Attention-based Ensemble for Deep Metric Learning

ECCV 2018

2021.05.27
Sebin Lee

Review: Meta Batch-Instance Normalization

- **Propose MetaBIN that improve generalization ability by unsuccessful generalization scenarios in a meta-learning manner.**

- **MetaBIN:** $\mathbf{y} = \rho\left(\gamma_B \cdot \hat{\mathbf{x}}_B + \beta_B\right) + (1-\rho)\left(\gamma_I \cdot \hat{\mathbf{x}}_I + \beta_I\right)$

- **Meta-train stage**
  - **Over-style normalization: scatter loss, shuffle loss**
  - **Under-style normalization: triplet loss**

- **Meta-test stage**
  - $\theta_\rho \leftarrow \theta_\rho - \gamma \nabla_{\theta_\rho} \mathcal{L}_{\text{tr}}(\mathcal{X}_T; \theta_f, \theta'_\rho)$

Attention-based Ensemble for Deep Metric Learning

ECCV 2018

2021.05.27
Sebin Lee

Contents

- Background & Motivation
- Our Approach
- Results
- Summary

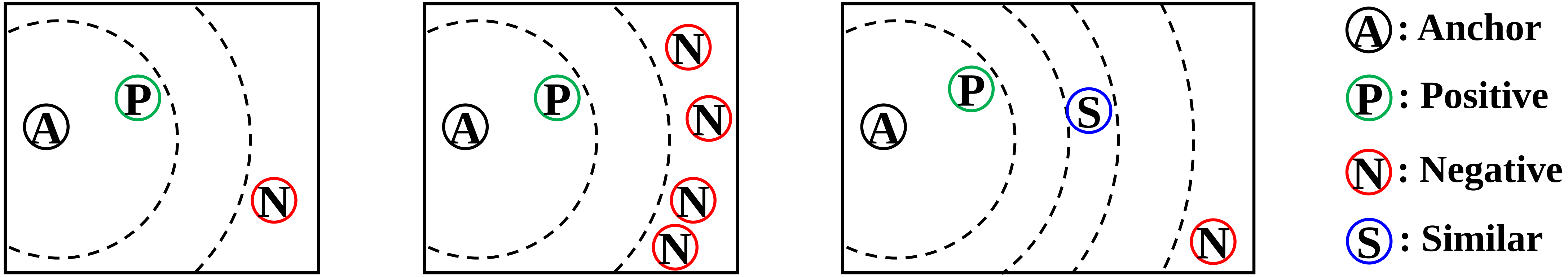
X : input space
Y : embedding space

Deep Metric Learning

- Goal: learn embedding function $f : X \to Y$

  In feature embedding space,
  positive samples are embedded as close as possible,
  negative samples are embedded as separated as possible.

- Ranking Loss Examples

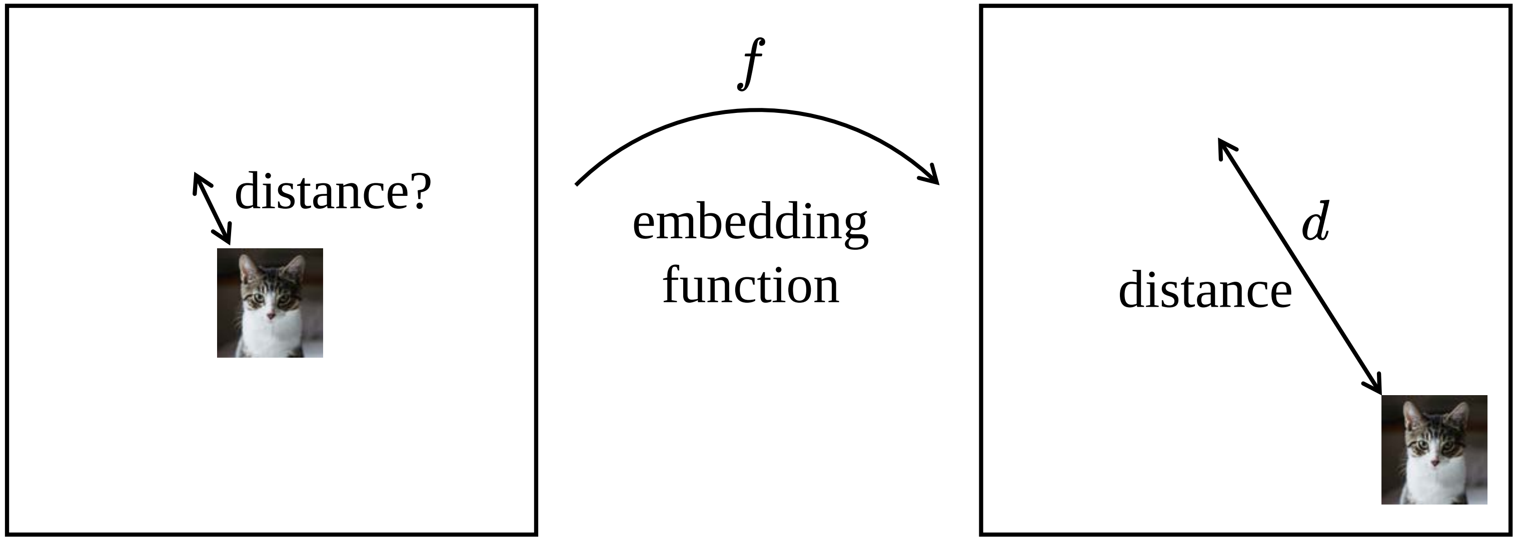


< Triplet Loss > < N-pair Loss > < Quadruplet Loss >

Cat/Dog pictures ack. https://medium.com/@crimy/

Deep Metric Learning: Distance

- **If embedding function maps an image in unknown metric space to known metric space.**
- **We can define distance function(metric function):** $d = ||f(x_i) - f(x_j)||_2$



