Attentive Weights Generation for Few Shot Learning via Information Maximization

Yiluan Guo, Ngai-Man Cheung etc
CITE:8
Singapore University of Technology and Design
CVPR 2020

2021.05.12.(수)
임진혁

Index

1. Introduction

2. Related

3. Proposed Approach

4. Experiments

임진혁<eeplearning@units.ac.kr>

Introduction: Why I choose this paper?

- **Task specific weigh에 대해서 “attention”이 반영된 논문을 찾다가**

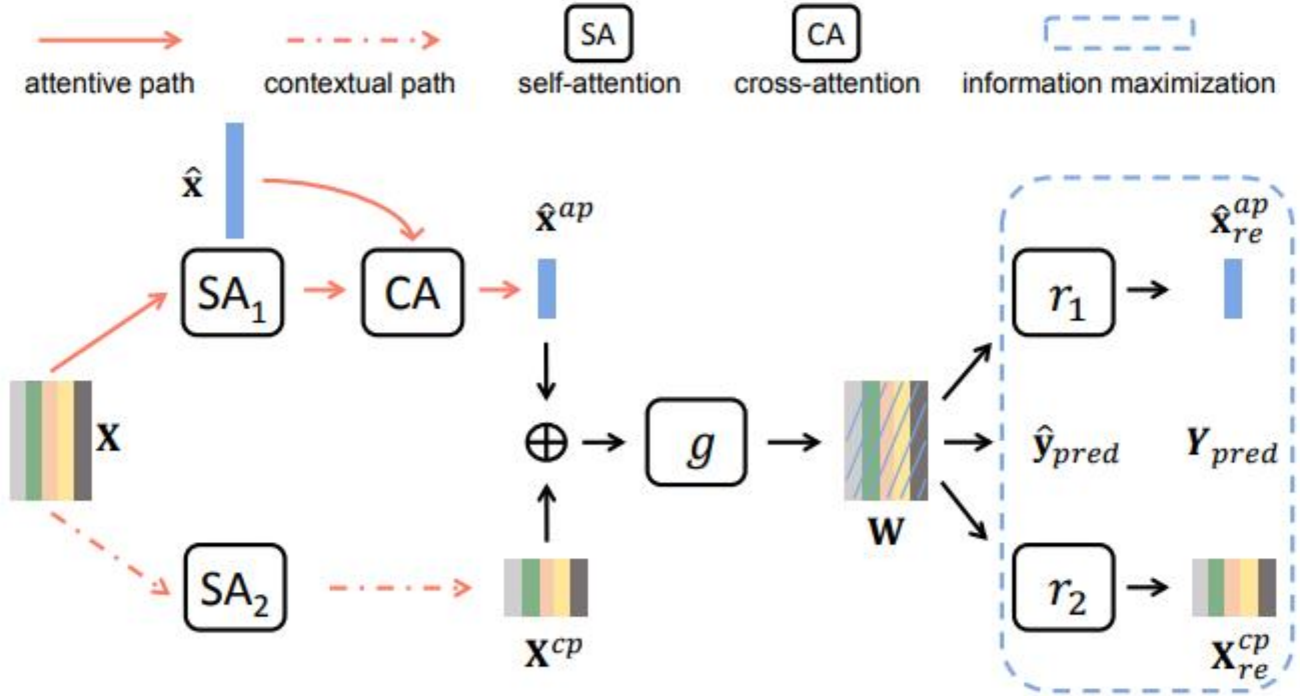
- **실험결과에서 attention의 효과를 정량적 제시**

임진혁<eeplearning@units.ac.kr>

Abstract

"Task Adaptive weights generating for Few shot image classification"

Attentive Weights Generation for Few Shot Learning via Information Maximization (AWGIM).

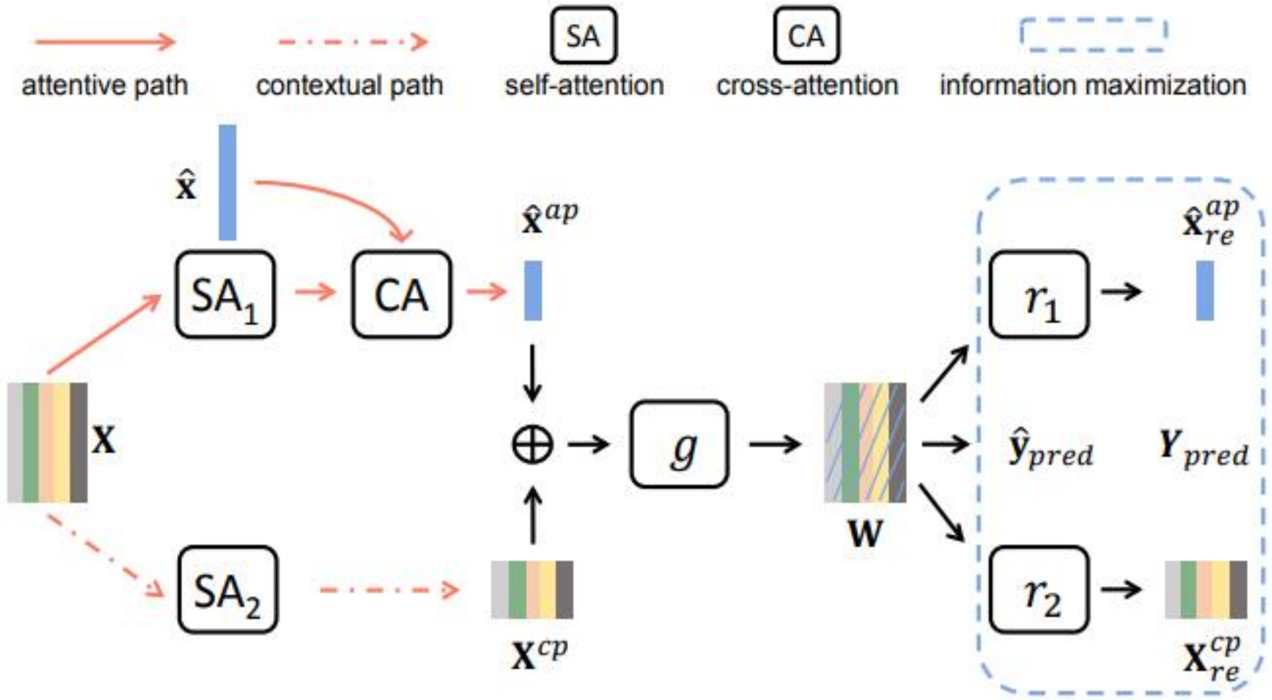


(i) Self-attention and cross attention paths to encode the context of the task and individual queries.

(ii) Mutual information maximization between generated weights and data within the task.

Introduction: Contribution

Attentive Weights Generation for Few Shot Learning via Information Maximization (AWGIM).



(i) Attention mechanism is applied to capture the context information of task & query.
- self attention:
  richer information about task
- cross attention:
  generated weights are adaptive to different task & different query

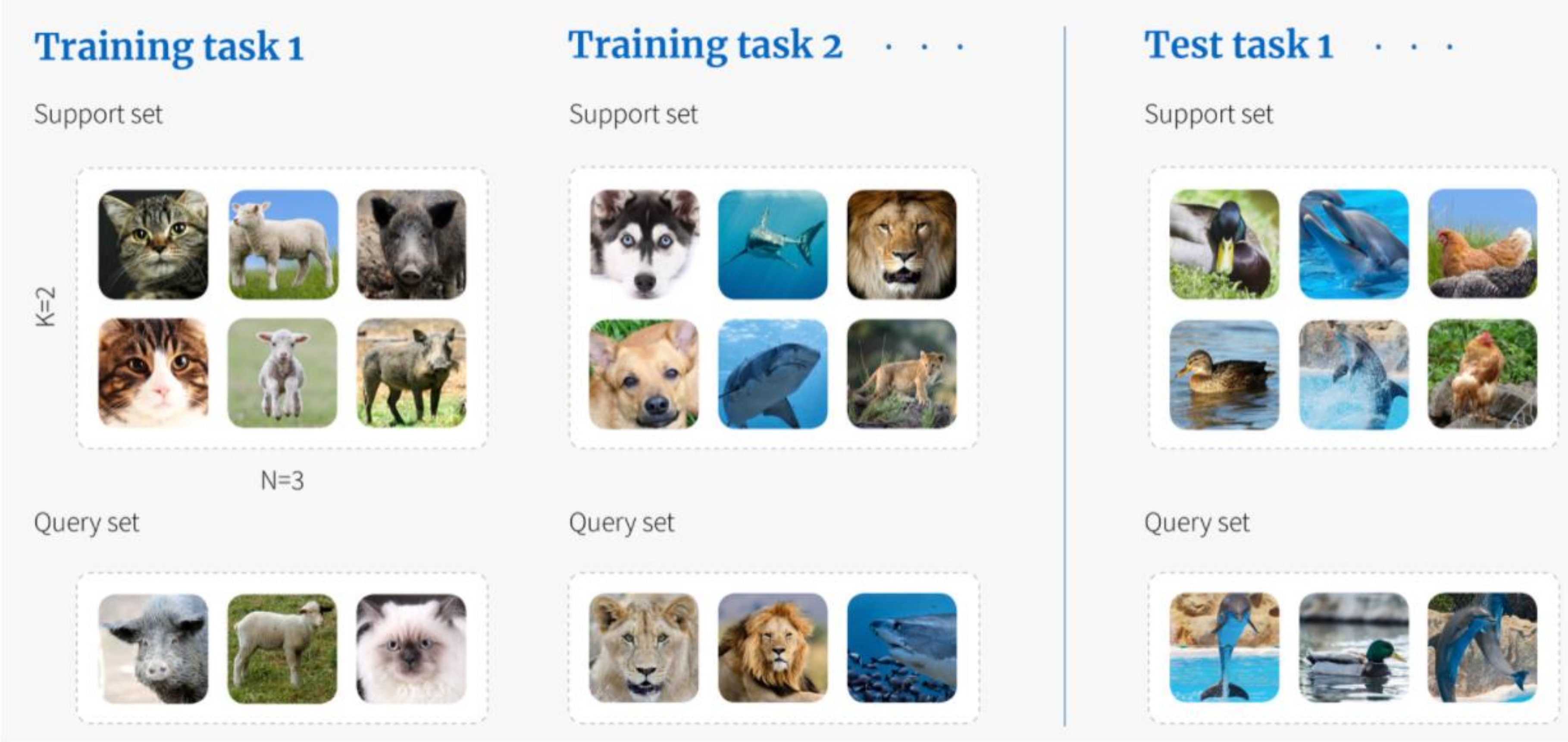
(ii) With Mutual information maximization , generated weights can adapt to diverse query samples.

