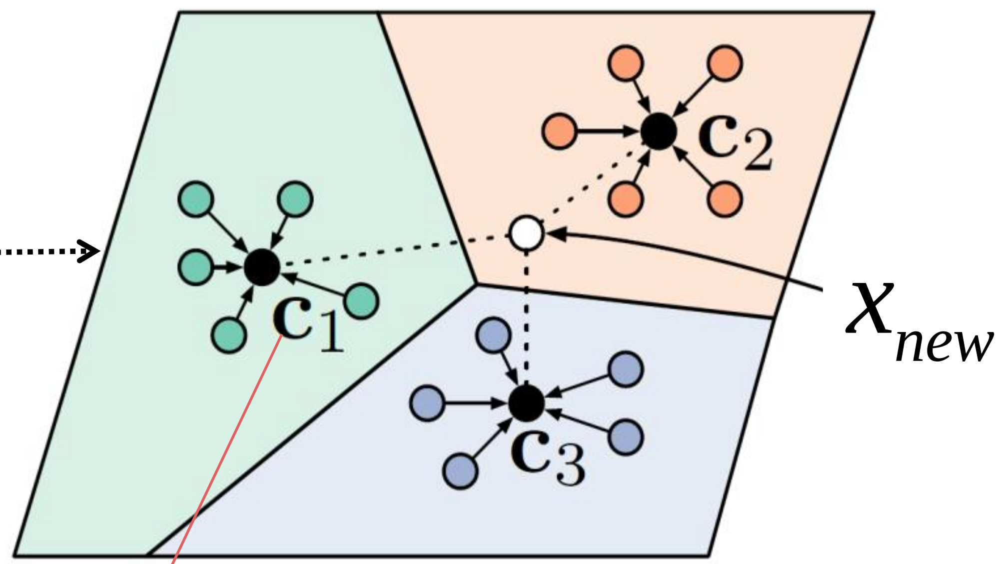
Prototypical Network [Snell et al. NIPS2017)]

Distance based approach (e.g.)

Snell, Swersky, and Zemel, 2017



Embedding space



(a) Few-shot

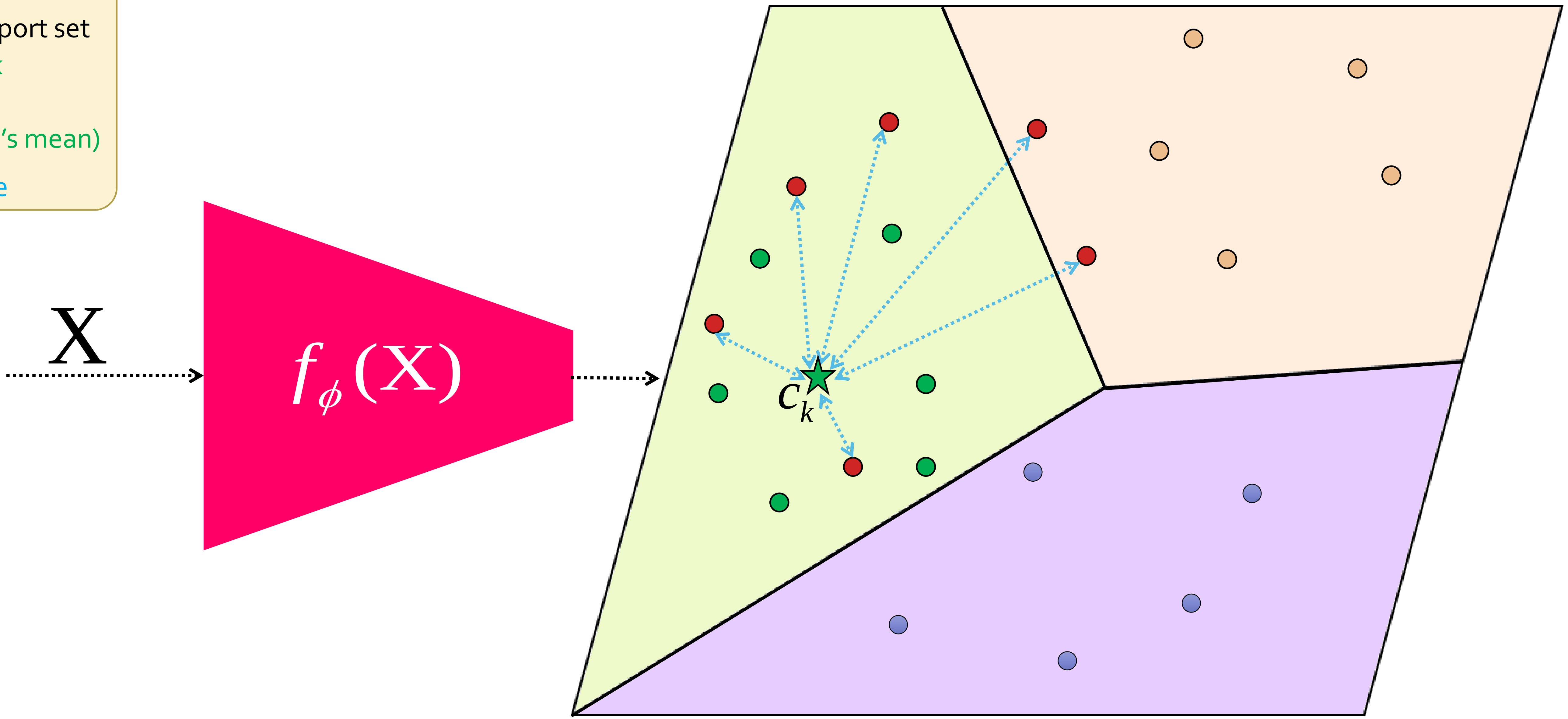
$$\boxed{c_k = \frac{1}{|S_k|} \sum_{(x_i, y_i) \in S_k} f_\phi(x_i)}$$

Prototypical Network [Snell et al. NIPS2017)]

Distance based approach (e.g.)

Ch.I

other classes support set
support set class k
query set class k
proto (support set's mean)
Euclidean distance



(k examples might not enough; overfitting problem)

Three approaches are proposed:

- Distance based:
  - Matching networks (Vinyals et al. 2016)
  - Prototypical Networks (Snell et al. 2017)
- Gradient descent based:
  - Meta-Learner LSTM (Ravi & Larochelle, 2017)
  - MAML (Finn et al. 2017)
- Black box neural network
  - MANN (Santoro et al. 2016)
  - SNAIL (Mishra et al. 2018)

Extending:

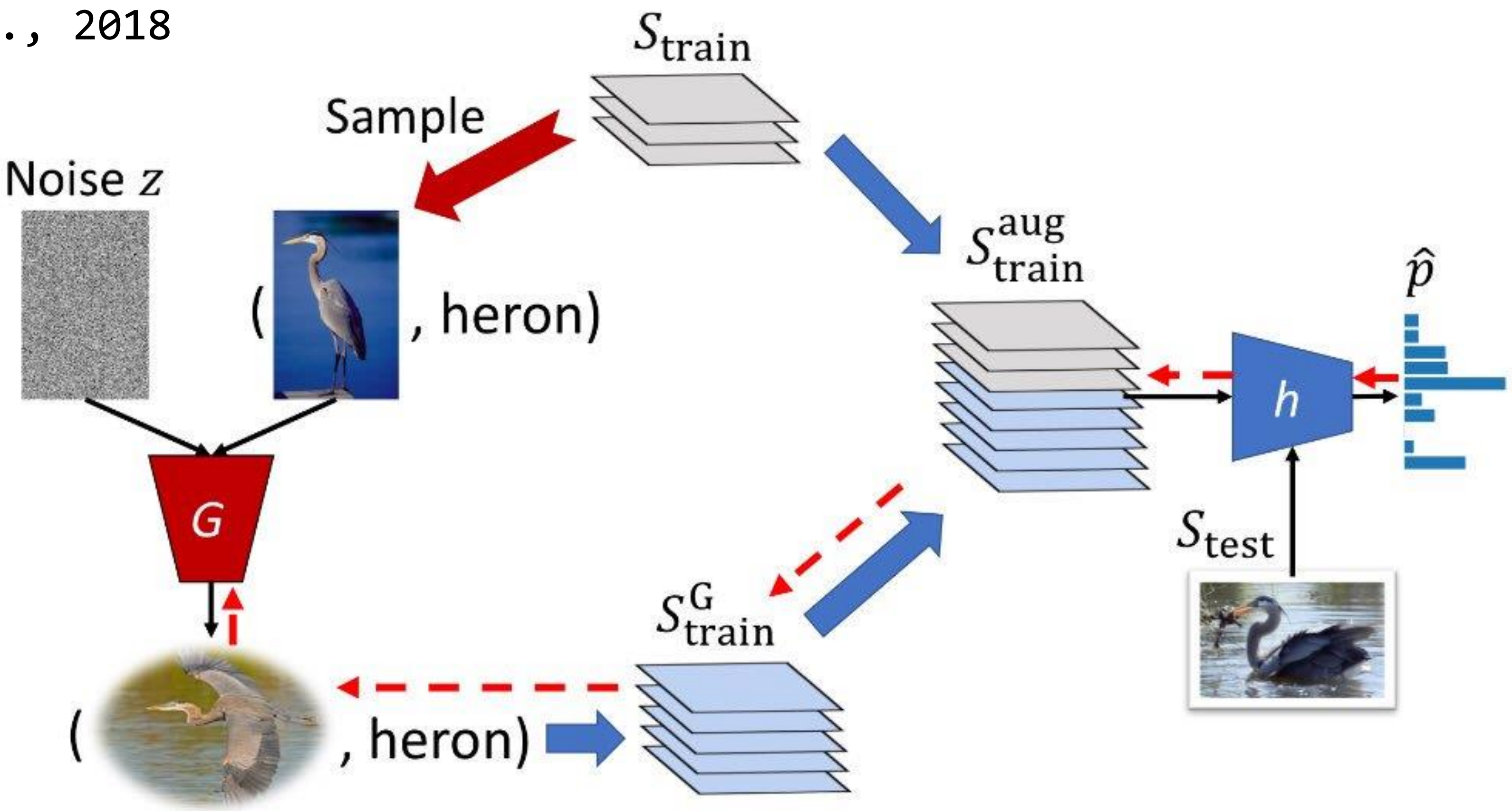
- Semi supervised learning
- Active Learning
- Data augmentation

Content

- Chapter I) Introduction
    - K-shot learning
    - Meta learning
    - Prototypical Neural Network
- Chapter II) Extending k-shot to k-plus shot
    - Preliminary work
    - Zero-Shot learning
    - Adding Decoder
- Chapter III) Interactive fast adaptation
    - Motivation
    - Active Meta-Learning
    - Memory-augmented model for quick adaptation
    - Proposal
- Chapter IV) Reducing Annotation Amount in Visual Learning
    - Motivation Active
    - Proposal