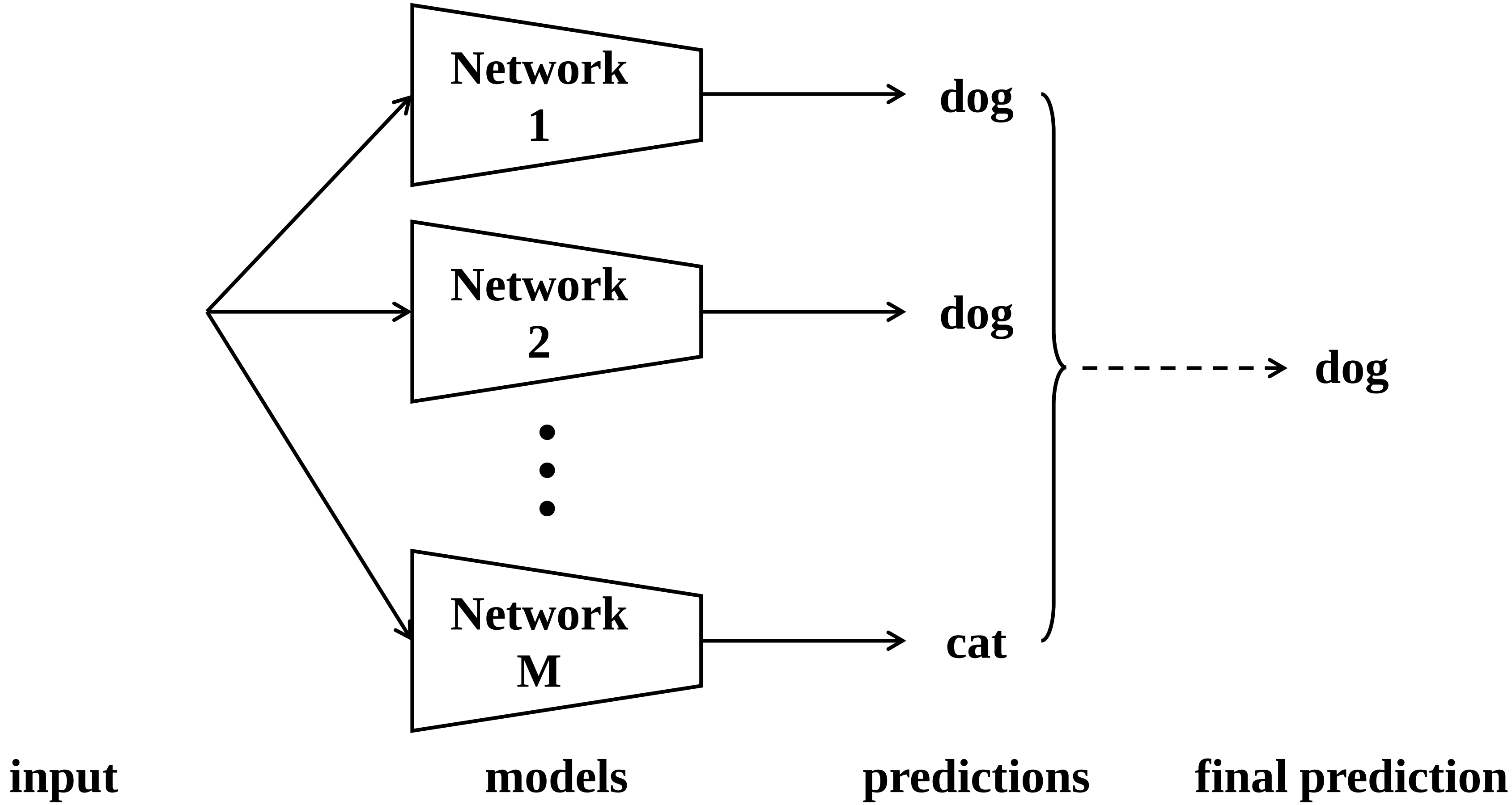
RGB space:
unknown metric space

Embedding space:
known metric space
(e.g., Euclidean space)