⇒ Solve weights generation problem

- Fixed weights for different query samples
- Query specific information lost during Generating

2 Related: Few shot problems

Few Shot Learning은 딥러닝이 사람처럼 very few samples로도 학습할 수 있게 하는 연구분야



Deep learning shows high performance while learning a very large amount of labeled data. (limitation)

In contrast, humans can quickly recognize the class with a small amount of data.

임진혁<eeplearning@unist.ac.kr>

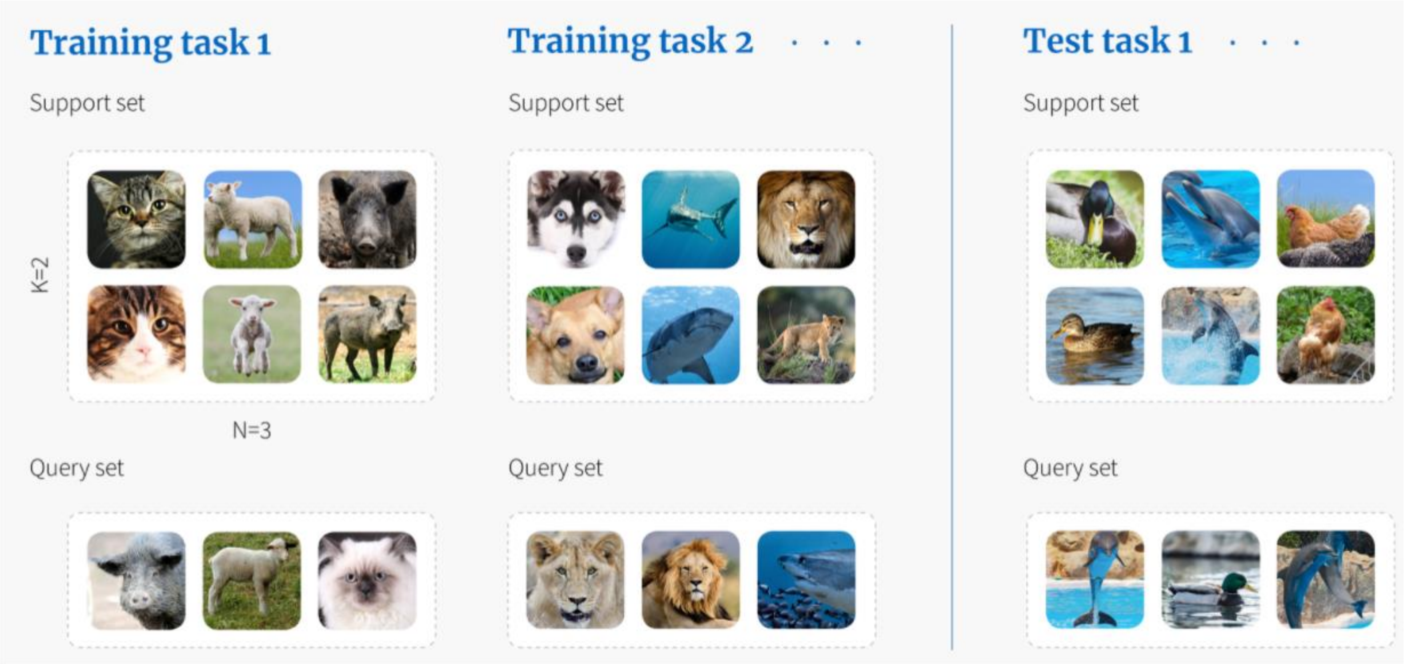
2 Related: Few shot problems

$$\mathcal{T} = (\mathcal{S}, \mathcal{Q})$$

$$\mathcal{S} = \left\{ \left(\mathbf{x}_{c_n}^{k}, \mathbf{y}_{c_n}^{k} \right) \mid k = 1, \ldots, K; n = 1, \ldots, N \right\}$$

$$\mathcal{Q} = \left\{ \left(\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_{|\mathcal{Q}|} \right) \right\}$$

"Meta Model" 's Performance: evaluated on $\mathcal{Q}\left(\hat{\mathbf{x}}, \hat{\mathbf{y}}\right)$ provided $labeled\ \mathcal{S}\left(\mathbf{x}_{c_n}, \mathbf{y}_{c_n}\right)$



N Way K Shot problem:

Task: (N class, K samples)로 이루어진
Support Data로(label 0) 학습하고
Query Data로 테스트한다.

Training task(meta-train):
Query's loss로 업데이트

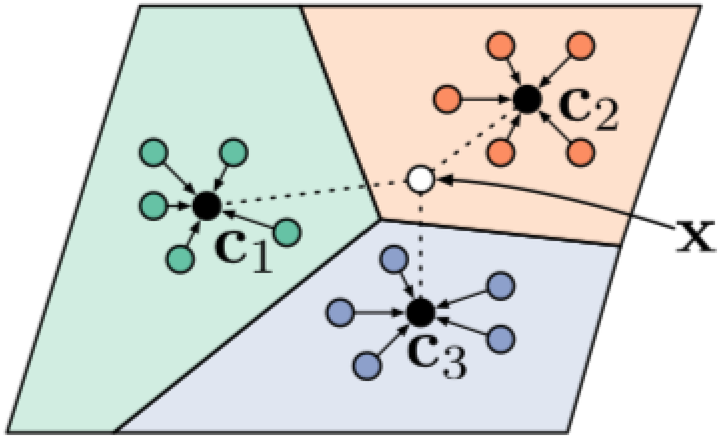
Test task(meta-test):
meta model's performance

2 Related: meta learning

There are many meta learning methods

Metric-based approach

- Support data <-> Query data 거리 유사도를 사용
- Query data를 가장 가까운 거리(유사도)의 support data class로 분류



(a) Few-shot

Optimization Based approach

- parameter optimization problem
- Optimization methods for few samples

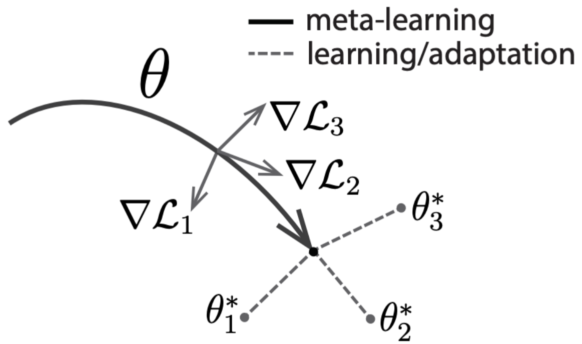


Figure 1. Diagram of our model-agnostic meta-learning algorithm (MAML), which optimizes for a representation θ that can quickly adapt to new tasks.

*Prototypical networks for few-shot learning (2017' NIPS)
*Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

2 Related: Latent Embedding Optimization (Weight Generalization)

Weight Generation Methods:

LEO(“META-LEARNING WITH LATENT EMBEDDING OPTIMIZATION”, Rao etc. ,ICLR 2019)

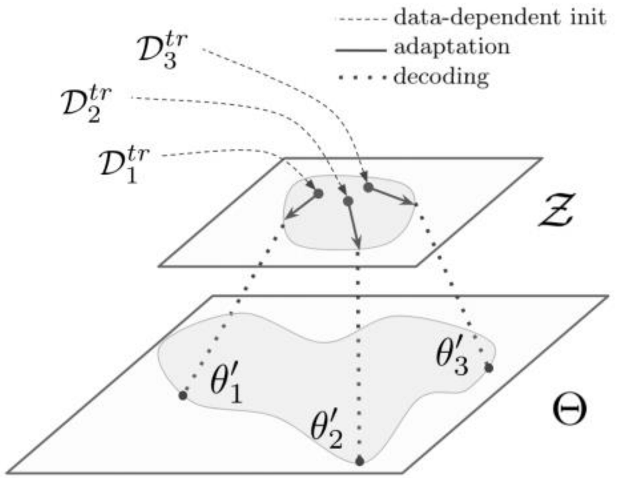


Figure 1: High-level intuition for LEO. While MAML operates directly in a high dimensional parameter space Θ, LEO performs meta-learning within a low-dimensional latent space $\mathcal{Z}$, from which the parameters are generated.