Extending Distance based Approach

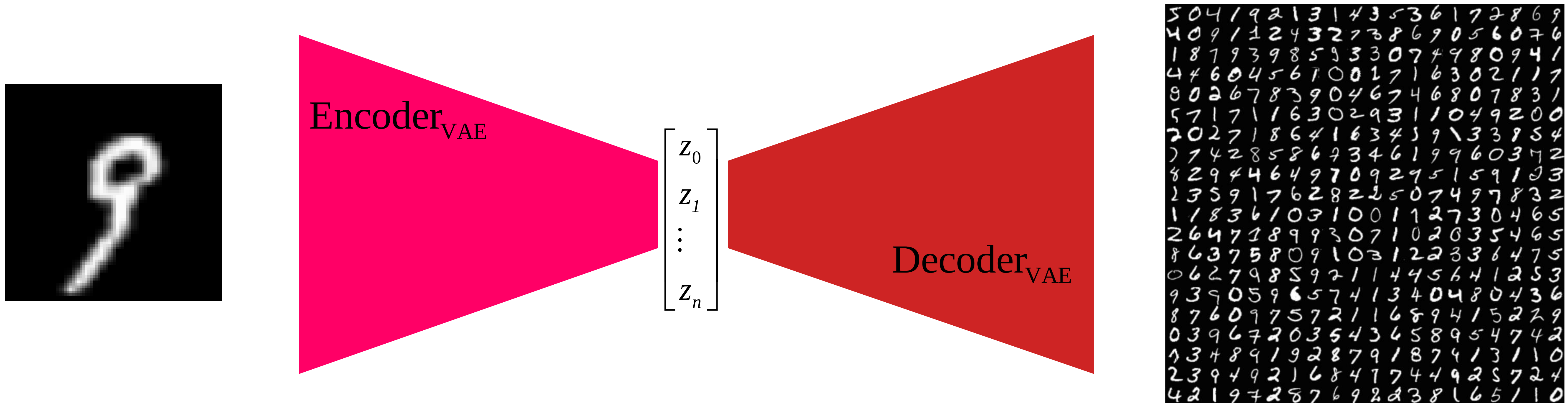
Wang et al., 2018



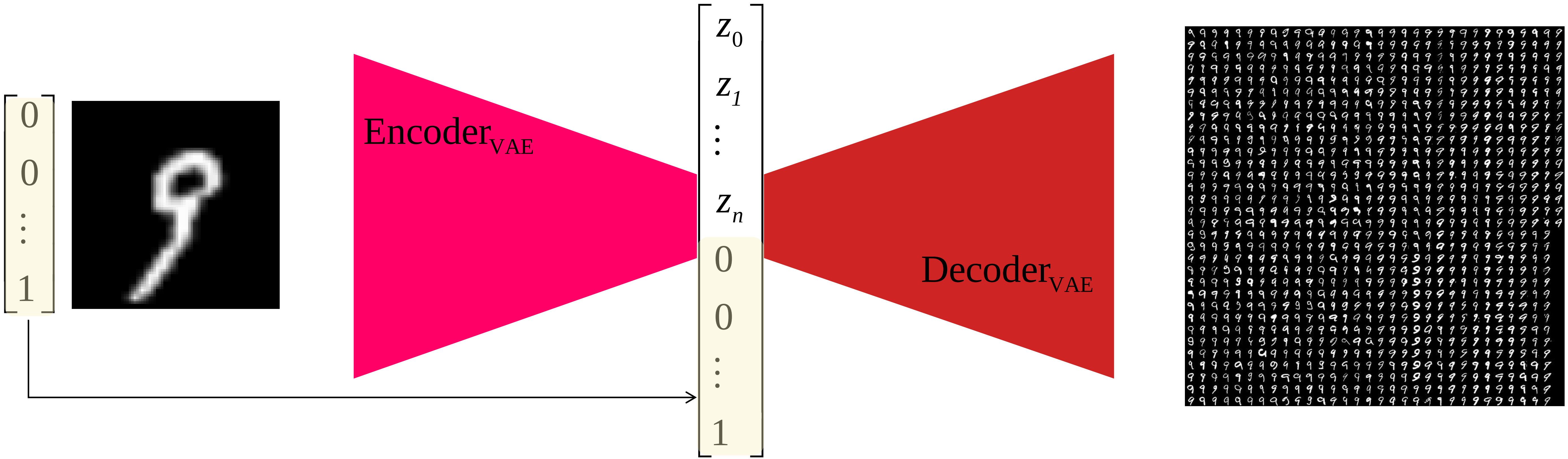
Unrealistic images in the image space???????

Why **NOT** realistic data augmentation?

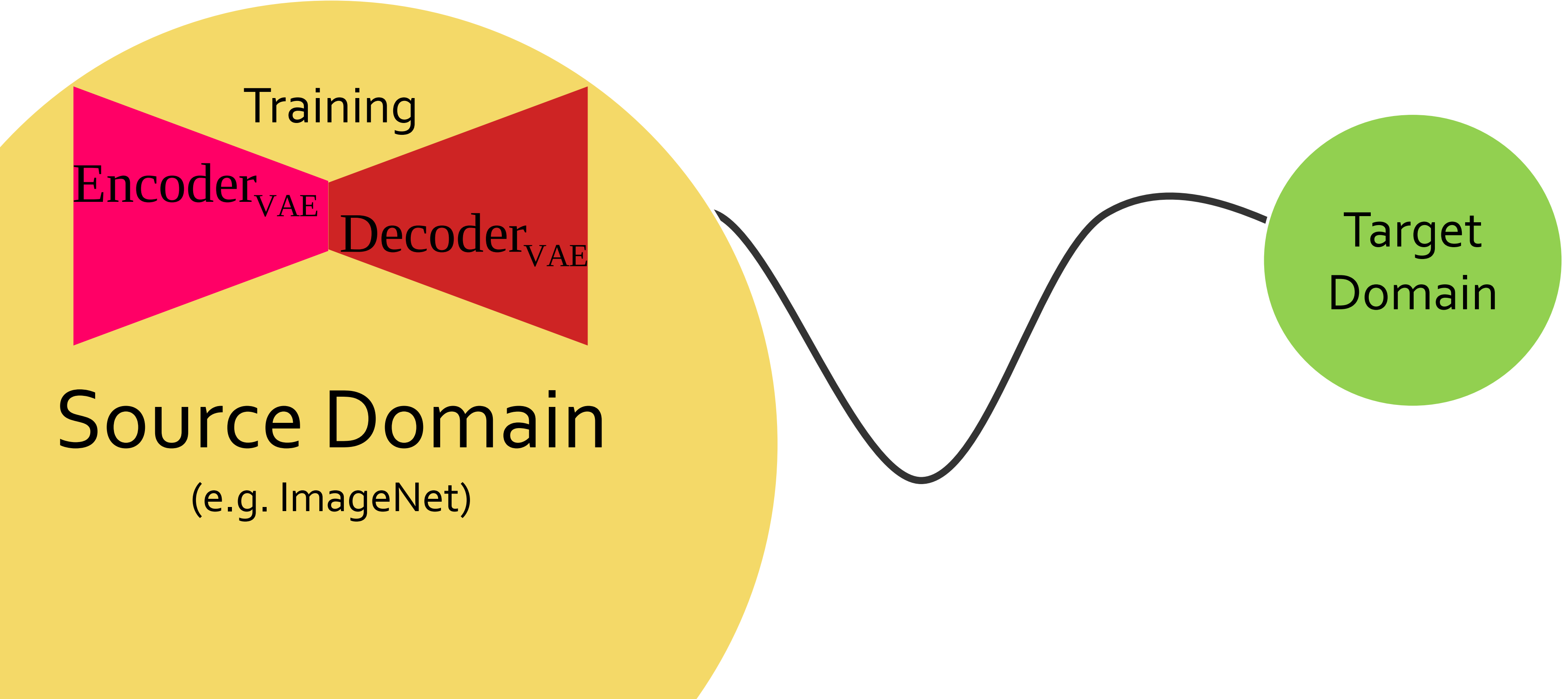
Small challenge: Supervised Generative Models!



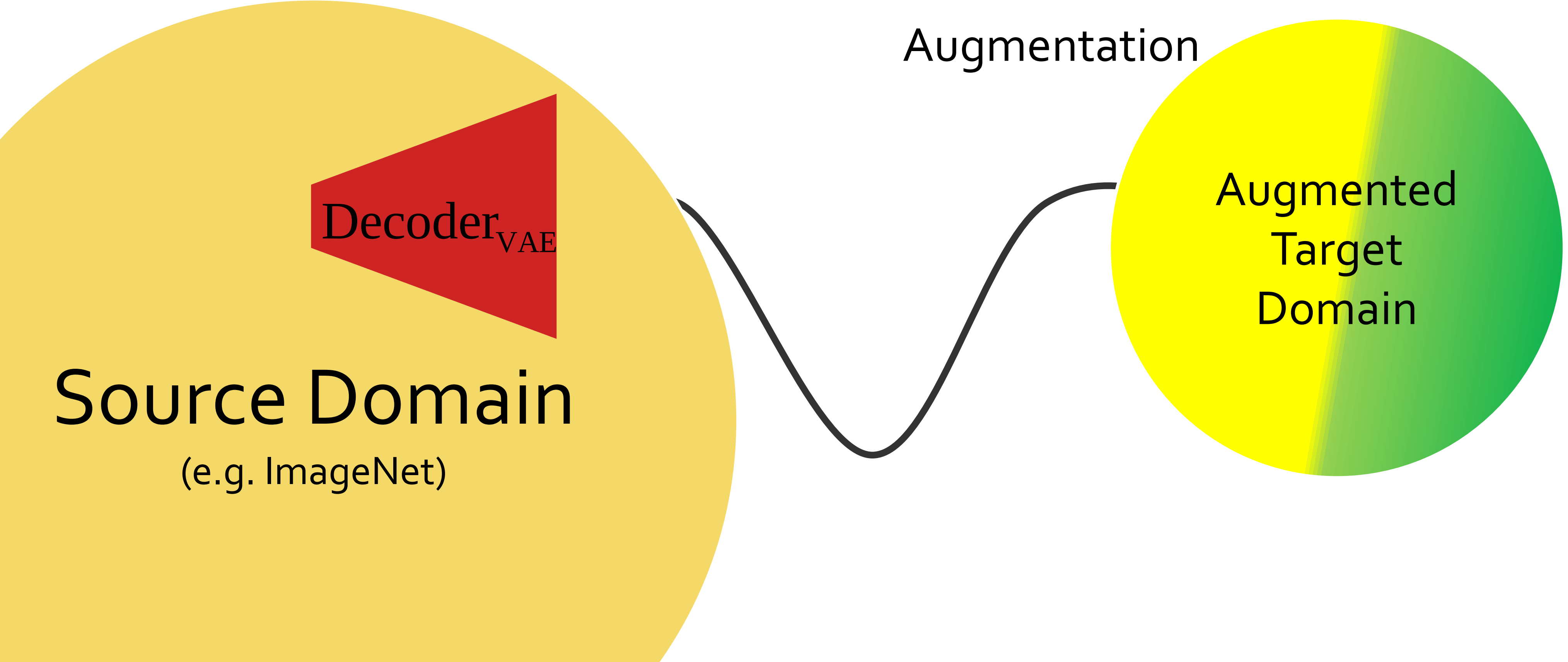
Small challenge: Supervised Generative Models!