Dog picture ack. https://medium.com/@crimy/

Ensemble

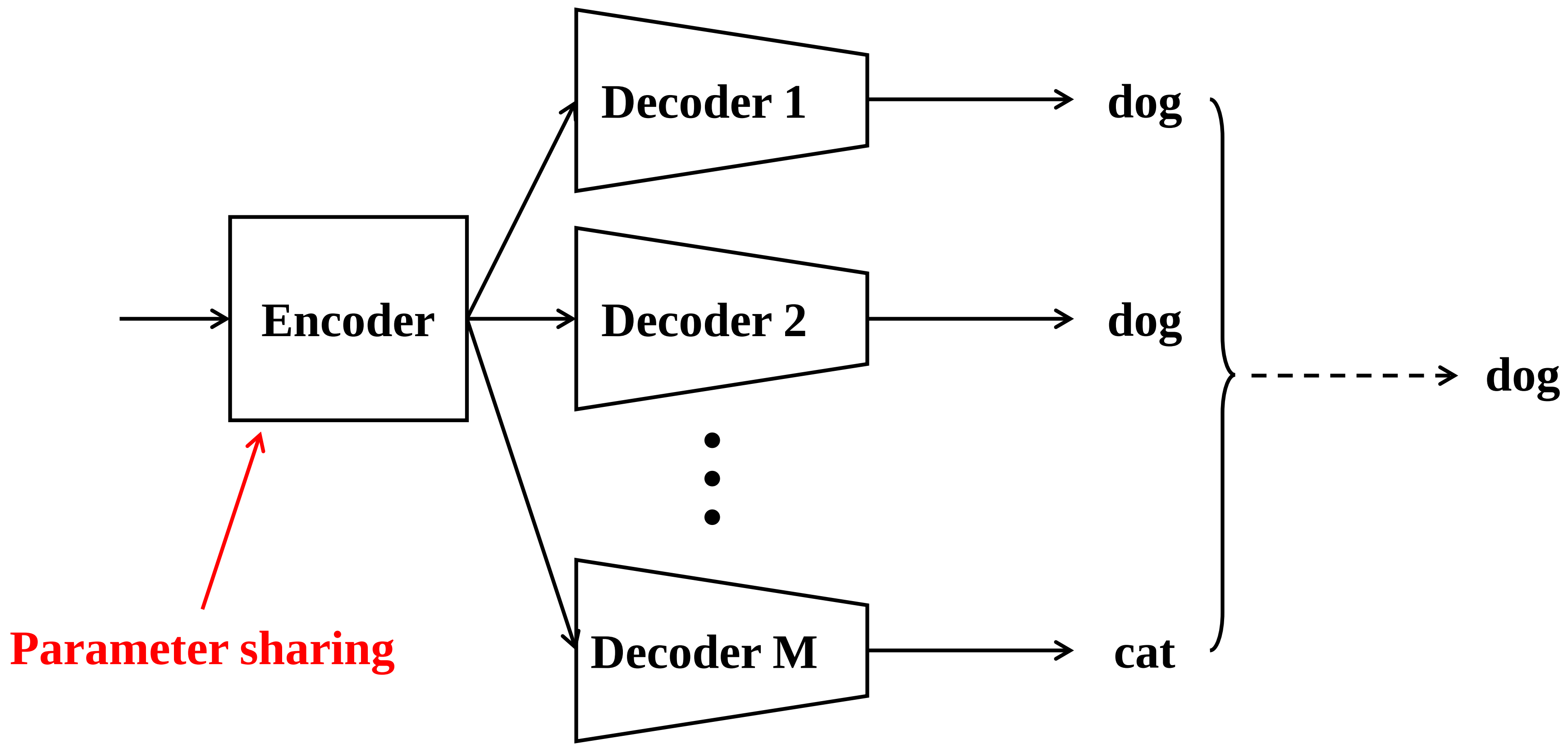
- **Use multiple models to obtain better performance.**



Dog picture ack. https://medium.com/@crimy/

Ensemble: Parameter Sharing

- For efficiency, use parameter sharing.



Ensemble: Required Property

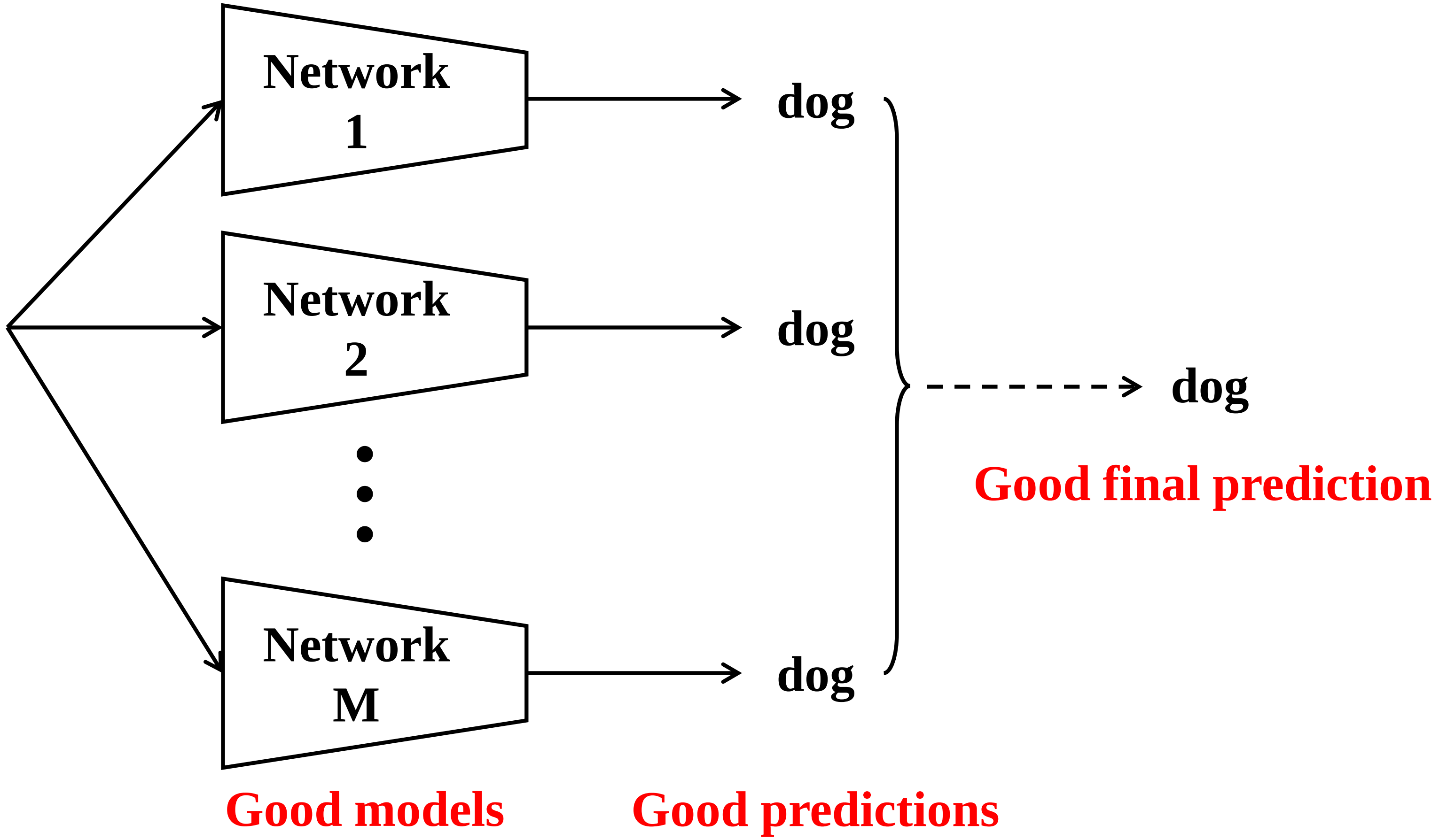
(1) Individual models should have high-performance.