1.learning a data-dependent latent generative representation of model parameters

2.performing gradient-based meta-learning in this low dimensional latent space

- a latent code z
- Conditioned u(encoding)
- generating function v(decoding) : x’ = v(z)
- Classification weights w
- The updated latent code z’
  (decode to new classification weights w’)

임진혁<eeplearning@unist.ac.kr>

Related: Latent Embedding Optimization (Weight Generalization)

Weight Generation Problem:

- **Fixed weights on any query set because it's conditioned on the support set of one task**
- **Query specific information lost while generating weights**
- **It is weak about generating weights for diverse query sample**

Goal of AWGIS:

Retain the information of support/query samples in the generated weights

임진혁<eeplearning@unist.ac.kr>

3 Related: Attention

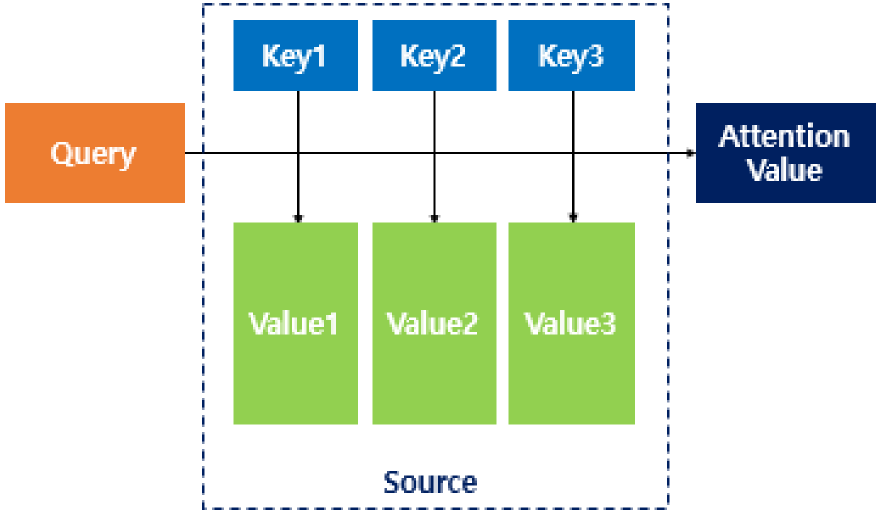
Attention mechanism shows great success in computer vision and natural language processing

It is effective in modeling the interaction between queries and key-value pairs from certain context

- Attention Value : Weighted(attention distribution) "Value" Sum

- attention distribution: *softmax*(attention score)

- attention score : 유사도 (주어진 Query & Key)



임진혁<eeplearning@unist.ac.kr>

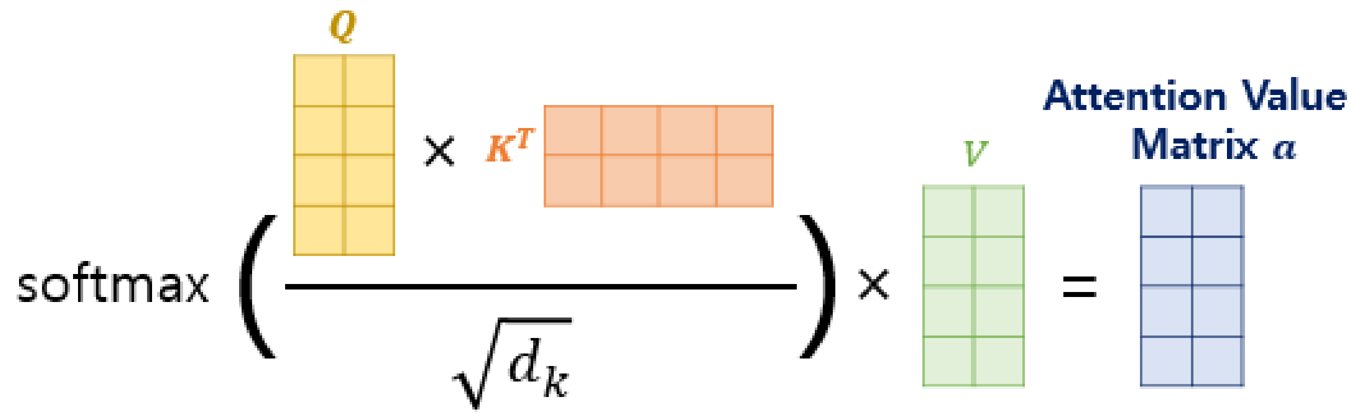
Related: Attention

- Attention Value : Weighted(attention distribution) "Value" Sum

- attention distribution: *softmax*(attention score)

- attention score : 유사도 (주어진 Query & Key)

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



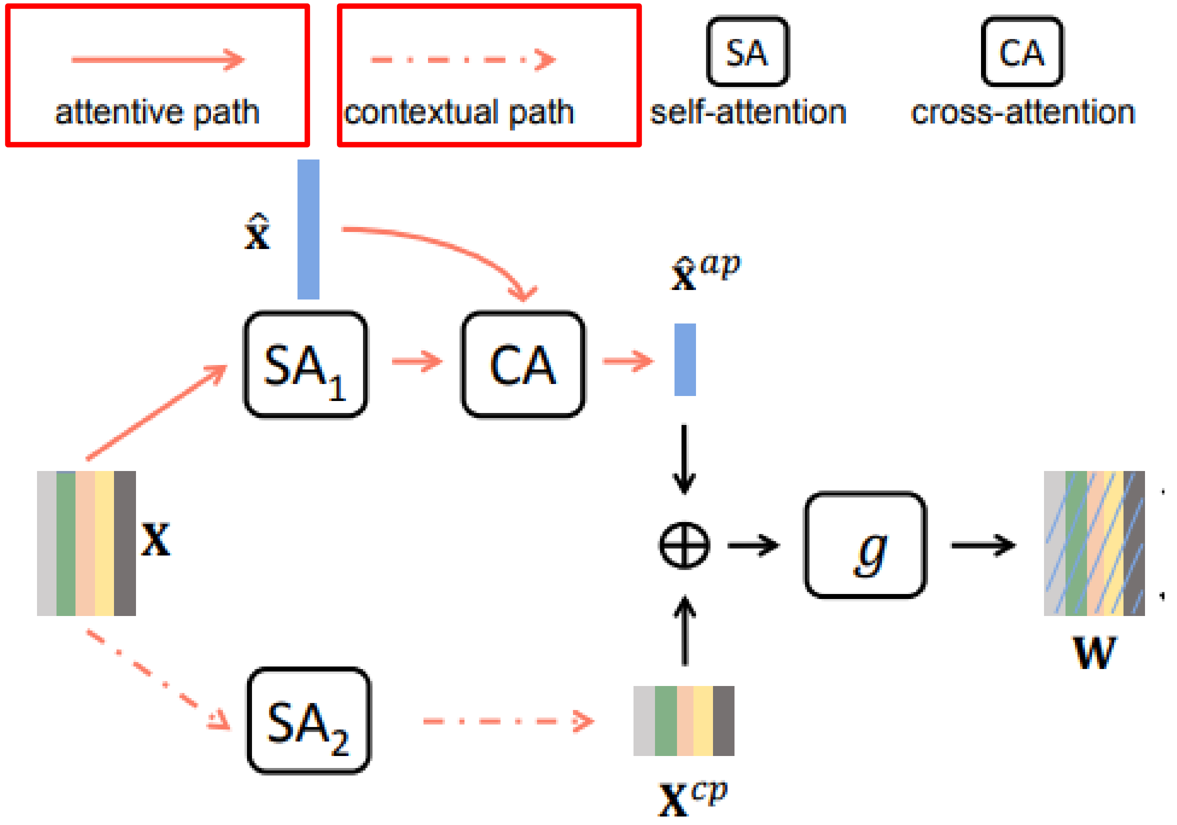
Attention in Proposed Method

Existing weights generation methods conditioned on the Support Set only (LEO Included)

AWGIS:
Separate 2 paths to encode task context and individual **query sample**

- **Contextual path**

- **Attentive path**



Attention in Proposed Method

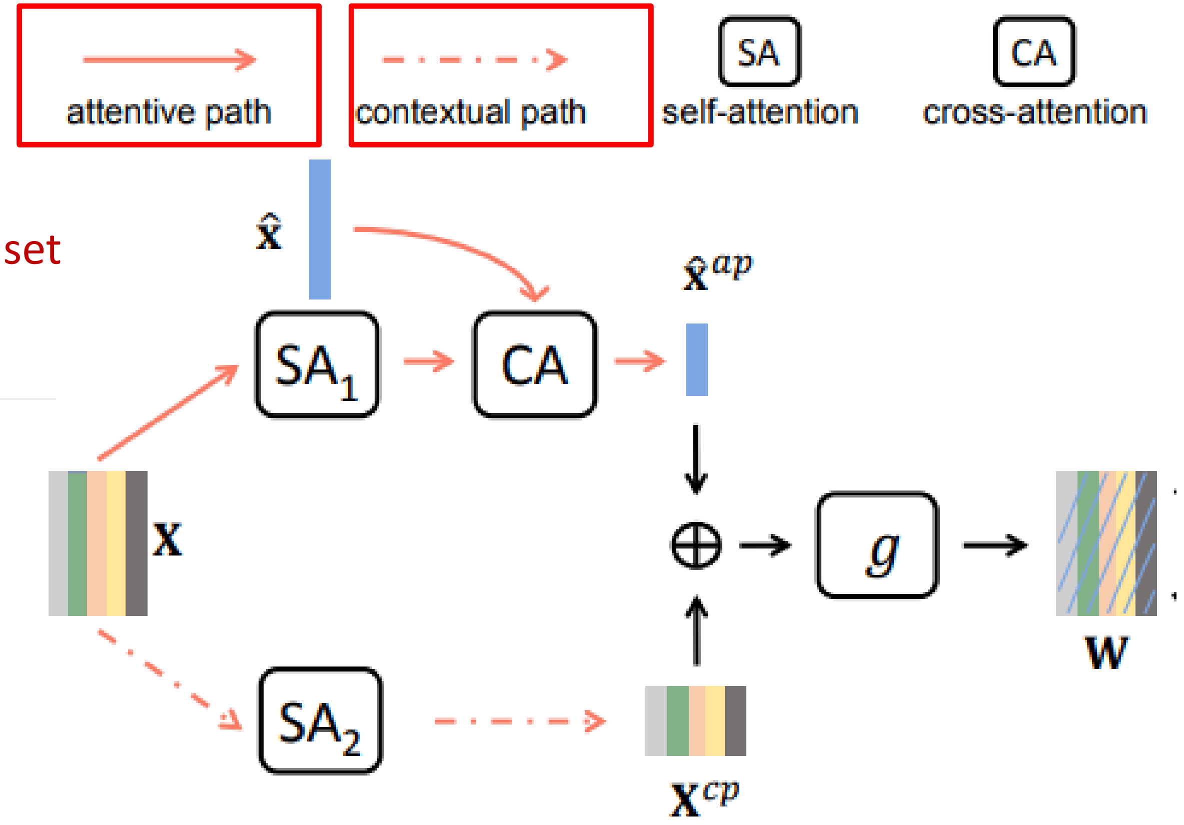
Existing weights generation methods conditioned on the Support Set only (LEO Included)

AWGIS:

Separate 2 paths to encode task context and individual **query sample**

- **Contextual path**
  - encode task context
  - learning representations for only the support set

$$\mathbf{X}^{cp} = f_{\theta^{sa}_{cp}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$



No optimal methods to adapt to different Query samples

Why?