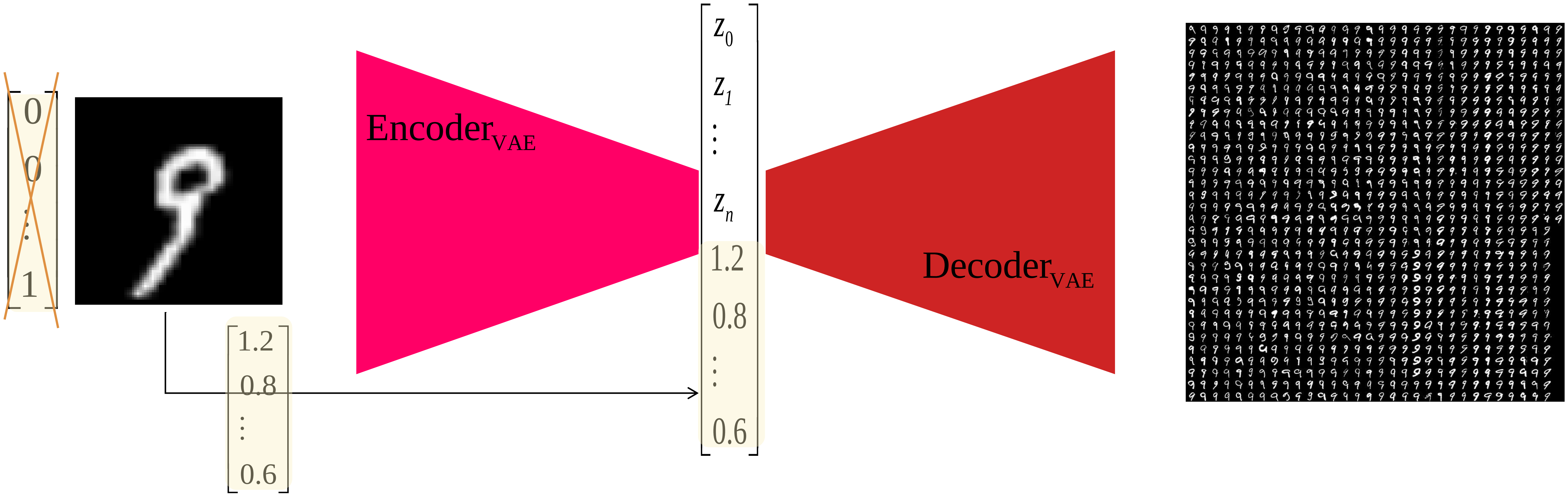
BIG challenge: Meta-Learning



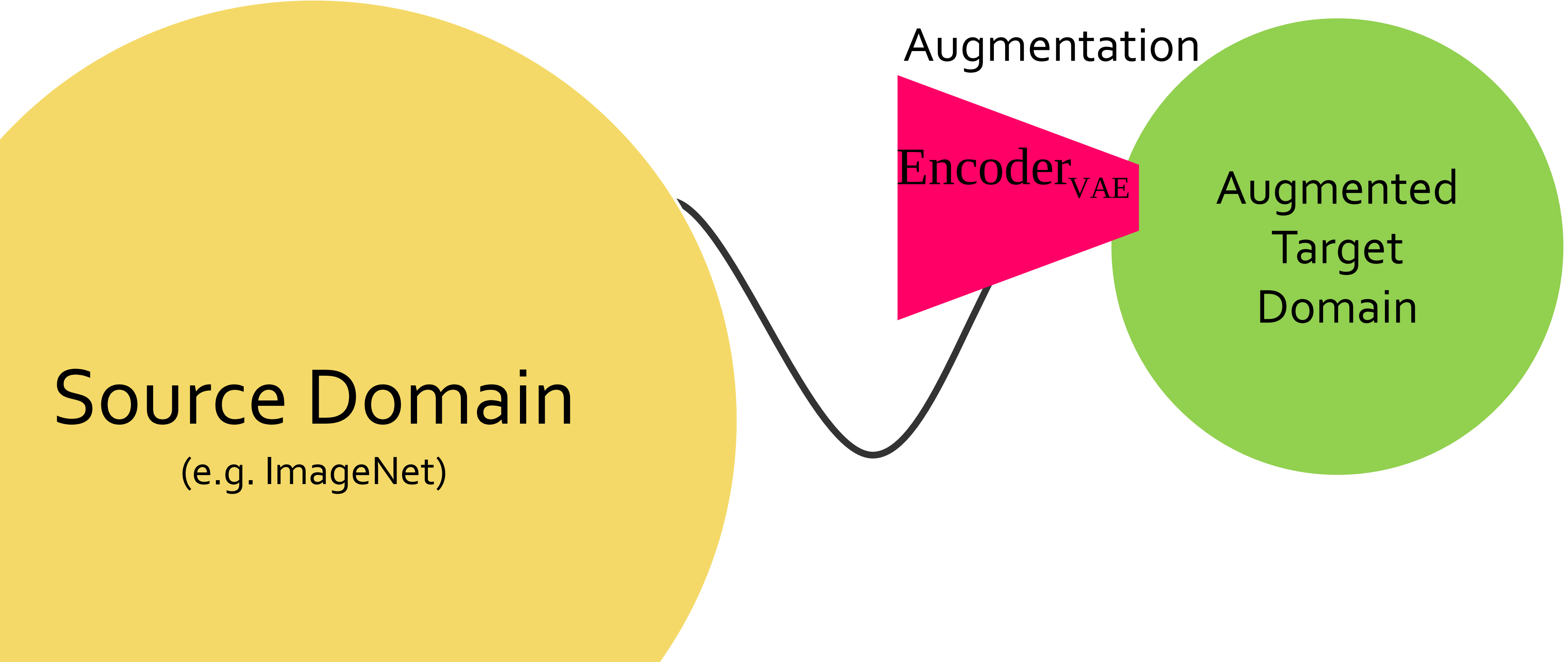
BIG challenge: Meta-Learning



Our Solution:

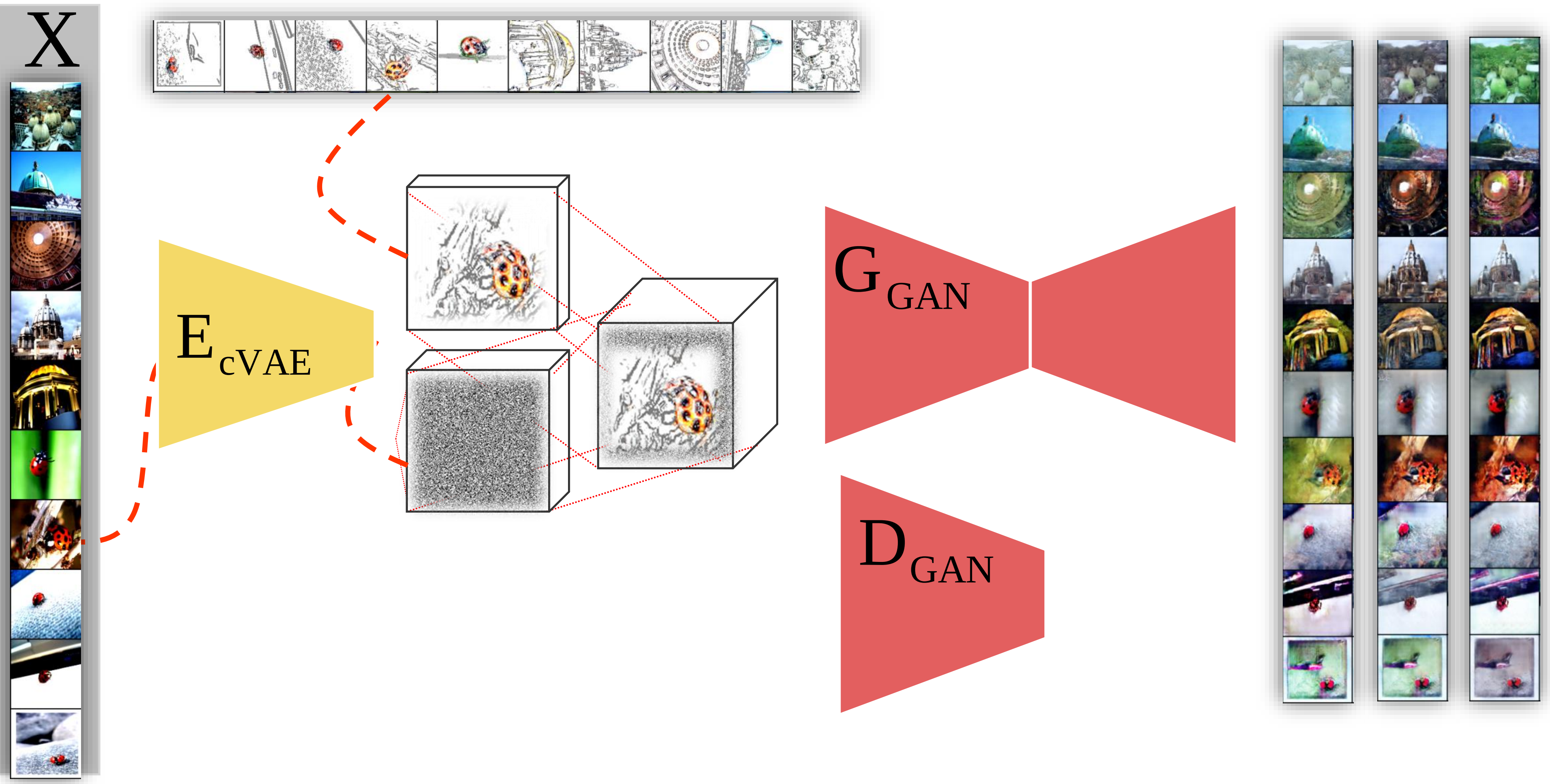


Our Solution:



Extending k-shot to k-pluse-shot Learning

Conditional Variational Autoencoder GAN

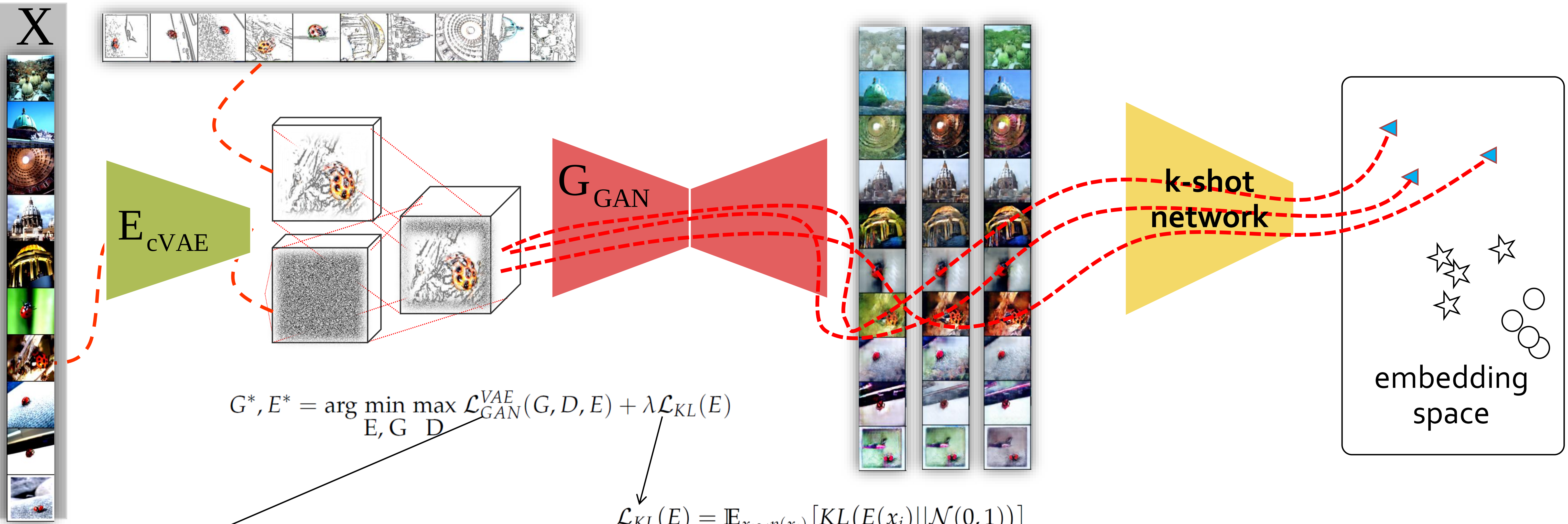


Consecutives Version

Inspired by:
Zhu, Jun-Yan, et al. "Toward multimodal image-to-image translation." *Advances in Neural Information Processing Systems*. 2017.