(2) Individual models should be diverse.

Dog picture ack. https://medium.com/@crimy/

Ensemble: Required Property(1)

(1) Individual models should have high-performance.

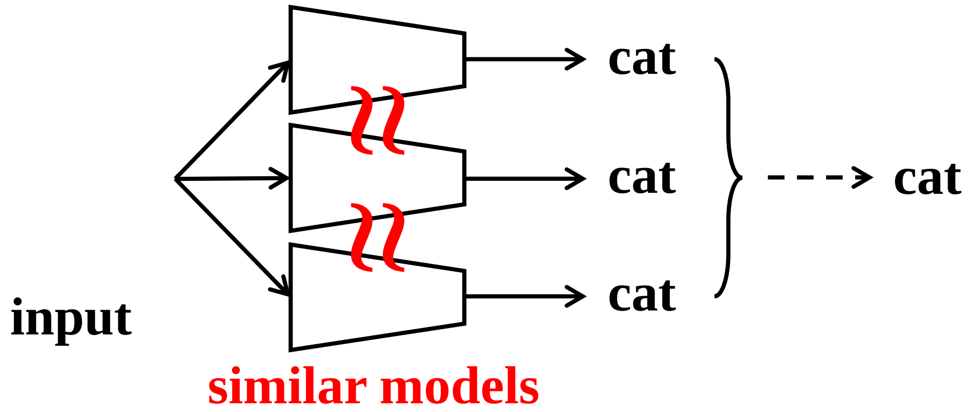


Dog picture ack. https://medium.com/@crimy/

Ensemble: Required Property(2)

(2) Individual models should be diverse.

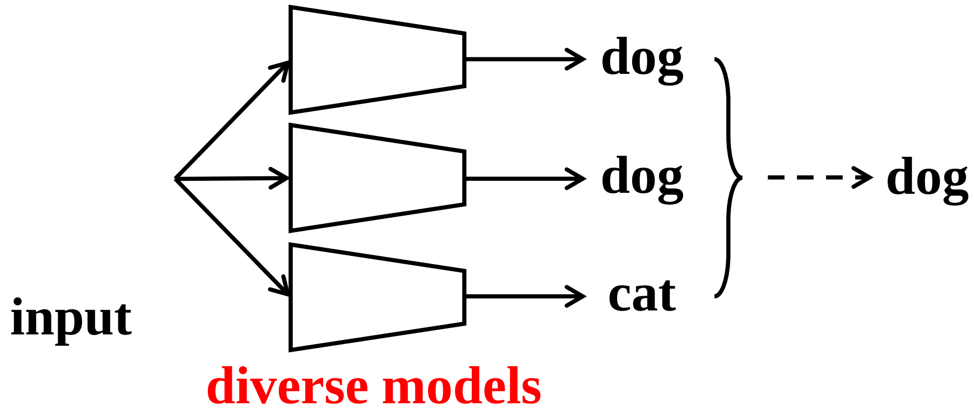
No diversity:



similar parameters
similar features
similar predictions

no advantage of ensemble

Diversity:



different parameters
different features
different predictions

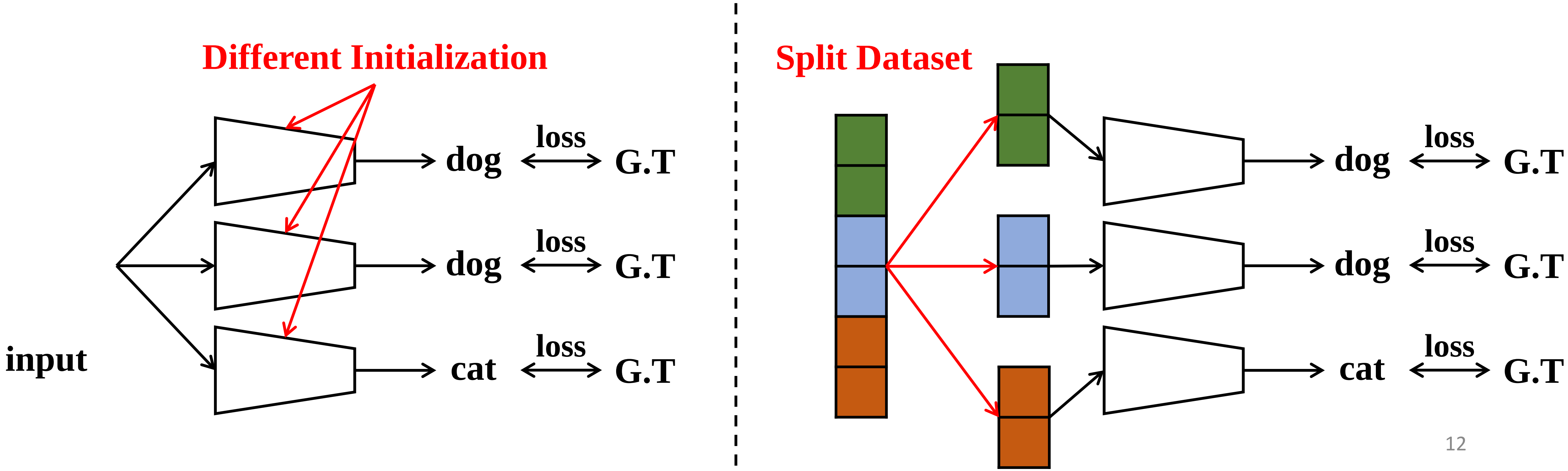
advantage of ensemble

Dog picture ack. https://medium.com/@crimy/

Ensemble for Model Diversity

- **Individual models should be diverse.**

- **Example of early works:**



Motivation & Goal

- **Required properties for Ensemble:**

  (1) Individual models should have high-performance.

  (2) Individual models should be diverse.

- **Ours:**

  (1) Apply **Attention** for high-performance.

  (2) Propose **Divergence Loss** for model diversity.

Contents

Dog pictures ack. https://medium.com/@crimy/

E : **encoder**
D_m : **decoder**

Baseline: M-heads

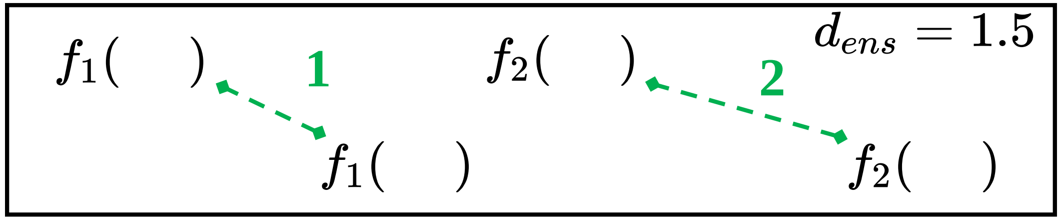
- **Use shared encoder for efficiency(i.e., parameter sharing)**

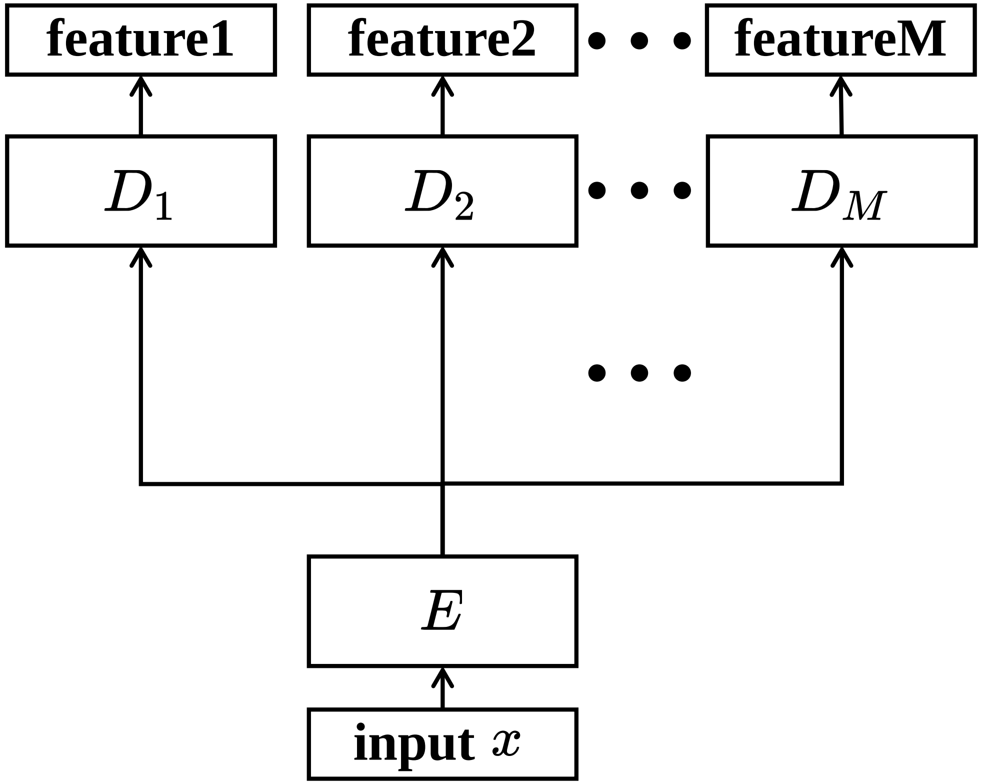
- **We have M models f_m by M decoders.**

$$f_1(x) = D_1(E(x))$$
$$\vdots$$
$$f_M(x) = D_M(E(x))$$

- **Ensemble distance function**

$$d_{ens}(x_i, x_j) = \frac{1}{M}\sum_{m=1}^{M} ||f_m(x_i) - f_m(x_j)||_2$$





< Baseline Architecture >

E : encoder
A_m : attention block
D_m : decoder

Ours: ABE-M(Attention-Based Ensemble with M learners)