임진혁<eeplearning@unist.ac.kr>

Attention in Proposed Method

AWGIS: Separate 2 paths to encode task context and individual **query sample**

- Contextual path

- encode task context
- learning representations for only the support set

$$\mathbf{X}^{cp} = f_{\theta^{sa}_{cp}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$

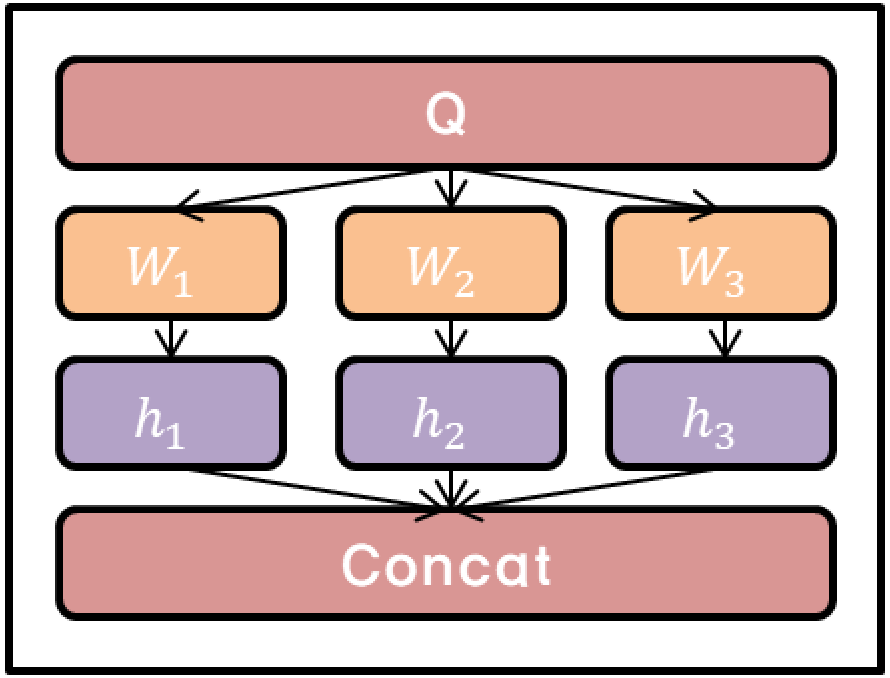
- Attentive path

- encode individual query sample
- learning representations

$$\mathbf{X}^{ap} = f_{\theta^{sa}_{ap}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$

$$\boxed{\hat{\mathbf{x}}^{ap} = f_{\theta^{ca}_{ap}}(Q = \hat{\mathbf{x}}, K = \mathbf{X}, V = \mathbf{X}^{ap})}$$

Multi – head Attention



Attention in Proposed Method

AWGIS: Separate 2 paths to encode task context and individual **query sample**

- Contextual path

$$\mathbf{X}^{cp} = f_{\theta^{sa}_{cp}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$

- Attentive path

$$\mathbf{X}^{ap} = f_{\theta^{sa}_{ap}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$

$$\boxed{\hat{\mathbf{x}}^{ap} = f_{\theta^{ca}_{ap}}(Q = \hat{\mathbf{x}}, K = \mathbf{X}, V = \mathbf{X}^{ap})}$$

Multi – head Attention

MultiHead $(Q, K, V) =$ Concat (head $_1, \ldots,$ head $_H$) W^O

$$\mathrm{head}_j(Q^j, K^j, V^j) = \mathrm{Attention}(Q^j, K^j, V^j),$$

Attention $(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right)$

$$Q^j = QW_Q^j, K^j = KW_K^j, V^i = VW_V^j$$

W_Q^j, W_K^j, W_V^j are the weight matrices for **j** th **head**

임진혁<eeplearning@unist.ac.kr>

Attention in Proposed Method

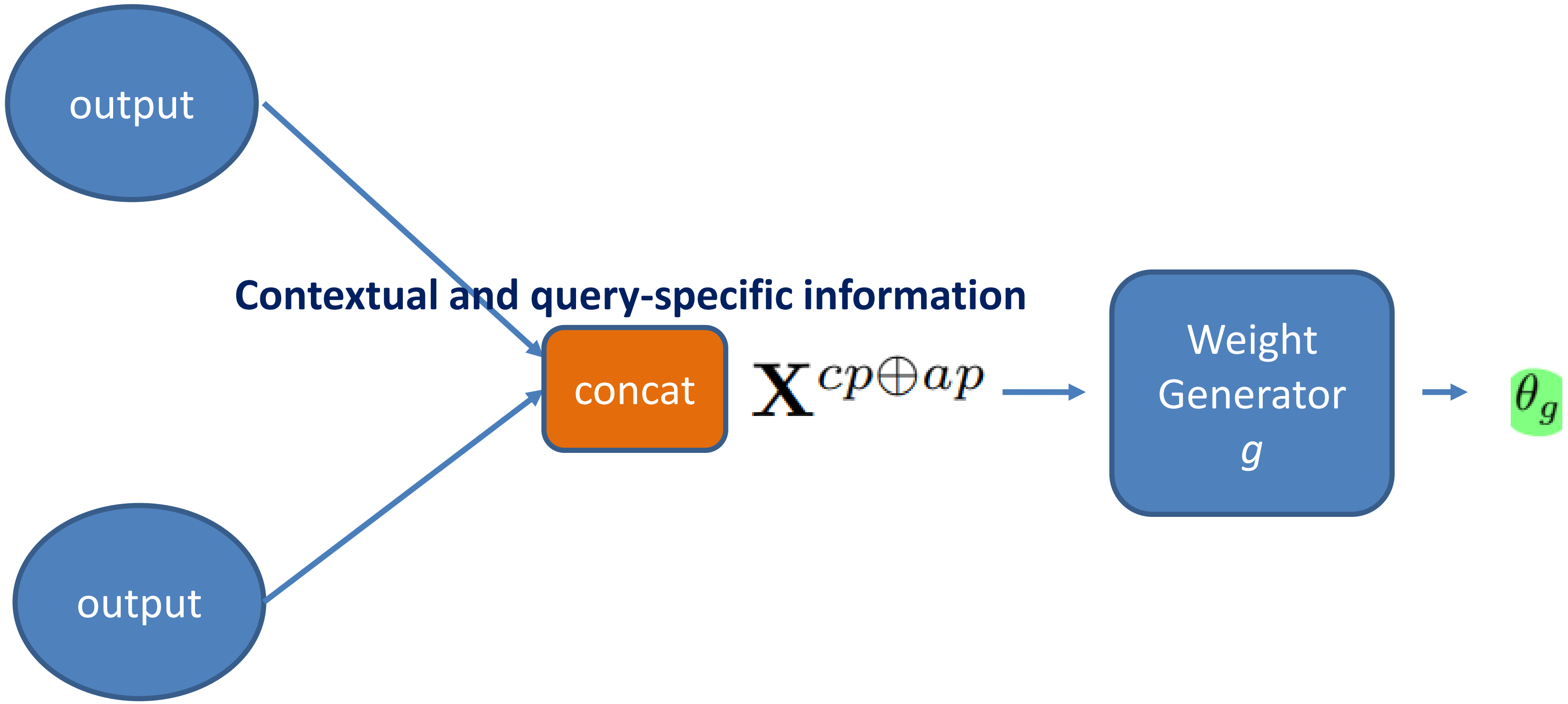
AWGIS: Separate 2 paths to encode task context and individual **query sample**

- **Contextual path**

$$\mathbf{X}^{cp} = f_{\theta^{sa}_{cp}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$

- **Attentive path**

$$\boxed{\hat{\mathbf{x}}^{ap} = f_{\theta^{ca}_{ap}}(Q = \hat{\mathbf{x}}, K = \mathbf{X}, V = \mathbf{X}^{ap})}$$