Extending k-shot to k-plus-shot Learning

Conditional Variational Autoencoder GAN



$$G^*, E^* = \arg\min_{E,G}\max_{D} \mathcal{L}^{VAE}_{GAN}(G, D, E) + \lambda \mathcal{L}_{KL}(E)$$

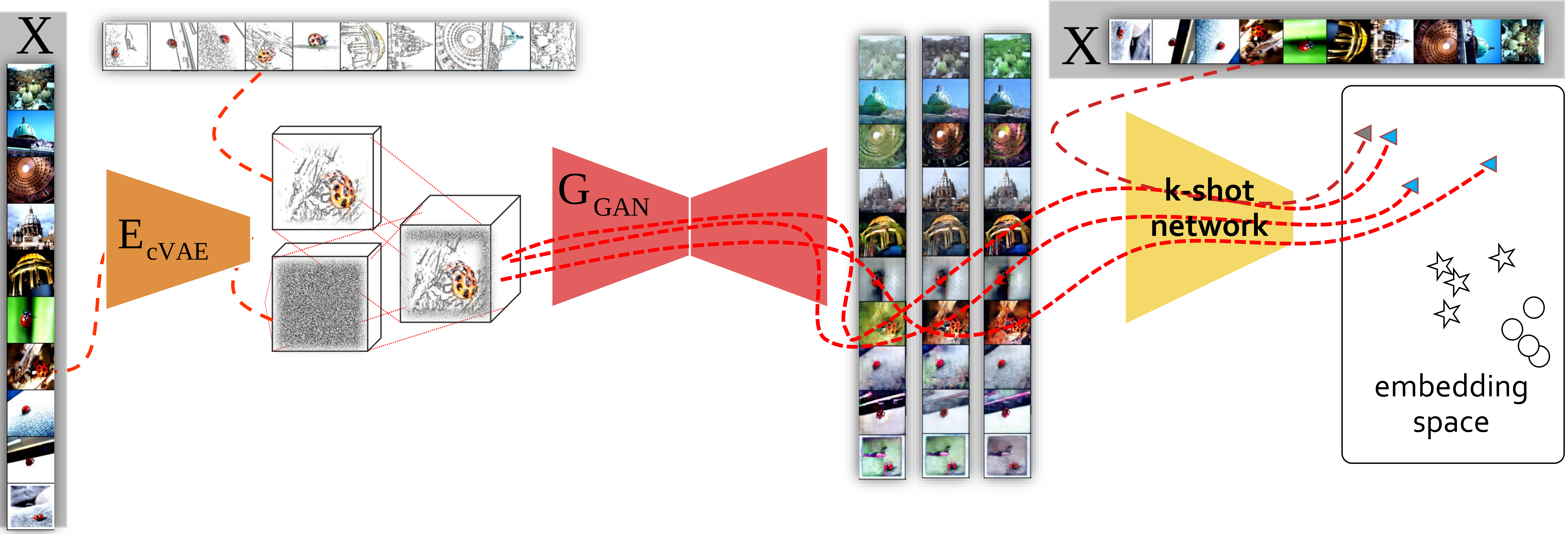
$$\mathcal{L}_{KL}(E) = \mathbb{E}_{x_i \sim p(x_i)}[KL(E(x_i) || \mathcal{N}(0,1))]$$

$$\mathcal{L}^{VAE}_{GAN} = \mathbb{E}_{x_i, x_i^c \sim p(x_i, x_i^c)}[log(D(x_i, x_i^c))] + \mathbb{E}_{x_i, x_i^c \sim p(x_i, x_i^c), z \sim E(x_i)}[log(1 - D(x_i^c, G(x_i^c, z)))]$$