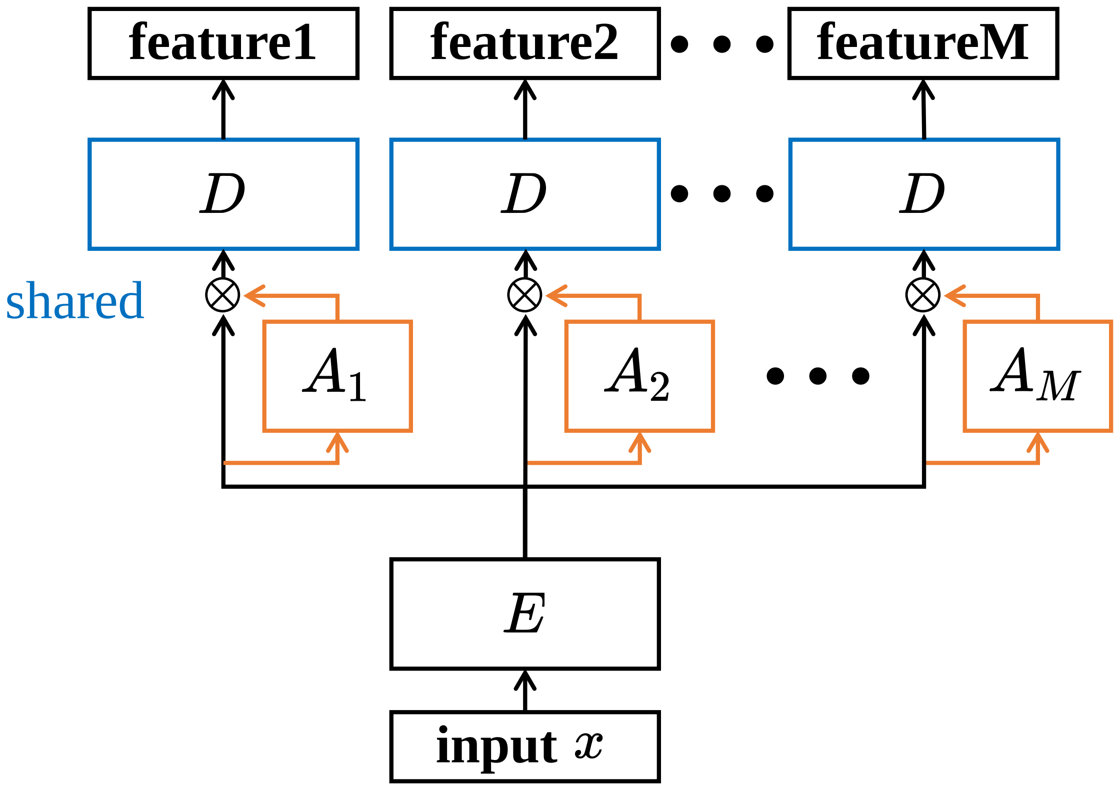
- Add attention block A_m .
- Use shared decoder for efficiency.

- We have M models f_m by M Attentions.

$$f_1(x) = D(E(x) \otimes A_1(E(x)))$$
$$\vdots$$
$$f_M(x) = D(E(x) \otimes A_M(E(x)))$$

- Ensemble distance function

$$d_{ens}(x_i, x_j) = \frac{1}{M} \sum_{m=1}^{M} ||f_m(x_i) - f_m(x_j)||_2$$



< Our Architecture >

Goal

- **Required properties for Ensemble:**

  ~~(1) Individual models should have high-performance.~~

  (2) Individual models should be diverse.

- **Ours:**

  ~~(1) Apply **Attention** for high-performance.~~

  (2) Propose **Divergence Loss** for model diversity.

Cat/Dog pictures ack. https://medium.com/@crimy/

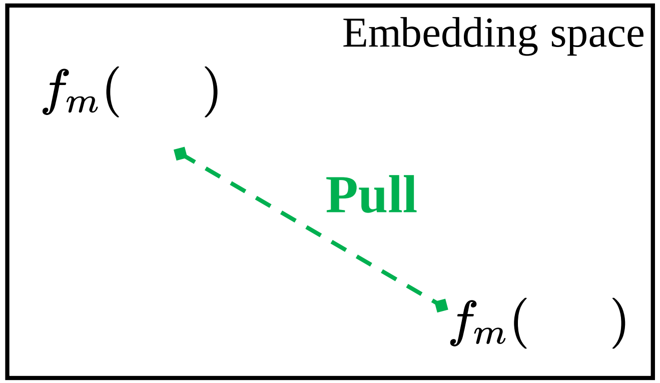
Loss Function: Pairwise Loss

- **Use Pairwise loss with m -th model as ranking loss**

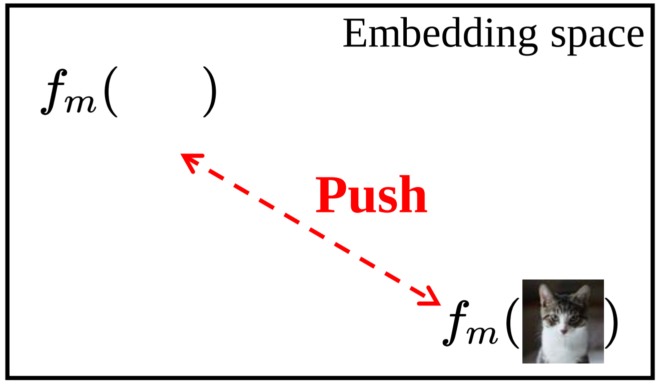
$$L_{pair,(m)} = \frac{1}{N}\sum_{i,j}(1 - y_{ij})\max(0, \text{margin} - ||f_m(x_i) - f_m(x_j)||_2 + y_{i,j}||f_m(x_i) - f_m(x_j)||_2)$$