임진혁<eeplearning@unist.ac.kr>

Attention in Proposed Method

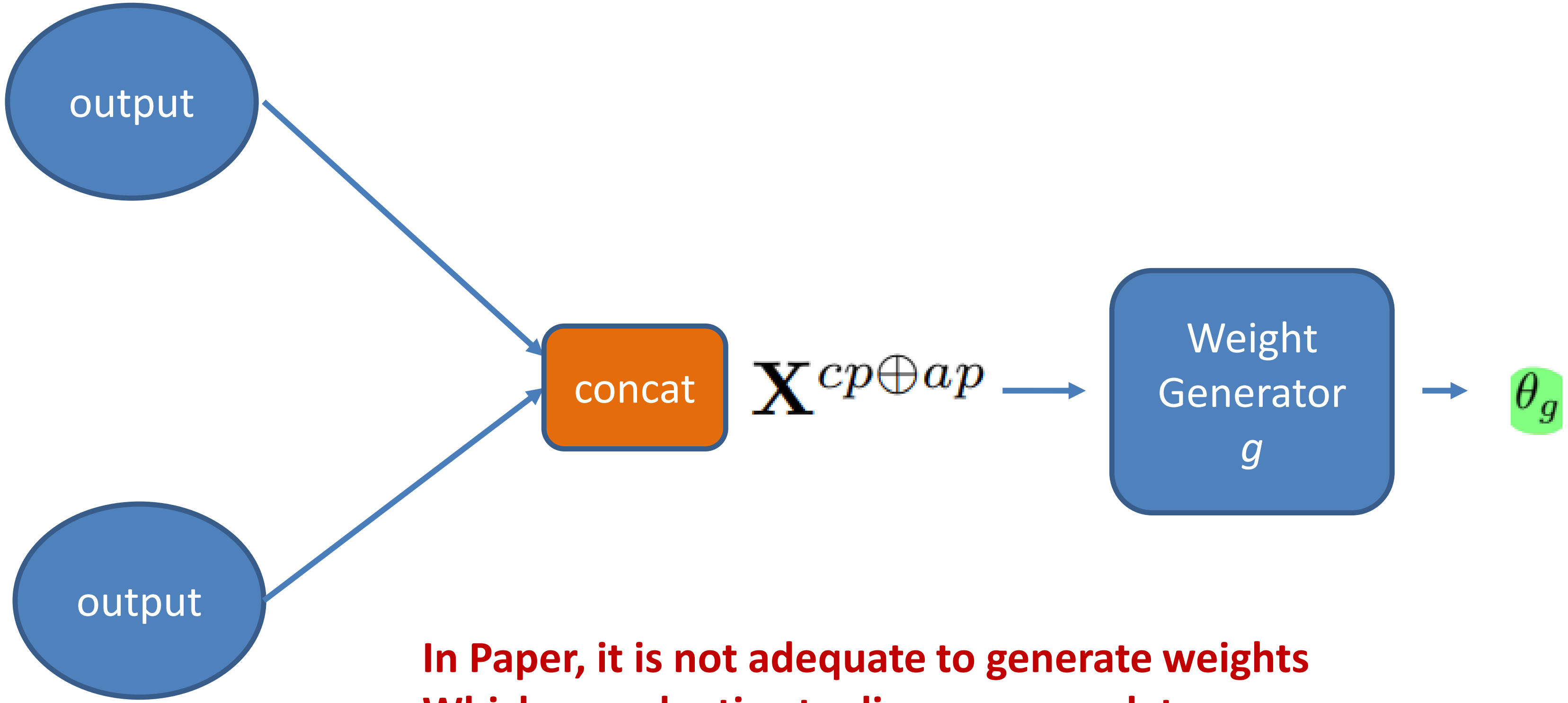
AWGIS: Separate 2 paths to encode task context and individual **query sample**

- **Contextual path**

$$\mathbf{X}^{cp} = f_{\theta_{cp}^{sa}}(Q = \mathbf{X}, K = \mathbf{X}, V = \mathbf{X})$$

- **Attentive path**

$$\boxed{\hat{\mathbf{x}}^{ap} = f_{\theta_{ap}^{ca}}(Q = \hat{\mathbf{x}}, K = \mathbf{X}, V = \mathbf{X}^{ap})}$$



In Paper, it is not adequate to generate weights
Which are adaptive to diverse query data

Why?

Related: Mutual information maximization

What is Mutual Information?

Given two random variables x and y, Mutual Information ***I*(x,y)** measures the decrease of uncertainty in one variable **when another is known.**

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- X에서 Y로부터 설명될 수 있는 정보량
- 또는 Y가 관측되었을 때, X에서 사라지는 불확실성

Q: 언제 상호정보량이 0일까?

임진혁<eeplearning@unist.ac.kr>

Related: Mutual information maximization

What is Mutual Information?

Given two random variables x and y, Mutual Information ***I*(x,y)** measures the decrease of uncertainty in one variable **when another is known.**

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

- X에서 Y로부터 설명될 수 있는 정보량
- 또는 Y가 관측되었을 때, X에서 사라지는 불확실성

Q: 언제 상호정보량이 0일까?

$$I(\mathrm{x};\mathrm{y}) = D_{\mathrm{KL}}(p(\mathrm{x},\mathrm{y}) \| p(\mathrm{x}) \otimes p(\mathrm{y})).$$

$$p(\mathrm{x},\mathrm{y}) = p(\mathrm{x}) \otimes p(\mathrm{y})$$

$$I(\mathrm{x},\mathrm{y}) = 0,$$

임진혁<eeplearning@unist.ac.kr>

Related: Mutual information maximization

InfoGAN's methods can help this problem

(Infogan: Interpretable rep- resentation learning by information maximizing generative adversarial nets. NeurIPS 2016)

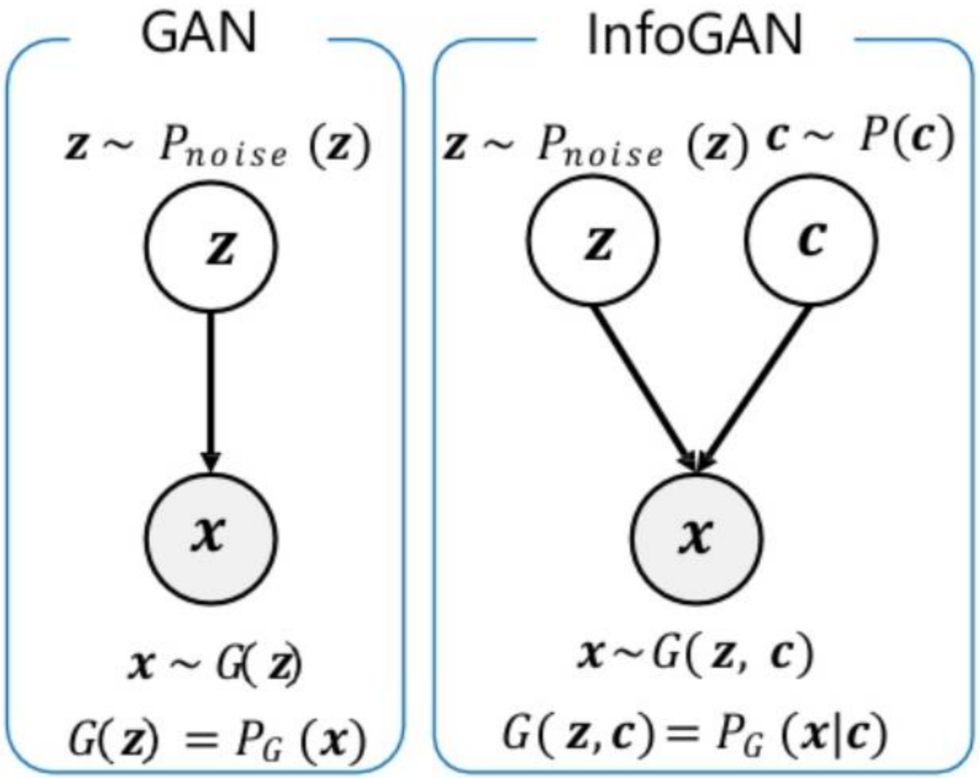
- 생성용 벡터(노이즈) Z를 z & code c로 분할
- Z & C 의 Mutual Information maximization

기존 GAN VS InfoGAN

- GAN은 entangled(얽혀있는) representation을 학습하기에 생성 벡터 z의 어떤 부분이 이미지의 어떤 부분을 관여하는지 알 수 없음(알기힘듬)
  G(z)

- InfoGan은 latent code c로 컨트롤 가능한 disentangled(엉킨것이 풀어진) representation 학습을 제안한다
  G(z,c)

IS THIS WORK?

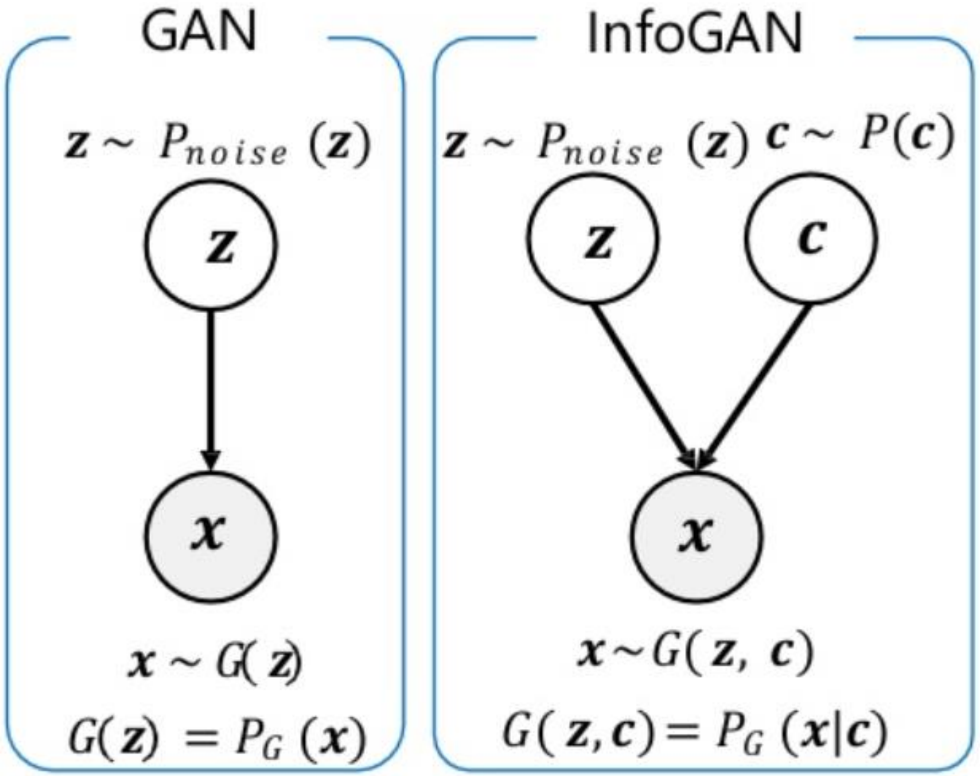


임진혁<eeplearning@unist.ac.kr>

Related: Mutual information maximization

InfoGAN's methods can help this problem

(Infogan: Interpretable rep- resentation learning by information maximizing generative adversarial nets. NeurIPS 2016)