Consecutives Version

Extending k-shot to k-plus-shot Learning

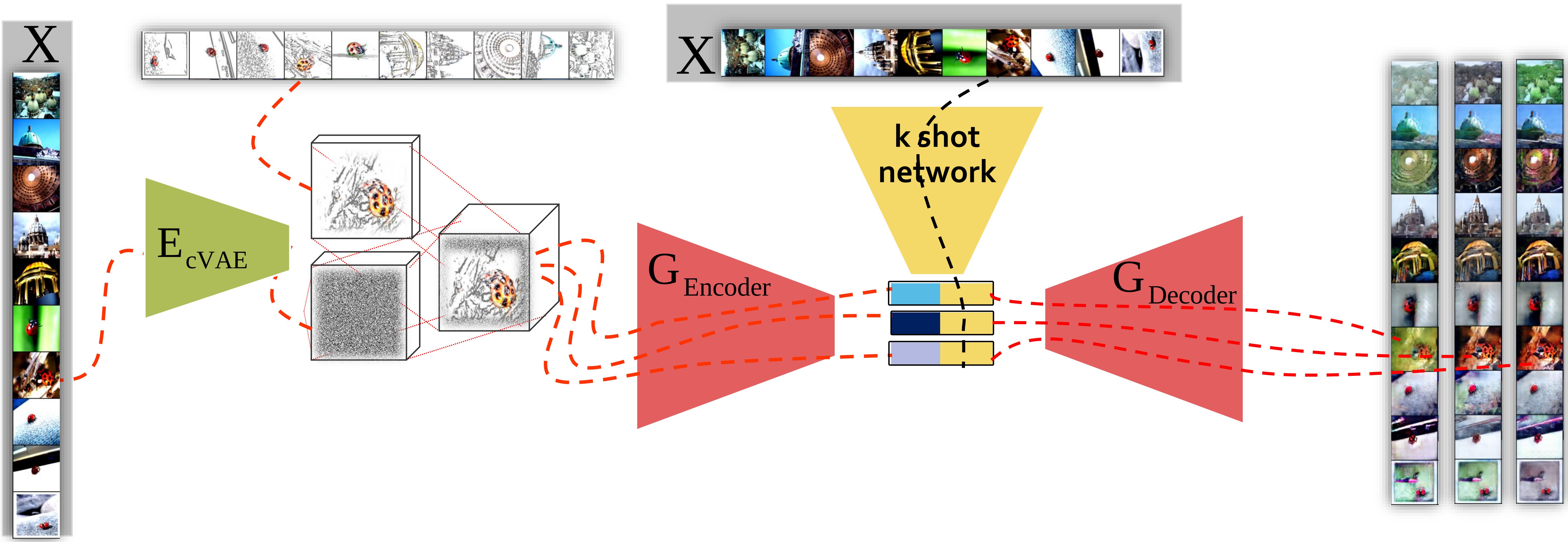
Conditional Variational Autoencoder GAN



Consecutives Version

Extending k-shot to k-plus-shot Learning

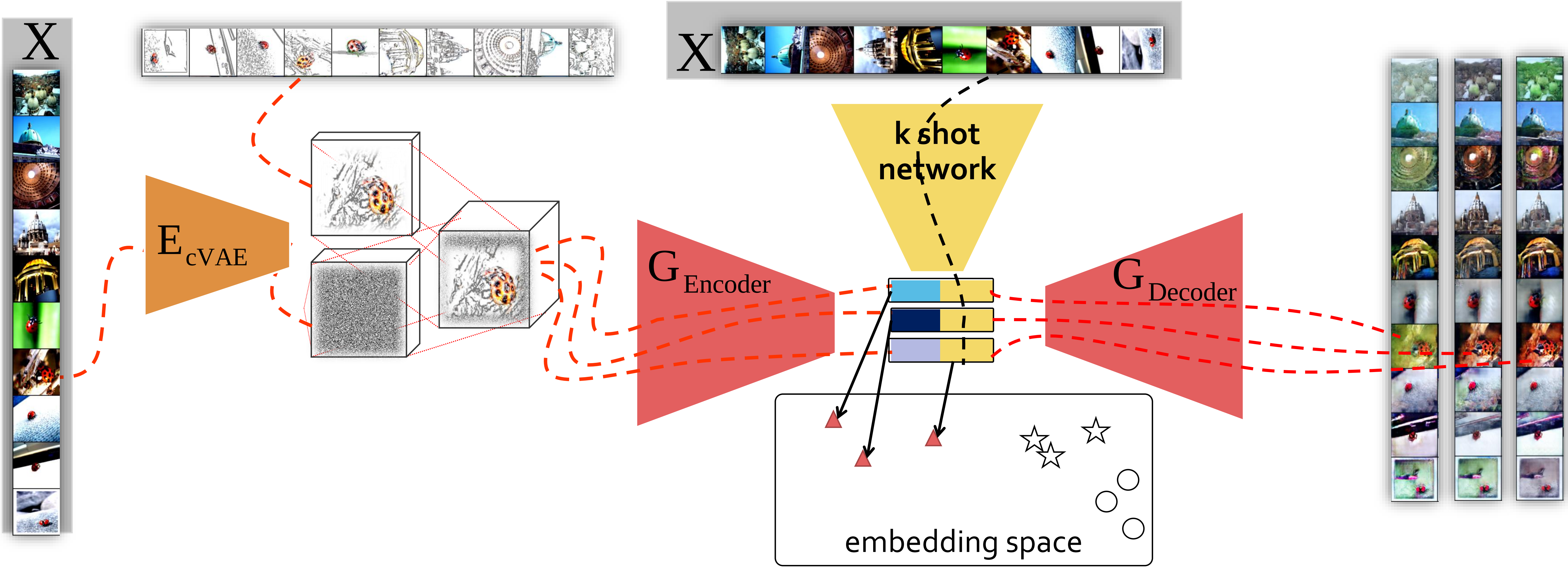
Conditional Variational Autoencoder GAN



Lateral Version

Extending k-shot to k-plus-shot Learning

Conditional Variational Autoencoder GAN



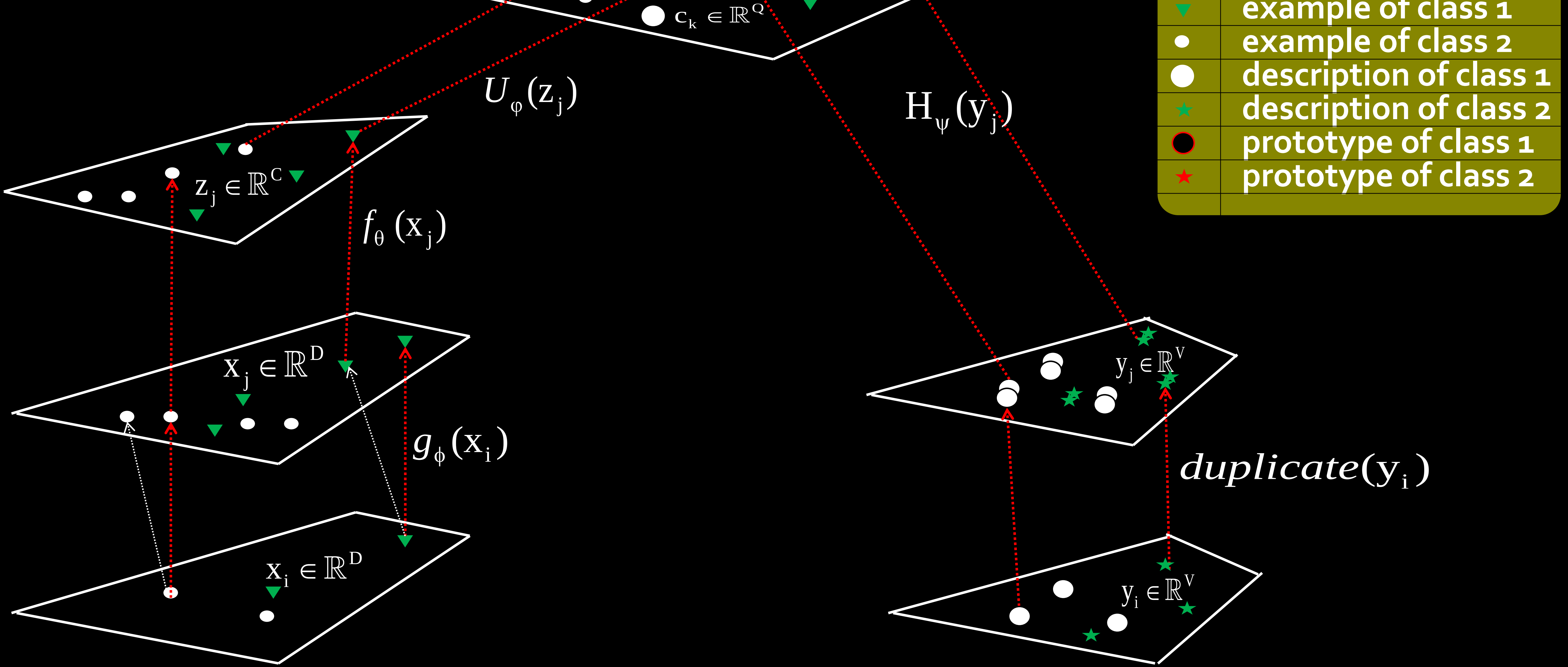
Lateral Version

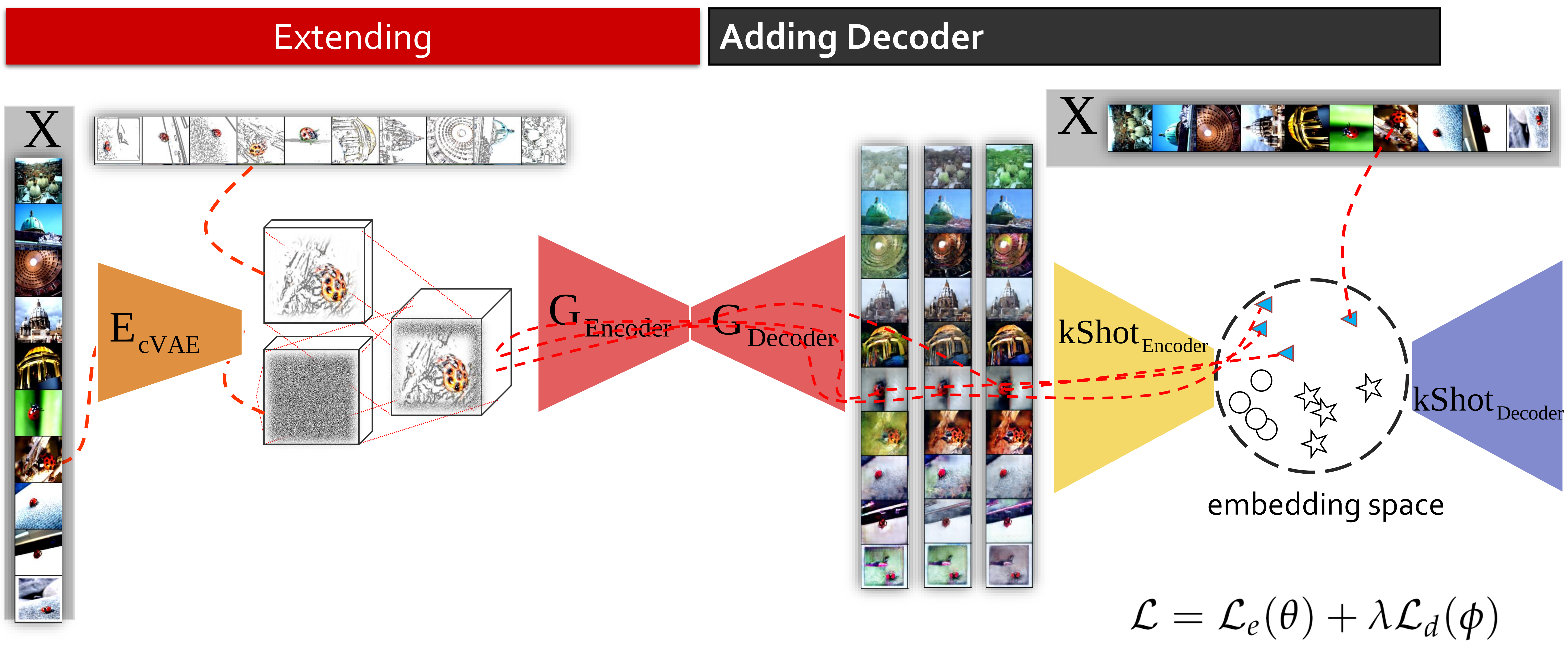
TABLE 2.1: Average classification accuracies on *mini*ImageNet in 5 ways, 5 Shot (with 5 query) setting

Method	Accuracy
MN (Vinyals et al., 2016)	55.3%
MAML (Finn, Abbeel, and Levine, 2017)	63.1%(±0.92%)
PN (Snell, Swersky, and Zemel, 2017)	63.20%(±0.04%)
Ours(Consecutive)	62.4%(±0.03%)
Ours(Lateral)	**65.7** % (±0.04%)

Extending Zero-Shot learning

Ch.II





Extending
Adding Decoder
X
X
E_{cVAE}
$G_{Encoder}$
$G_{Decoder}$
$kShot_{Encoder}$
$kShot_{Decoder}$
embedding space
$\mathcal{L} = \mathcal{L}_e(\theta) + \lambda \mathcal{L}_d(\phi)$

Summary of the contributions

1. Extending cVAE-GAN to meta-learning framework
2. k-plus learning with realistic and variational data augmentation
3. Zero-shot learning (k-plus learning)
4. Adding Decoder for distance based k-shot learning

Content

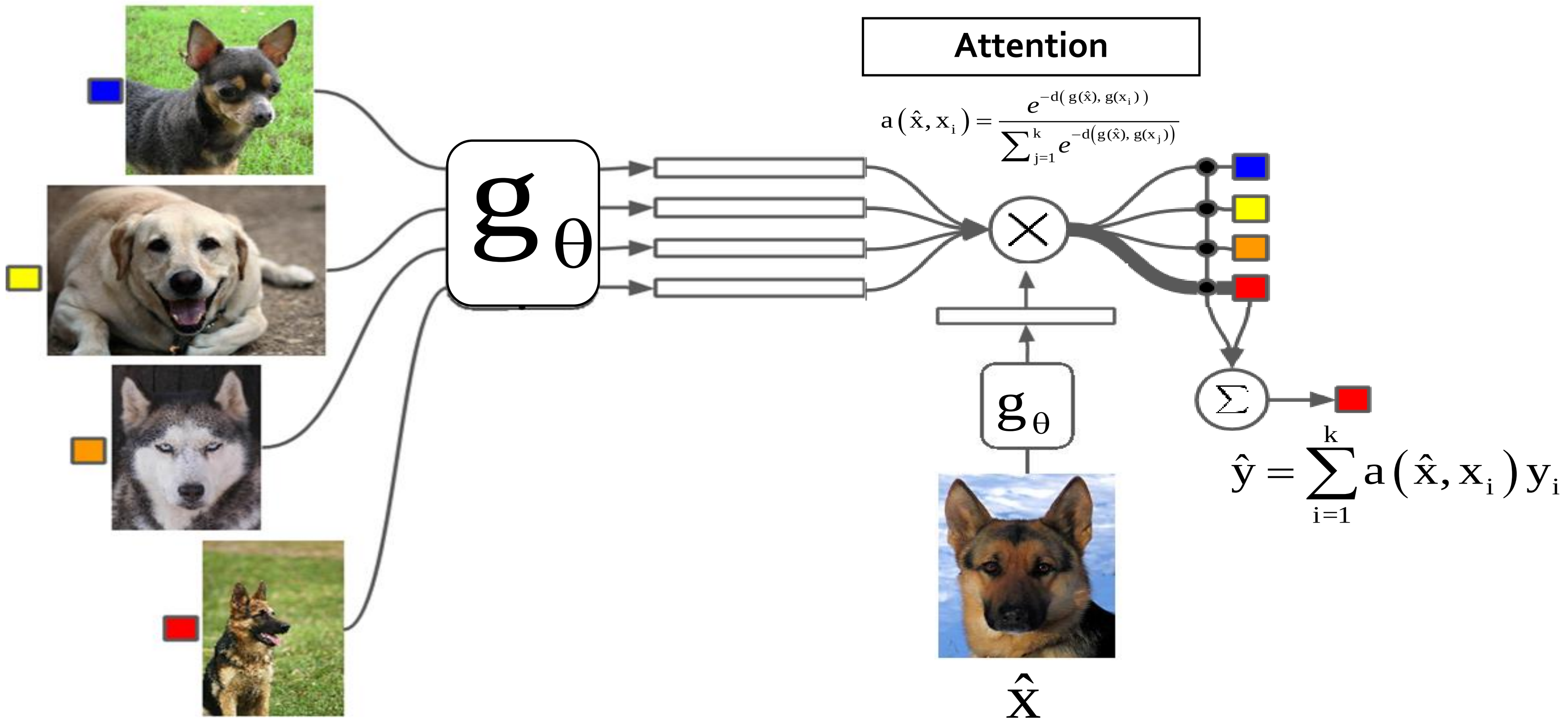
- Chapter I) Introduction
    - K-shot learning
    - Meta learning
    - Prototypical Neural Network
- Chapter II) Extending k-shot to k-plus shot
    - Preliminary work
    - Zero-Shot learning
    - Adding Decoder
- Chapter III) Interactive fast adaptation
    - Motivation
    - Active Meta-Learning
    - Memory-augmented model for quick adaptation
    - Proposal
- Chapter IV) Reducing Annotation Amount in Visual Learning
    - Motivation Active
    - Proposal

LSTM and Matching Network

Meta-learned LSTMs

- could rapidly extend to never-before-seen quadratic functions (Hochreiter, Younger, and Conwell, 2001)
- could successfully find the contextual pattern in k-shot learning (Vinyals et al., 2016), and (Wang et al., 2018)

Vinyals et al., 2016



LSTM and Neural Turing Machine

Meta-learned LSTMs

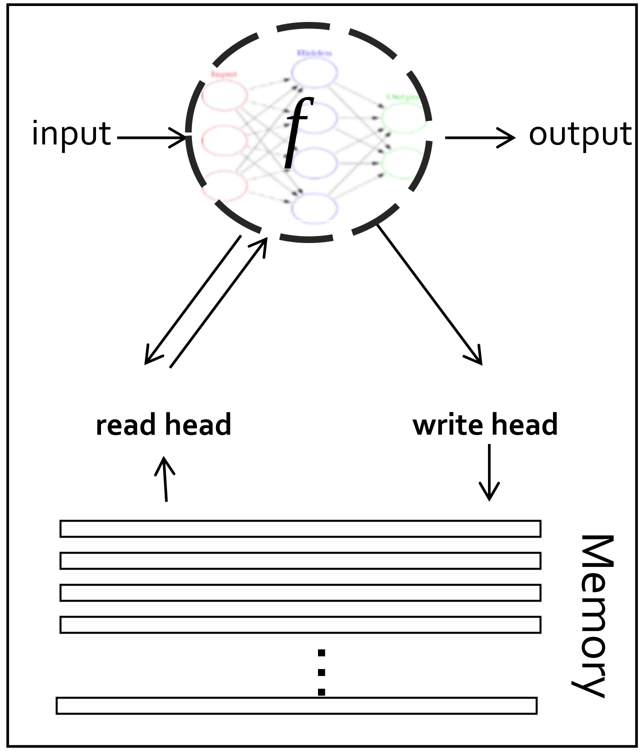
- could rapidly extend to never-before-seen quadratic functions (Hochreiter, Younger, and Conwell, 2001)
- could successfully find the contextual pattern in k-shot learning (Vinyals et al., 2016), and (Wang et al., 2018)

Problems with LSTM:

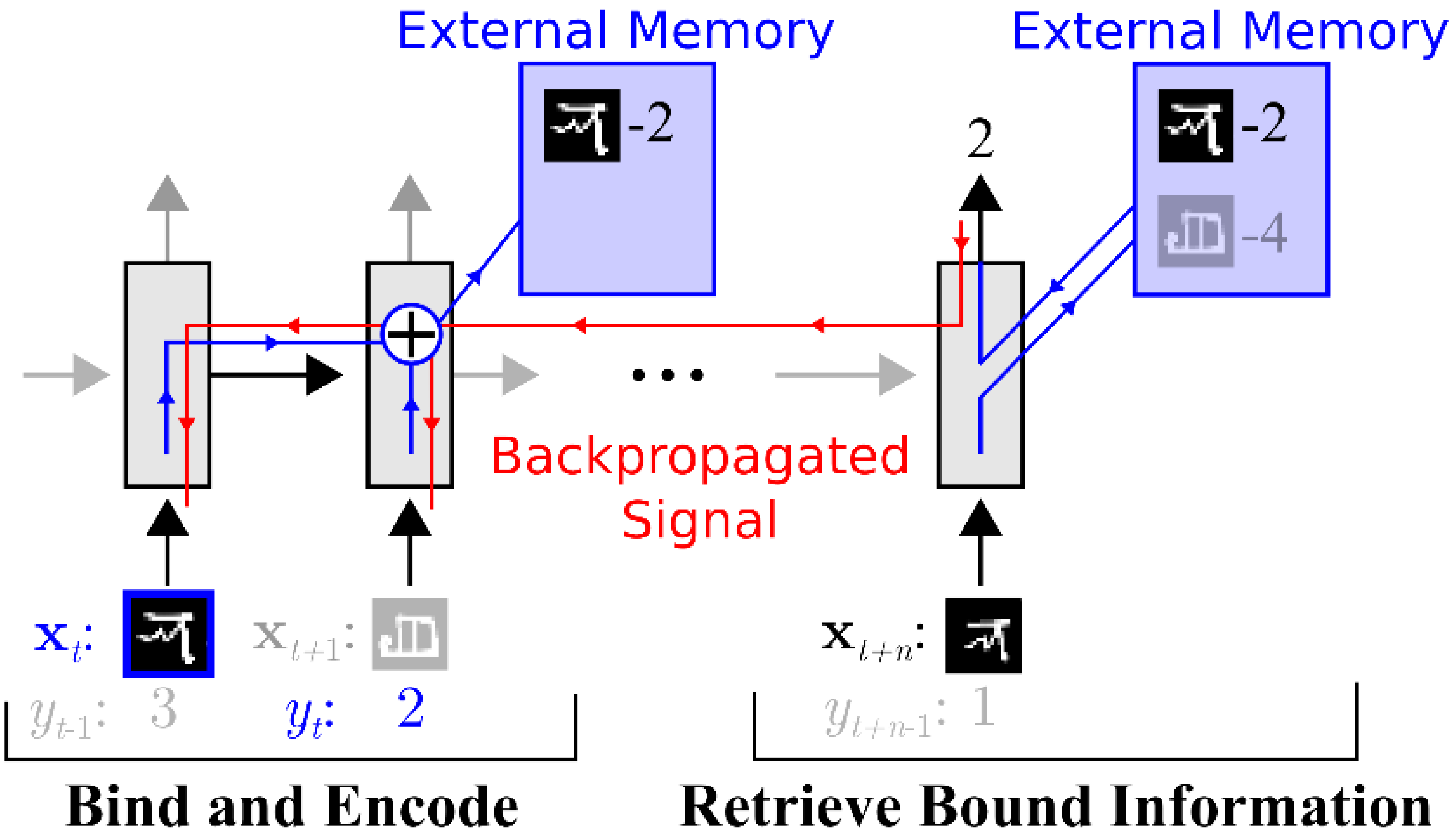
- LSTM's memory cannot extend to the classification of examples in many situations
- LSTM's memory is not addressable and retrievable when we need information.

Solution:

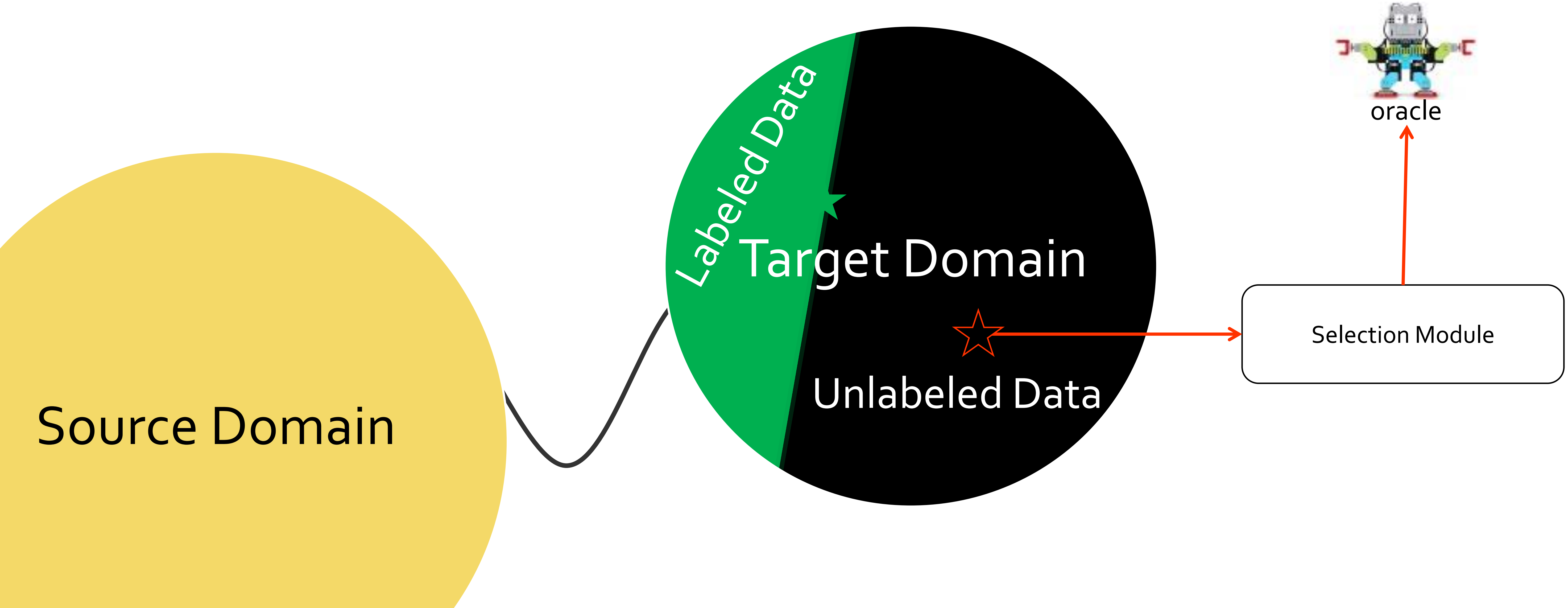
- Neural Turing machines (NTM) (Graves, Wayne, and Danihelka, 2014)



Memory-augmented model for quick adaptation (Santoro et al., 2016)

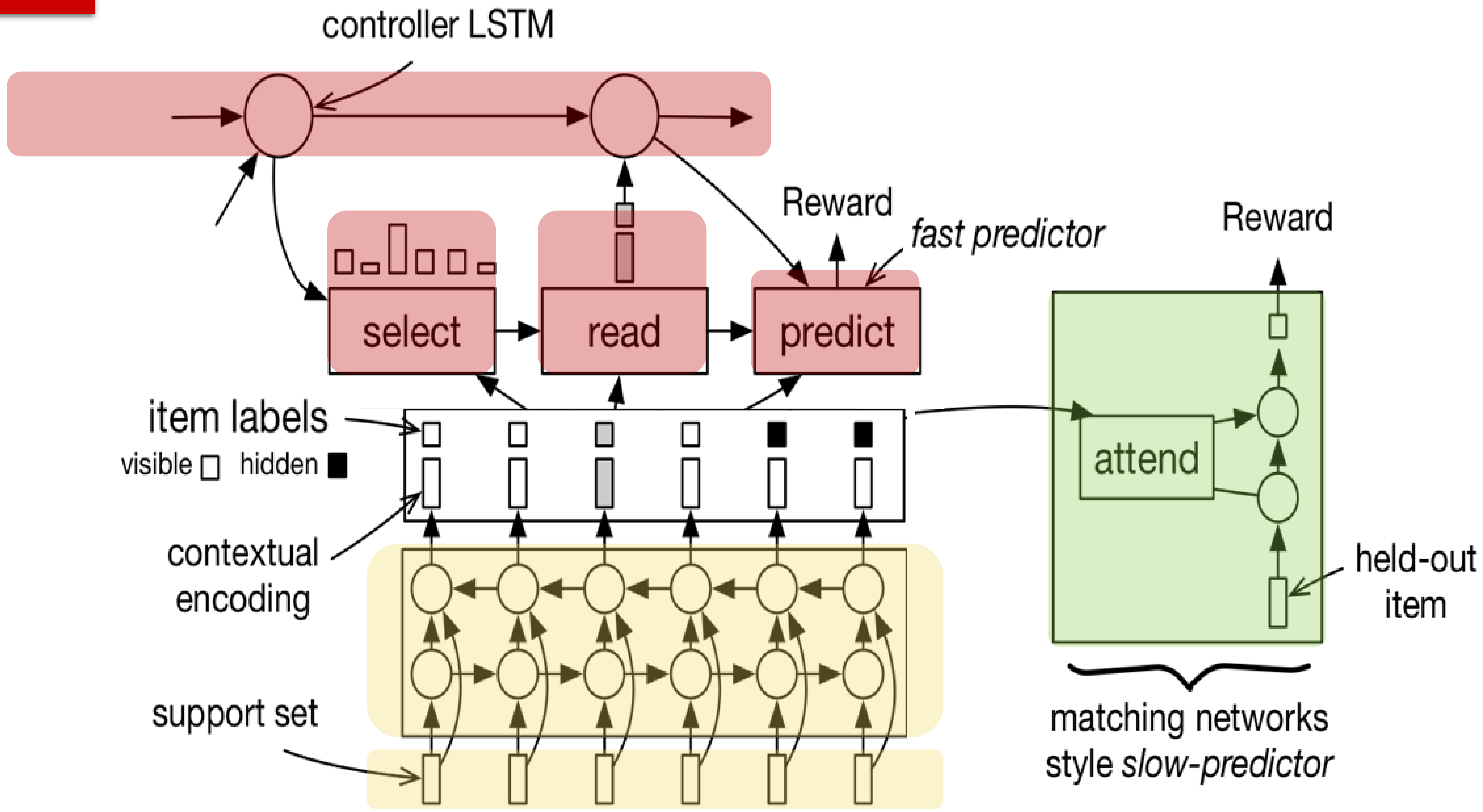


Learning Active Learning (LAL)



Learning active learning (LAL)