where $y_{i,j}$ is 1 if x_i , x_j is belong to same class, otherwise 0.

- **Example for Pairwise Loss with m-th model**



if positive samples



if negative samples

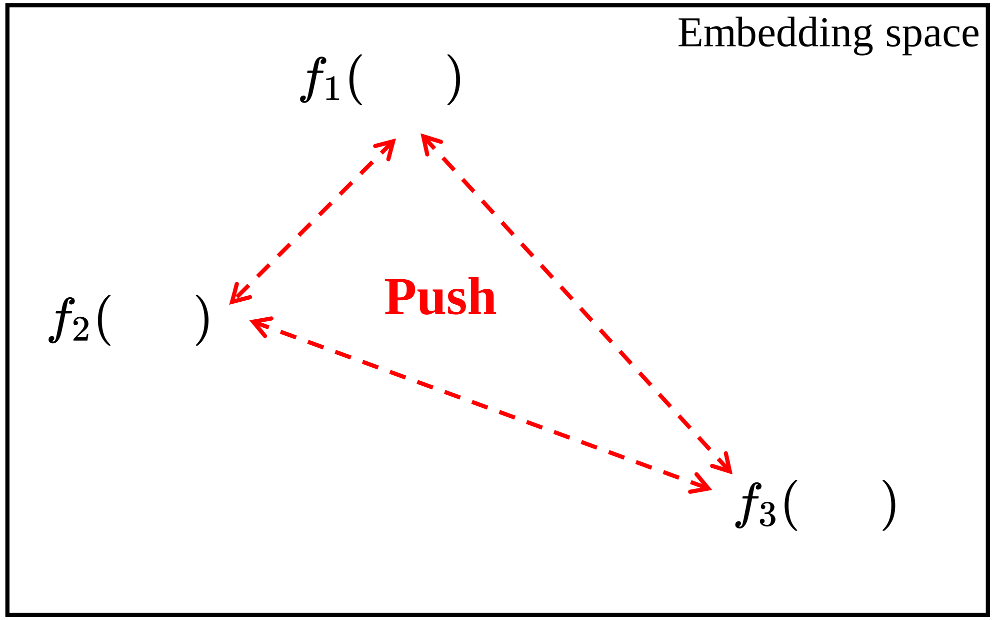
Dog pictures ack. https://medium.com/@crimy/

Loss Function: Divergence Loss

- **Proposed Divergence loss for diversity of individual models.**

$$L_{div} = \sum_i \sum_{p,q} \max(0, \text{margin} - ||f_p(x_i) - f_q(x_i)||_2)$$

- **Divergence loss aims to push the features of same sample from different models.**



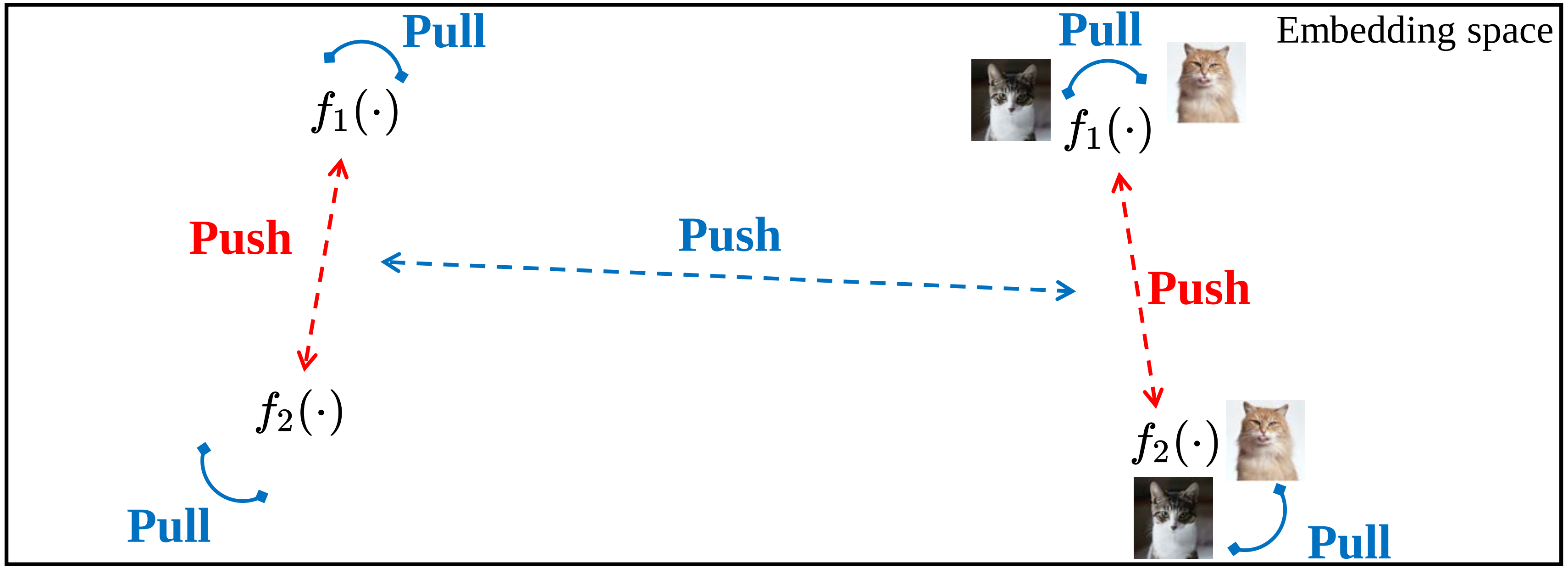
Dog/Cat pictures ack. https://medium.com/@crimy/