단순히 GAN에 Code c를 추가하기만 하면 기존의 GAN과 달라지지 않는다.
C의 값과 상관없이 학습을 하기 때문(무시)

Generated output에 code c의 정보가 유지되어야 한다

따라서 InfoGAN은 Code C와 Z관 서로 관련이 있도록 Mutual Infroamtion(MI) Maximization을 사용한다.

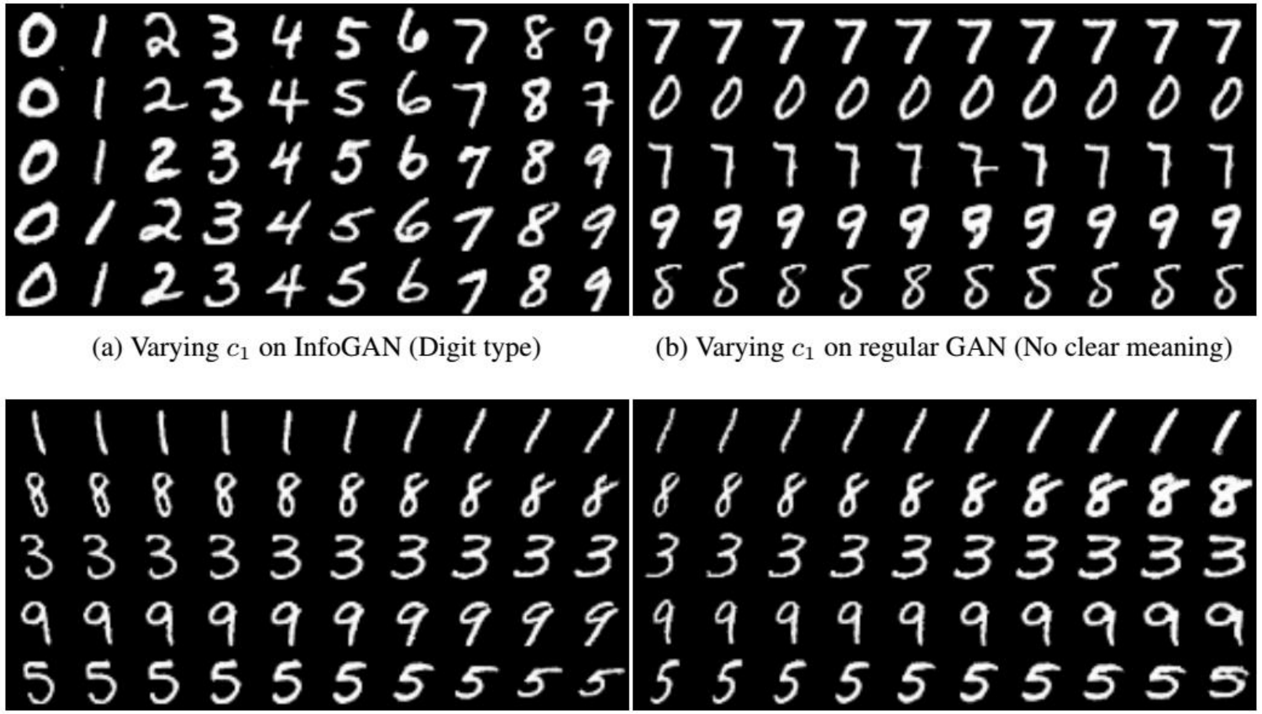
$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$

임진혁<eeplearning@unist.ac.kr>

Related: Mutual information maximization

InfoGAN's methods can help this problem

(Infogan: Interpretable rep- resentation learning by information maximizing generative adversarial nets. NeurIPS 2016)



(a) Varying c_1 on InfoGAN (Digit type) (b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation) (d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Regular Gan's $G(z,c) => G(z)$????무시

$$P_G(x \mid c) = P_G(x)$$

How it can help

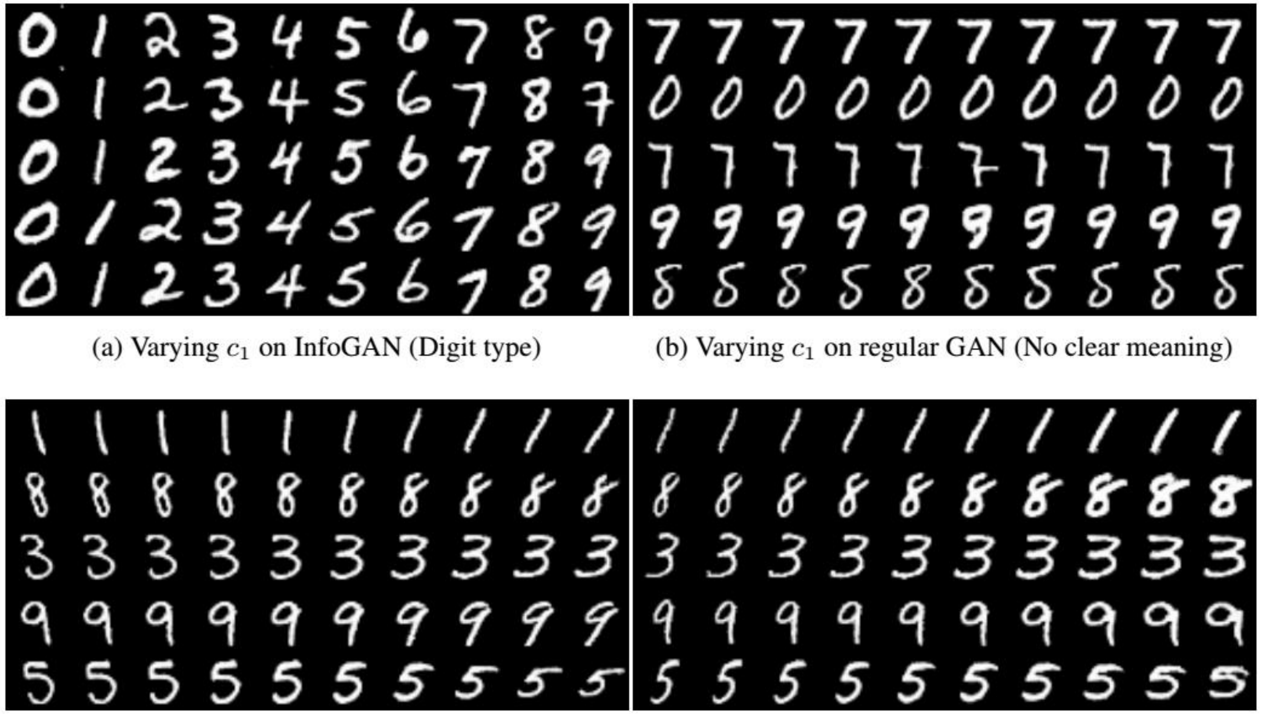
"weight generation for few samples" problem?

임진혁<eeplearning@unist.ac.kr>

3 Related: Mutual information maximization

InfoGAN's methods can help this problem

(Infogan: Interpretable rep- resentation learning by information maximizing generative adversarial nets. NeurIPS 2016)



(a) Varying c_1 on InfoGAN (Digit type)

(b) Varying c_1 on regular GAN (No clear meaning)

(c) Varying c_2 from −2 to 2 on InfoGAN (Rotation)

(d) Varying c_3 from −2 to 2 on InfoGAN (Width)

How it can help

"weight generation for few samples" problem?