Bachman, Sordoni, and Trischler, 2017



$$R(E,S_t,h_t) = \sum_{(\hat{x},\hat{y})\in E} logp(\hat{y}|\hat{x},h_t,S_t)$$

$$\underset{\theta}{maximize} \underset{(S,E)\sim\mathcal{D}}{\mathbb{E}} \left[\underset{\pi(S,E)}{\mathbb{E}} \left[\sum_{t-1}^{T} R(E,S_t,h_t) \right] \right]$$

for t = 1 ... T **do**

$i \leftarrow SELECT(S^u_{t-1}, S^k_{t-1}, h_{t-1})$

$h_t \leftarrow UPDATE(h_{t-1}, x_i, y_i)$

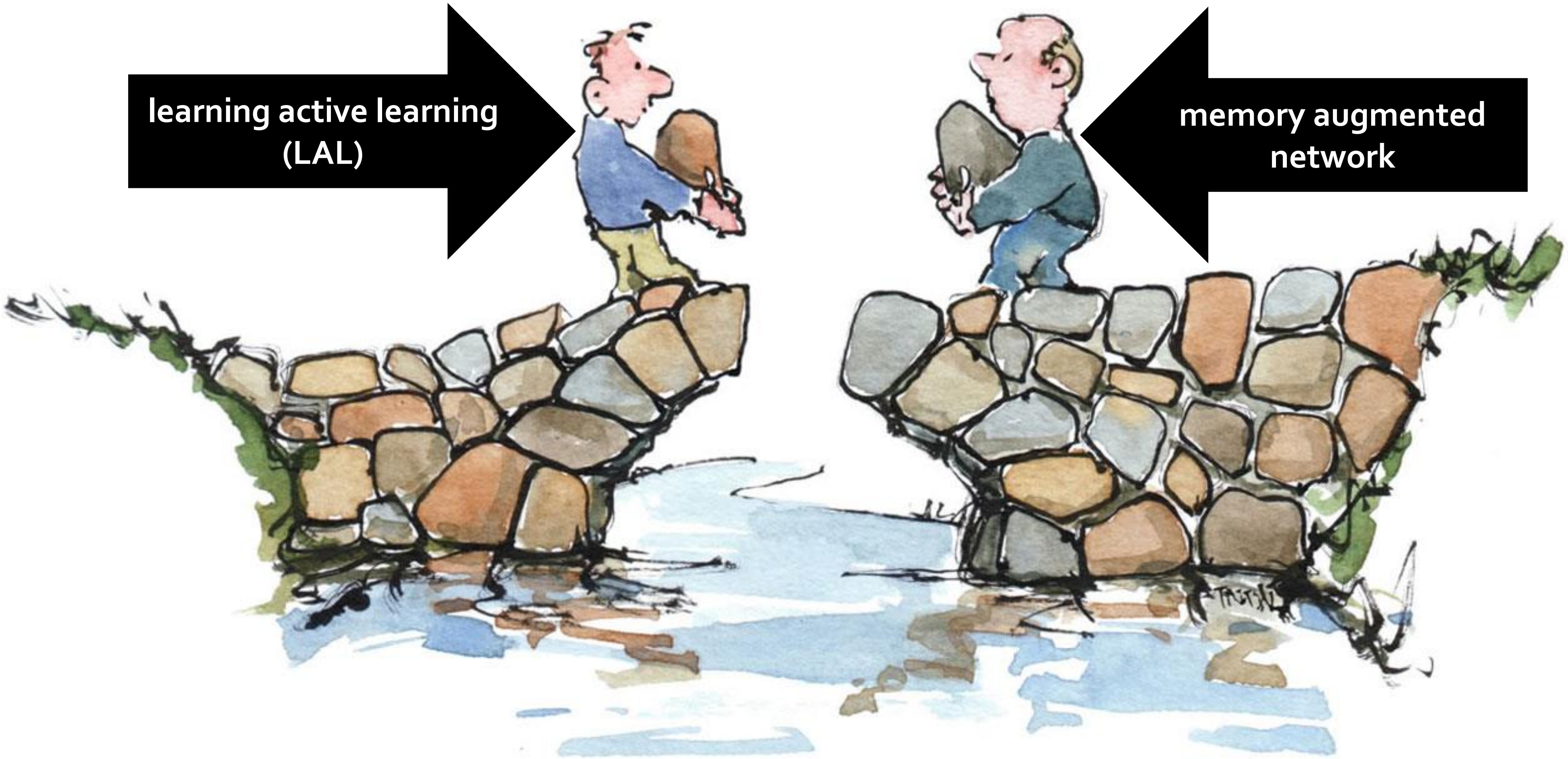
$S^k_t \leftarrow S^{t-1}_k \cup (x_i, y_i), S^k_t \leftarrow S^{t-1}_u / (x_i, .)$

$L^S_t \leftarrow \mathrm{FAST-Pred}(S, S^u_t, S^k_t, h^t)$

$L^E_T \leftarrow \mathrm{SLOW-Pred}(E, S^u_T, S^k_T, h_T)$

Our proposal

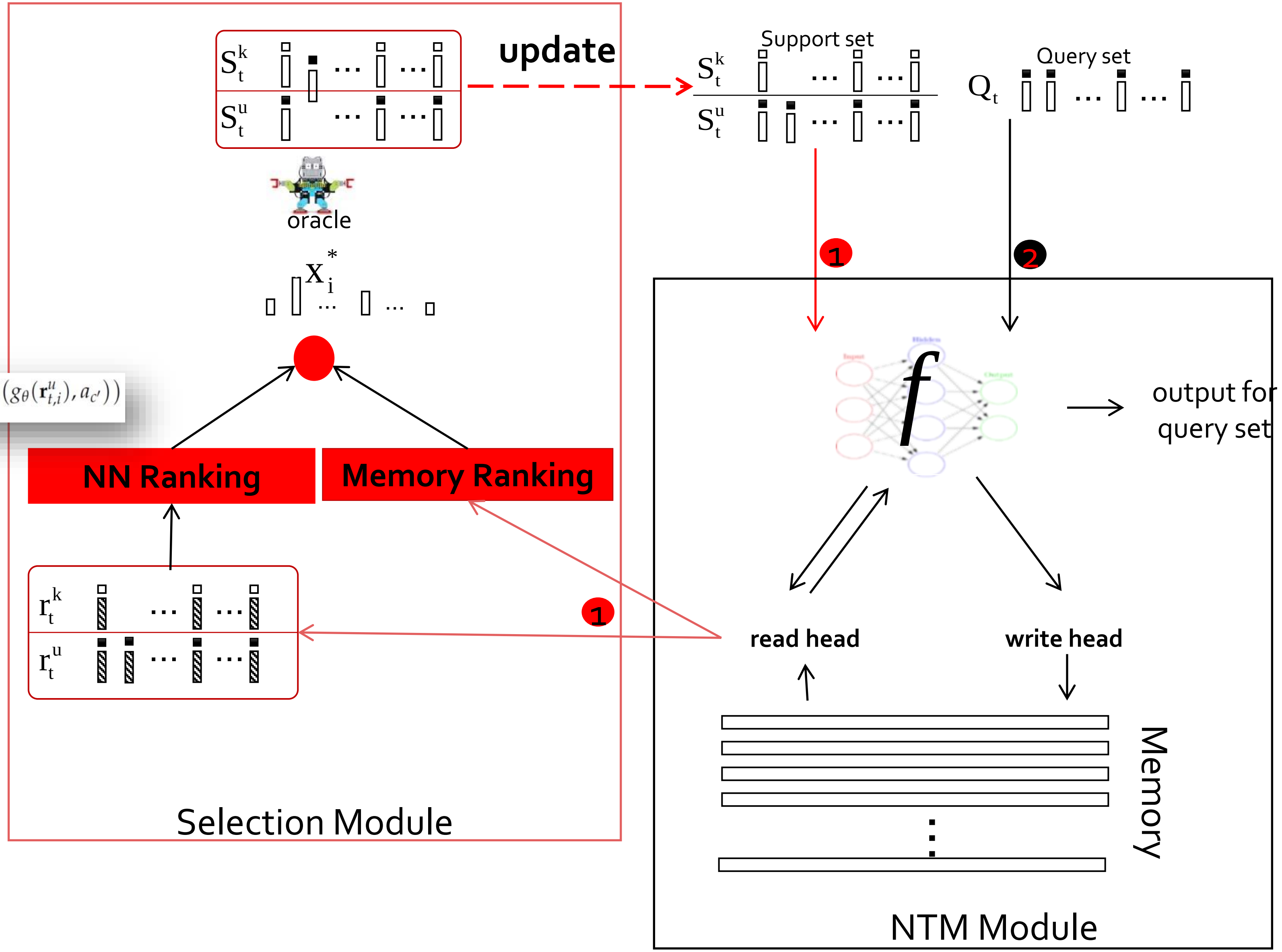
LAL by memory augmented neural networks



By Frits Ahlefeldt

LAL by memory augmented neural networks

$$\mathcal{L}_\theta = \exp(d(g_\theta(\mathbf{r}^u_{t,i}), a_c)) + log\sum_{c'}(exp(-d(g_\theta(\mathbf{r}^u_{t,i}), a_{c'}))$$



read

$$d_{t,i} = d(k_{t,i}, M_t) = \frac{k_{t,i}.M_t}{||k_{t,i}||.||M_t||}$$

$$R^i_t = \frac{exp(k_{t,i})}{\sum_j exp(k_{t,j})}$$

$$\mathbf{r}_{t,i} \leftarrow \sum_i \mathbf{w}^r_t(i)\mathcal{M}_t(i)$$

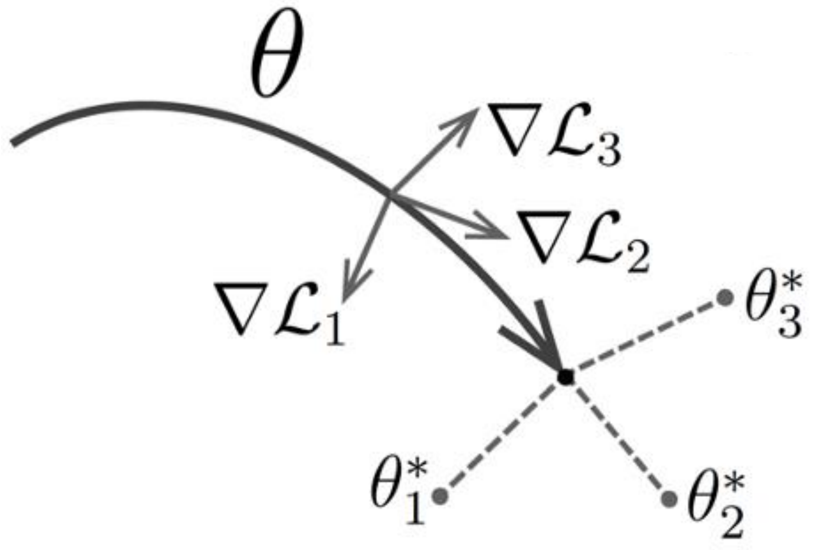
write

$$\mathbf{w}^r_t(i) \leftarrow \frac{exp(d(k_{t,i}, \mathcal{M}_t(i)))}{\sum_j exp(d(k_{t,i}, \mathcal{M}_t(j)))}$$

MAML- Background

Algorithm 2 MAML's algorithm proposed in (Finn, Abbeel, and Levine, 2017)

while not convergence **do** ▷ We have the answer if r is 0
  Randomly select N number of tasks $\mathcal{T}_N$ from all tasks
  for $\mathcal{T}_i \in \mathcal{T}_s$ **do**
    $D_{base} \leftarrow$ {S, E} sample k labeled and n unlabeled examples from N classes
    $G_{base} \leftarrow \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f_\theta)$ ▷ Compute gradient of $\mathcal{L}$ at base level using D_{base}
    $\theta'_i \leftarrow \theta - \alpha G_{base}$ ▷ Update Parameters
    $D^t_{meta} \leftarrow$ {S, E} sample examples for meta training
  endfor
  $G_{meta} \leftarrow \nabla_\theta \sum_{\mathcal{T}_i \sim \mathcal{T}_S} \mathcal{L}_{\mathcal{T}_i}(f_{\theta'_i})$ ▷ Compute meta-gradient using D_{meta}
  $\theta \leftarrow \theta - \beta G_{meta}$ ▷ Update meta-parameters
endwhile



MAML for our proposed model

Algorithm 3 Adapted MAML's algorithm for memory augmented learning active *learning* (changes are in red color line)

while not convergence **do**
    Randomly select N number of tasks $\mathcal{T}_N$ from all tasks
    for $\mathcal{T}_t \in \mathcal{T}_N$ **do**
        $D_{base} \leftarrow$ {S, E} sample k labeled and n unlabeled examples from N classes
        $\theta'_t \leftarrow \theta - \alpha G_{base}$ ▷ Update Parameters
        $D^t_{meta} \leftarrow$ {S, E} sample examples for meta training
        endfor
    endfor
    $G_{meta} \leftarrow \nabla_\theta \sum_{\mathcal{T}_t \sim \mathcal{T}_S} \mathcal{L}_{\mathcal{T}_t}(f_{\theta'_t})$ ▷ Compute meta-gradient using D^t_{meta}
    $\theta \leftarrow \theta - \beta G_{meta}$ ▷ Update meta-parameters
endwhile

Summary of the contributions

1. Memory augmented learning active learning (LAL) for fast adaptation

2. Two levels of gradient for memory augmented neural network

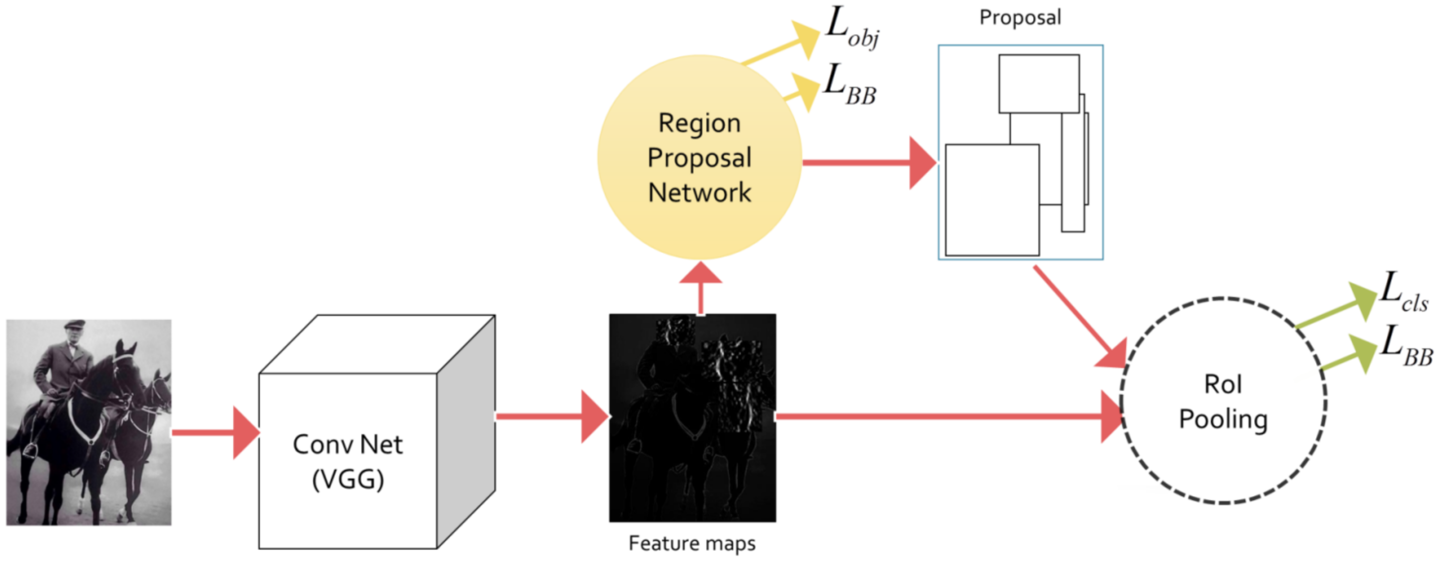
Content

- Chapter I) Introduction
    - K-shot learning
    - Meta learning
    - Prototypical Neural Network
- Chapter II) Extending k-shot to k-plus shot
    - Preliminary work
    - Zero-Shot learning
    - Adding Decoder
- Chapter III) Interactive fast adaptation
    - Motivation
    - Active Meta-Learning
    - Memory-augmented model for quick adaptation
    - Proposal
- Chapter IV) Reducing Annotation Amount in Visual Learning
    - Motivation
    - Proposal

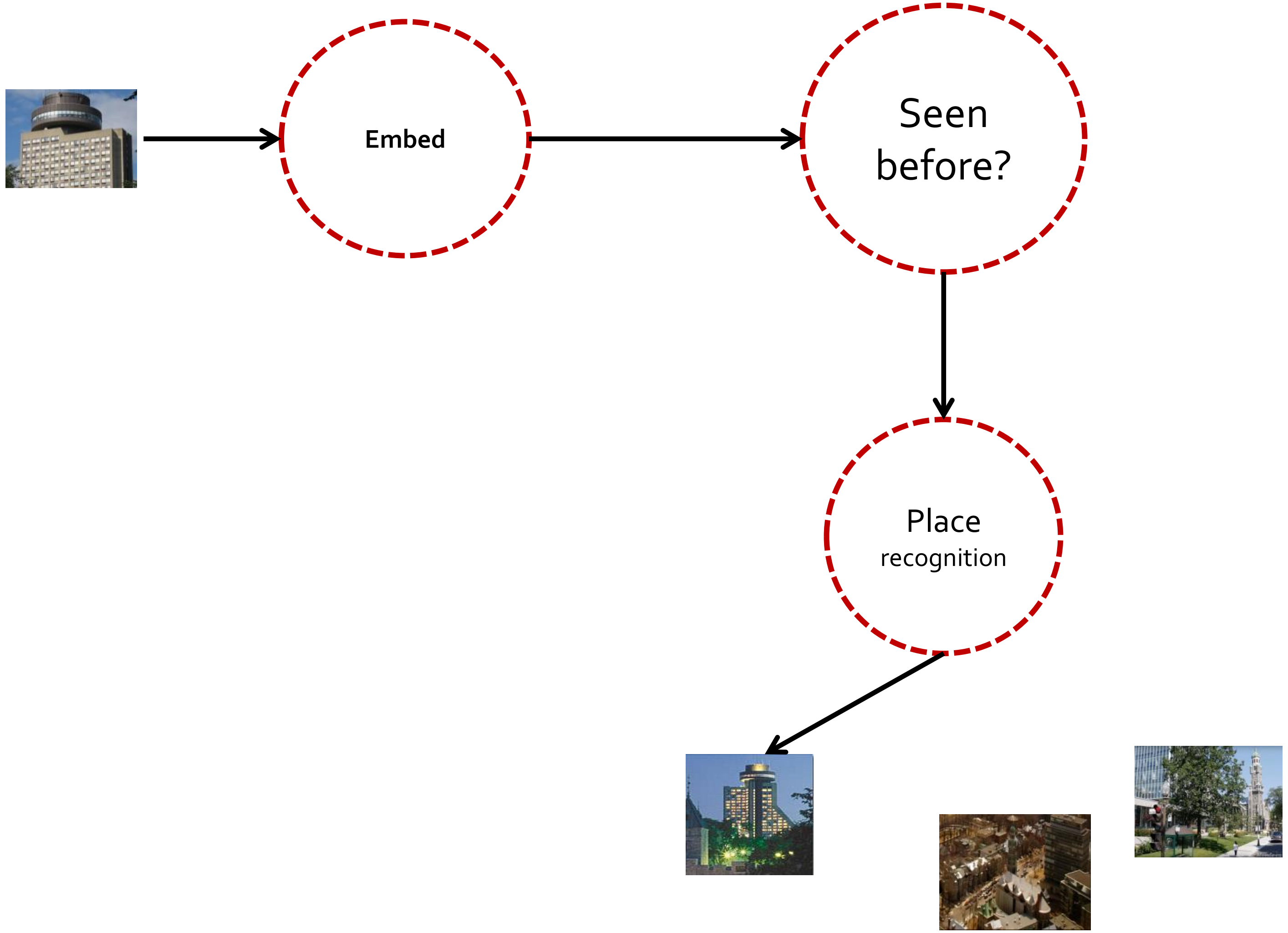
Motivation

Extend interactive fast adaptation

object detection



visual place recognition



Direction I)

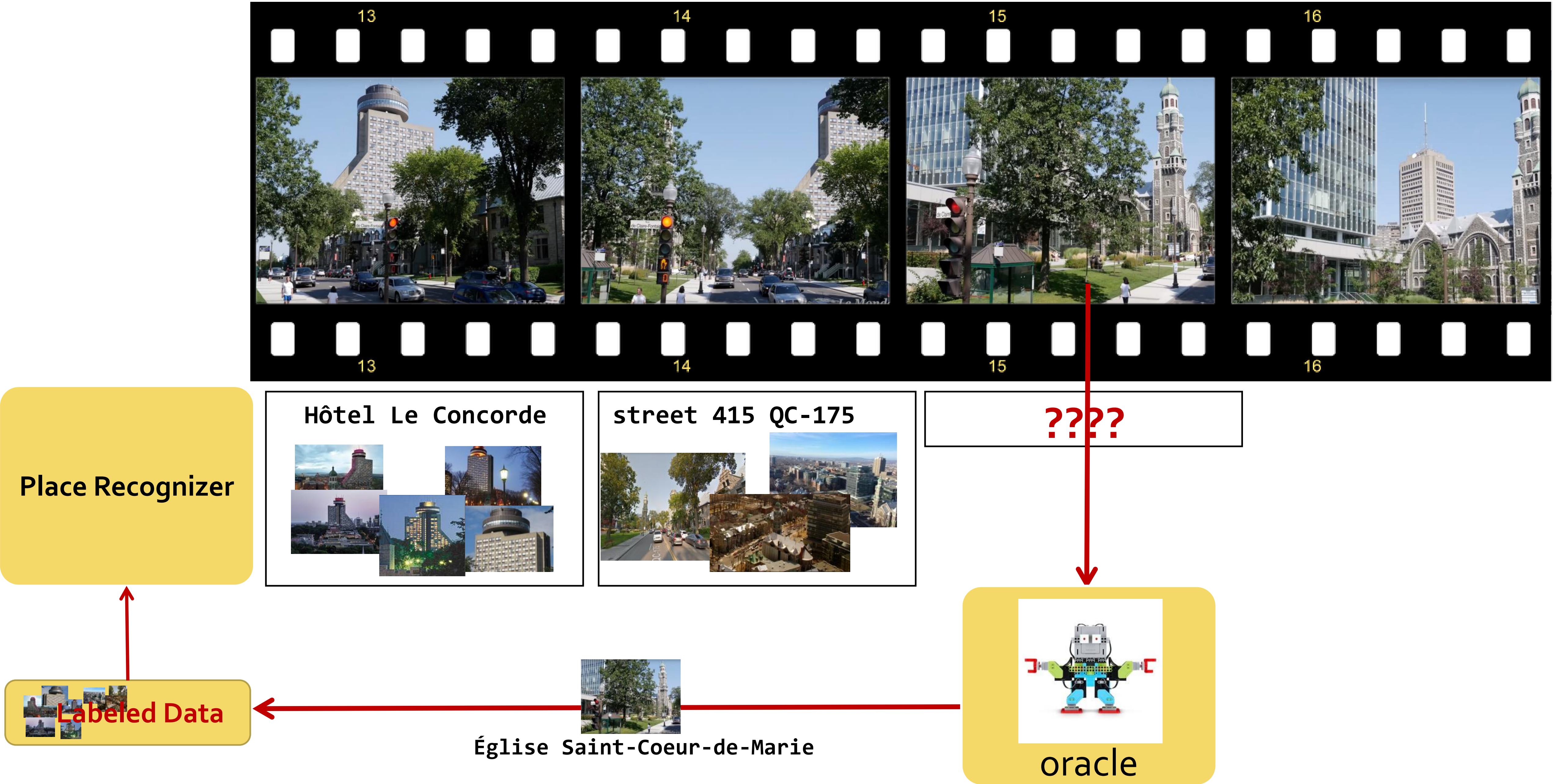
k-shot object recognition

Direction II)

k-shot based active learning for visual place recognition

Visual Place Recognition

Learning visual learning



Conclusion

- cVAE-GAN is extendable to meta-learning framework
- k-plus learning with realistic data augmentation has advantages
- memory augmented networks for learning active learning (LAL) with two level of gradient
- quick adaptation for real world problem is necessary

Thanks

- Christian Gagné
- Ihsen Hedhli
- Marc-André Gardner
- Azadeh Sadat Mozafari
- Alexandre Hains
- Hugo Siqueira Gomes

- Gabriel Leclerc
- Ba Diep Nguyen
- Changjian Shui
- Mahdieh Abbasi
- Louis-Émile Robitaille
- Sophie Baillargeon
- Sébastien De Blois
- Jonathan Marek

Thank You!

Question!