AWGIS:

Generated Output에 Q & S information이 남도록

$$I((x', y'); W_i) + I((x_{c_i}, y_{c_i}); W_i)$$

InfoGAN:

Generated image에 Latent code c의 정보가 남도록

$$I(c; G(z, c))$$

임진혁<eeplearning@unist.ac.kr>

Mutual information maximization in Proposed Method

기존의 Weight Generation :

$p(\mathbf{w} \mid \mathcal{T})$ for one task $\mathcal{T}$. $p(\mathbf{w} \mid \mathcal{S})$

AWGIS : encode the query-specific information during generation of weights and learn the model

$p(\mathbf{w} \mid \hat{\mathbf{x}}, \mathcal{S})$

Mutual information maximization

$$\max I((\hat{\mathbf{x}}, \hat{\mathbf{y}}); \mathbf{w}_i) + \frac{1}{K} \sum_K I((\mathbf{x}_{c_i}, \mathbf{y}_{c_i}); \mathbf{w}_i).$$

Chain Rule: $I((\hat{\mathbf{x}}, \hat{\mathbf{y}}); \mathbf{w}_i) = I(\hat{\mathbf{x}}; \mathbf{w}_i) + I(\hat{\mathbf{y}}; \mathbf{w}_i \mid \hat{\mathbf{x}})$

$$\max I(\hat{\mathbf{x}}; \mathbf{w}_i) + I(\hat{\mathbf{y}}; \mathbf{w}_i | \hat{\mathbf{x}}) + \frac{1}{K} \sum_K [I(\mathbf{x}_{c_i}; \mathbf{w}_i) + I(\mathbf{y}_{c_i}; \mathbf{w}_i | \mathbf{x}_{c_i})].$$

Still don't know true posteriori distribution like $p(\hat{\mathbf{y}} \mid \hat{\mathbf{x}}, \mathbf{w}_i), p(\hat{\mathbf{x}} \mid \mathbf{w}_i)$

use **Variational Information Maximization (approximation of lower bound of MI)**

임진혁<eeplearning@unist.ac.kr>

3 Variational Information Maximization

$$
\begin{aligned}
I(\hat{\mathbf{x}};\mathbf{w}_i) &= H(\hat{\mathbf{x}}) - H(\hat{\mathbf{x}}|\mathbf{w}_i) \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p(\hat{\mathbf{x}}|\mathbf{w}_i)]] \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[D_{\mathrm{KL}}(p(\hat{\mathbf{x}}|\mathbf{w}_i)\|p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)) \\
&\quad +\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]] \\
&\geq H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]] \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i,\hat{\mathbf{x}}\sim p(\mathbf{w},\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)] \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}})}[\mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w}|\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]]
\end{aligned} \tag{8}
$$

$$
\begin{aligned}
I(\hat{\mathbf{y}};\mathbf{w}_i|\hat{\mathbf{x}}) \geq\, & H(\hat{\mathbf{y}}|\hat{\mathbf{x}}) + \\
& \mathbb{E}_{\hat{\mathbf{y}}\sim p(\hat{\mathbf{y}}|\hat{\mathbf{x}})}[\mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w}|\hat{\mathbf{y}},\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{y}}|\hat{\mathbf{x}},\mathbf{w}_i)]].
\end{aligned} \tag{9}
$$

$$
\begin{aligned}
\max_\theta \mathbb{E}[\,&\log p_\theta(\hat{\mathbf{y}}|\hat{\mathbf{x}},\mathbf{w}_i) + \log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i) + \\
&\frac{1}{K}\sum_K \log p_\theta(\mathbf{y}_{c_i}|\mathbf{x}_{c_i},\mathbf{w}_i) + \log p_\theta(\mathbf{x}_{c_i}|\mathbf{w}_i)].
\end{aligned} \tag{10}
$$

maximizing the log likelihood can be achieved by minimizing L2 reconstruction loss

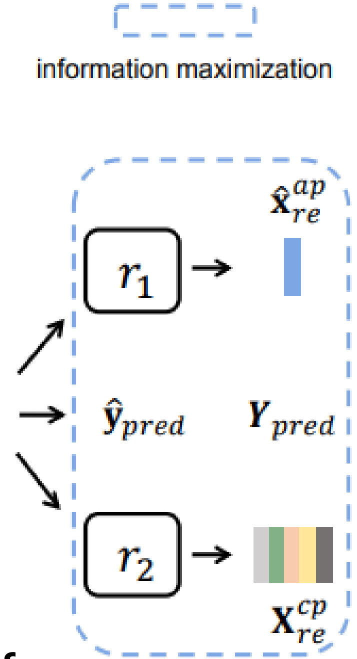
임진혁<eeplearning@unist.ac.kr>

Variational Information Maximization

$$
\begin{aligned}
I(\hat{\mathbf{x}};\mathbf{w}_i) &= H(\hat{\mathbf{x}}) - H(\hat{\mathbf{x}}|\mathbf{w}_i)\\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p(\hat{\mathbf{x}}|\mathbf{w}_i)]]\\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[D_{\mathrm{KL}}(p(\hat{\mathbf{x}}|\mathbf{w}_i)\|p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i))\\
&\quad +\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]]\\
&\geq H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]]\\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i,\hat{\mathbf{x}}\sim p(\mathbf{w},\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]\\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}})}[\mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w}|\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]]
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
I(\hat{\mathbf{y}};\mathbf{w}_i|\hat{\mathbf{x}}) \geq\, & H(\hat{\mathbf{y}}|\hat{\mathbf{x}}) +\\
& \mathbb{E}_{\hat{\mathbf{y}}\sim p(\hat{\mathbf{y}}|\hat{\mathbf{x}})}[\mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w}|\hat{\mathbf{y}},\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{y}}|\hat{\mathbf{x}},\mathbf{w}_i)]].
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
\max_\theta \mathbb{E}[&\log p_\theta(\hat{\mathbf{y}}|\hat{\mathbf{x}},\mathbf{w}_i) + \log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i) +\\
&\frac{1}{K}\sum_K \log p_\theta(\mathbf{y}_{c_i}|\mathbf{x}_{c_i},\mathbf{w}_i) + \log p_\theta(\mathbf{x}_{c_i}|\mathbf{w}_i)].
\end{aligned}
\tag{10}
$$



maximizing the log likelihood can be achieved by minimizing L2 reconstruction loss

Variational Information Maximization

$$
\begin{aligned}
I(\hat{\mathbf{x}};\mathbf{w}_i) &= H(\hat{\mathbf{x}}) - H(\hat{\mathbf{x}}|\mathbf{w}_i) \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p(\hat{\mathbf{x}}|\mathbf{w}_i)]] \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[D_{\mathrm{KL}}(p(\hat{\mathbf{x}}|\mathbf{w}_i)\|p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)) \\
&\quad +\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]] \\
&\geq H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w})}[\mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}}|\mathbf{w}_i)}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]] \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\mathbf{w}_i,\hat{\mathbf{x}}\sim p(\mathbf{w},\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)] \\
&= H(\hat{\mathbf{x}}) + \mathbb{E}_{\hat{\mathbf{x}}\sim p(\hat{\mathbf{x}})}[\mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w}|\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i)]]
\end{aligned}
\tag{8}
$$

$$
\begin{aligned}
I(\hat{\mathbf{y}};\mathbf{w}_i|\hat{\mathbf{x}}) \geq & H(\hat{\mathbf{y}}|\hat{\mathbf{x}}) + \\
& \mathbb{E}_{\hat{\mathbf{y}}\sim p(\hat{\mathbf{y}}|\hat{\mathbf{x}})}[\mathbb{E}_{\mathbf{w}_i\sim p(\mathbf{w}|\hat{\mathbf{y}},\hat{\mathbf{x}})}[\log p_\theta(\hat{\mathbf{y}}|\hat{\mathbf{x}},\mathbf{w}_i)]].
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
\max_\theta \mathbb{E}[\log p_\theta(\hat{\mathbf{y}}|\hat{\mathbf{x}},\mathbf{w}_i) + \log p_\theta(\hat{\mathbf{x}}|\mathbf{w}_i) + \\
\frac{1}{K}\sum_K \log p_\theta(\mathbf{y}_{c_i}|\mathbf{x}_{c_i},\mathbf{w}_i) + \log p_\theta(\mathbf{x}_{c_i}|\mathbf{w}_i)].
\end{aligned}
\tag{10}
$$

maximizing the log likelihood can be achieved by minimizing L2 reconstruction loss

Proposed Approach: AWGIM 전체적인 동작구조

Contribution: learn to generate optimal classification weights for each query samples

-setting: to generate classification weights for one sampled task with few labeled training data

Problem

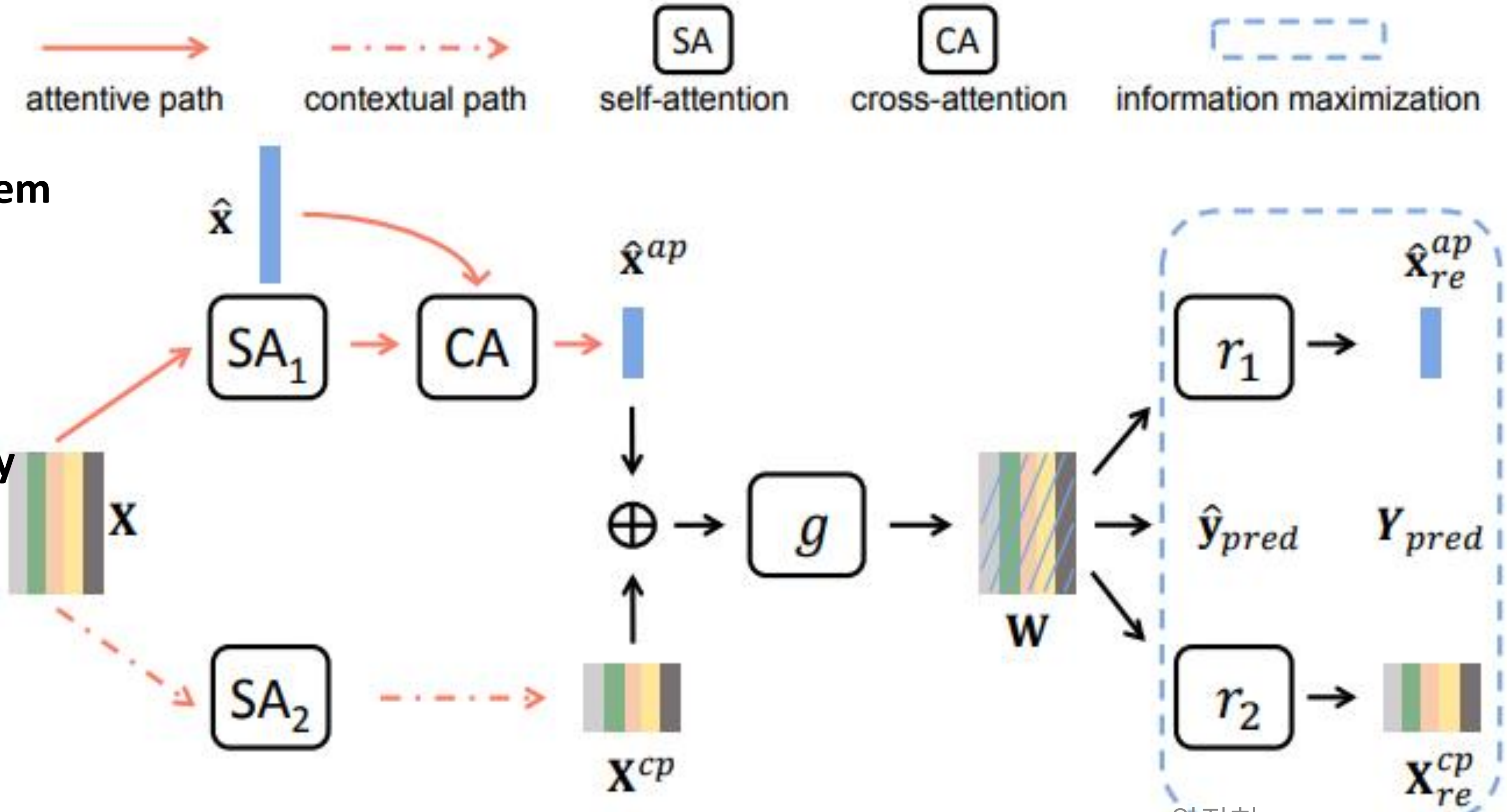
Weight generation for few shot problem

- **Query specific information loss**
- **Can't adopt to diverse query**

Proposed Methods

- Cross attention with individual query

- MI maximization query & weights



암진혁<eeplearning@units.ac.kr>

4 Experiments

데이터:
- mini-ImageNet (100 classes: 64(meta-train, 16(meta-validation), 20(meta-test))
- tiered-ImageNet (608 classes: 351(meta-train, 97(meta-validation), 160(meta-test))

- same image features extractor in LEO (28 layer wide Residual Net , output: 640 d)
- Randomly sample N classes from meta-training set
  - Support set: N * K
  - Query set: 15
- g, r1 and r2: 2-layer MLPs (256 hidden units)
- Number of heads: 4

- Evaluation: Average accuracy for Query set (600 N way K shot tasks sampled from meta-testing set)

  - 1. Train the model on meta- training set
  - 2. Choose the optimal hyper-parameters by meta- validation results
  - 3. Train the model on (meta- training & meta- validation) together using fixed hyper-parameters

Experiments

Model	Feature Extractor	5-way 1-shot	5-way 5-shot
Matching Networks [43]	Conv-4	46.60	60.00
MAML[11]	Conv-4	$48.70 \pm 1.84\%$	$63.11 \pm 0.92\%$
Meta LSTM [33]	Conv-4	$43.44 \pm 0.77\%$	$60.60 \pm 0.71\%$
Prototypical Nets [39]	Conv-4	$49.42 \pm 0.78\%$	$68.20 \pm 0.66\%$
Relation Nets [41]	Conv-4	$50.44 \pm 0.82\%$	$65.32 \pm 0.70\%$
SNAIL [28]	Resnets-12	$55.71 \pm 0.99\%$	$68.88 \pm 0.92\%$
TPN [26]	Resnets-12	59.46	75.65
MTL [40]	Resnets-12	$61.20 \pm 1.80\%$	75.50 ± 0.80
MetaOptNet [21]	Resnets-12	**$64.09 \pm 0.62\%$**	**$80.00 \pm 0.45\%$**
Dynamic [12]	WRN-28-10	$60.06 \pm 0.14\%$	$76.39 \pm 0.11\%$
Prediction [32]	WRN-28-10	$59.60 \pm 0.41\%$	$73.74 \pm 0.19\%$
DAE-GNN [13]	WRN-28-10	**$62.96 \pm 0.15\%$**	**$78.85 \pm 0.10\%$**
LEO [36]	WRN-28-10	$61.76 \pm 0.08\%$	$77.59 \pm 0.12\%$
AWGIM	WRN-28-10	**$63.12 \pm 0.08\%$**	**$78.40 \pm 0.11\%$**

<mini- Imagenet>

Table 2. Accuracy comparison with other approaches on *tiered*ImageNet. Top 3 results are highlighted.

Model	Feature Extractor	5-way 1-shot	5-way 5-shot
MAML [11]	Conv-4	$51.67 \pm 1.81\%$	$70.30 \pm 1.75\%$
Prototypical Nets [39]	Conv-4	$53.31 \pm 0.89\%$	$72.69 \pm 0.74\%$
Relation Nets [41]	Conv-4	$54.48 \pm 0.93\%$	$71.32 \pm 0.78\%$
TPN [26]	Conv-4	$59.91 \pm 0.96\%$	$72.85 \pm 0.74\%$
MetaOptNet [21]	Resnets-12	$65.81 \pm 0.74\%$	$81.75 \pm 0.53\%$
Dynamic [12]	WRN-28-10	**$67.92 \pm 0.16\%$**	**$83.10 \pm 0.12\%$**
DAE-GNN [13]	WRN-28-10	**$68.18 \pm 0.16\%$**	**$83.09 \pm 0.12\%$**
LEO [36]	WRN-28-10	$66.33 \pm 0.05\%$	$81.44 \pm 0.09\%$
AWGIM	WRN-28-10	**$67.69 \pm 0.11\%$**	**$82.82 \pm 0.13\%$**

<tired- Imagenet>

임진혁<eeplearning@units.ac.kr>

Q:IS THERE Attention effects ?

임진혁<eeplearning@units.ac.kr>

Experiments Q:IS THERE Attention effects ?

Model	*mini*ImageNet 5-way 1-shot	*mini*ImageNet 5-way 5-shot	*tiered*ImageNet 5-way 1-shot	*tiered*ImageNet 5-way 5-shot
LEO	61.76 %	77.59 %	66.33%	81.44 %
Generator in LEO	60.33 %	74.53 %	65.17%	78.77 %
Only context PATH(S) Generator conditioned on S only	61.02%	74.33%	66.22%	79.66%
Generator conditioned on S with IM	62.04%	77.54%	66.43%	81.73%

Self Attention **works well**

임진혁<eeplearning@units.ac.kr>

Experiments Q:IS THERE Attention effects ?

NO attention

MLP encoding (*i.e.* no attention)	62.26%	76.91%	65.84%	79.24%
MLP encoding, $\lambda_1 = \lambda_2 = \lambda_3 = 0$	58.95%	71.68%	63.92%	75.80%
AWGIM (ours)	**63.12%**	**78.40%**	**67.69%**	**82.82%**

Attention **works well**

임진혁<eeplearning@units.ac.kr>

5 읽으면서 든 의문점 (1)

Query sample 하나 하나 갖고 generation 하면 시간이 너무 많이 들지 않을까?(cost)

- All these experiments are conducted on the same computing device
- 다른 소타들이 더 cost가 크다.
- 왜냐? 해당 알고리즘의 시간복잡도는 Q의 크기로 결정되는데
  보통 few shot problem에서 "Q" 크기가 작음, 또한 Inner update가 없어서 생각보다 compution 부담이 적음
- 물론 Q가 커지면 더 느려진다

Convergence speed

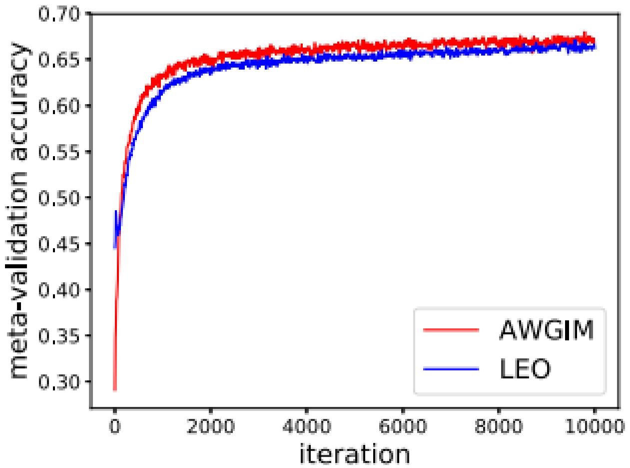


Figure 2. The meta-validation accuracy during meta-training.

Table 5. Inference time cost of AWGIM and MLP encoding.

Method	5-way 1-shot $\|Q\| = 5$	5-way 1-shot $\|Q\| = 50$	5-way 5-shot $\|Q\| = 5$	5-way 5-shot $\|Q\| = 50$
MLP	0.015s	0.031s	0.021s	0.076s
LEO	0.029s	0.032s	0.033s	0.039s
AWGIM	0.019s	0.036s	0.025s	0.079s

5 MIL 읽으면서 든 의문점 (1)

Multi head을 사용했을 때 차이가 있는가?

- single head vs multi-head attention
- “mini-imagenet” 벤치마크에서 실험
- Single head는 그냥 mlp 인코딩과 큰 차이가 없었다(1 shot problem) (충격..)
- Labeld support data가 희박할 수록 Multi head 유무의 차이가 큰 것으로 분석된다.

Multi head attention

Table 4. Accuracy results on *mini*ImageNet with 4 heads or single head in attention networks.

Method	5-way 1-shot	5-way 5-shot
4 heads	63.12%	78.40%
single head	62.35%	77.75%

임진혁<eeplearning@unist.ac.kr>

감사합니다

임진혁<eeplearning@unist.ac.kr>