Loss Function: Total

- **Total Loss: Pairwise Loss + Divergence Loss**

$$L = \sum_{m=1}^{M} L_{pair,(m)} + \lambda_{div} L_{div}$$

- **Example) 4 samples(, , , ) and 2 models($f_1(\cdot)$, $f_2(\cdot)$)**



Goal

- **Required properties for Ensemble:**

  ~~(1) Individual models should have high-performance.~~

  ~~(2) Individual models should be diverse.~~

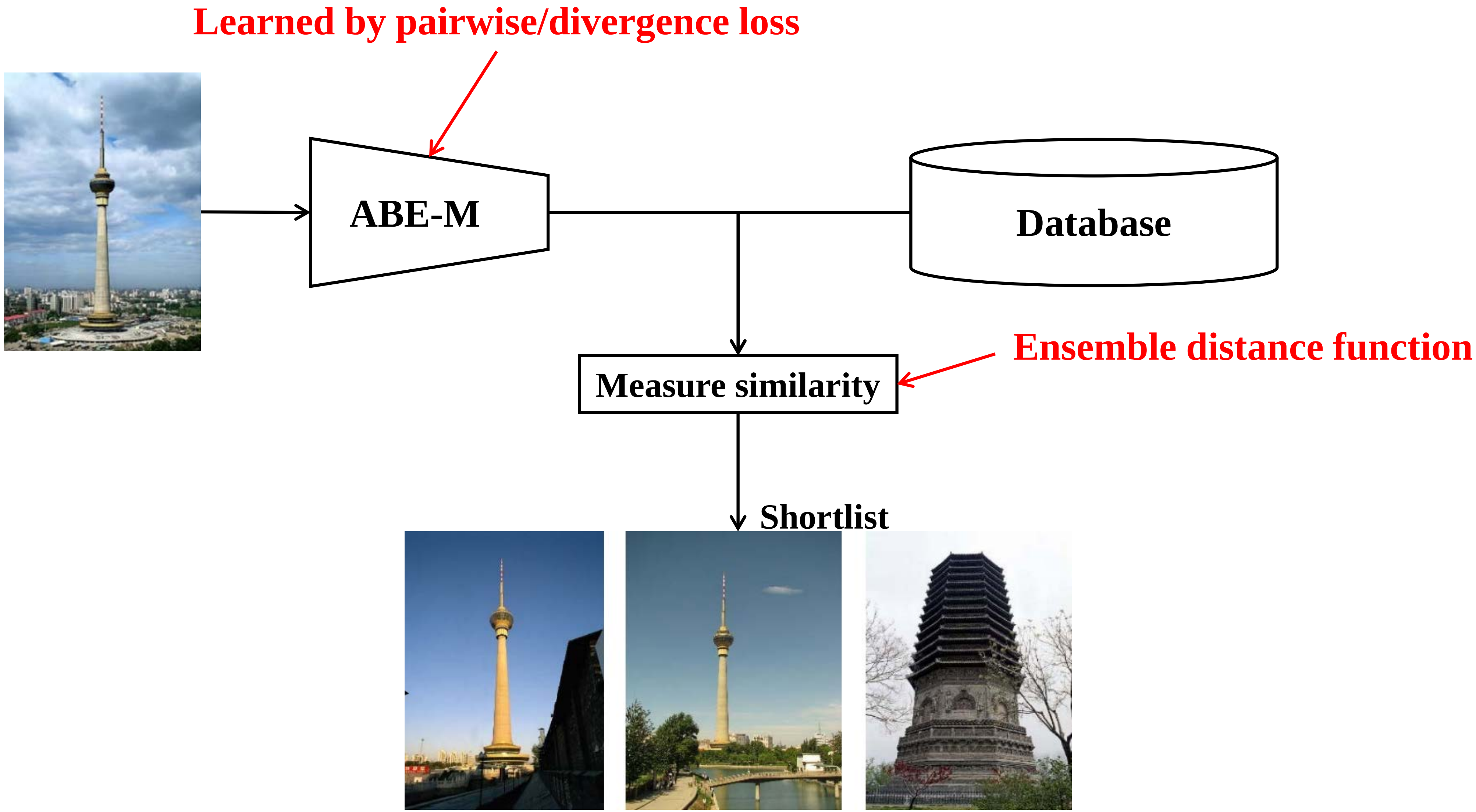
- **Ours:**

  ~~(1) Apply **Attention** for high-performance.~~

  ~~(2) Propose **Divergence Loss** for model diversity.~~

Tower pictures ack. Prof Sung-eui Yoon's lecture slide

Apply to Image Retrieval

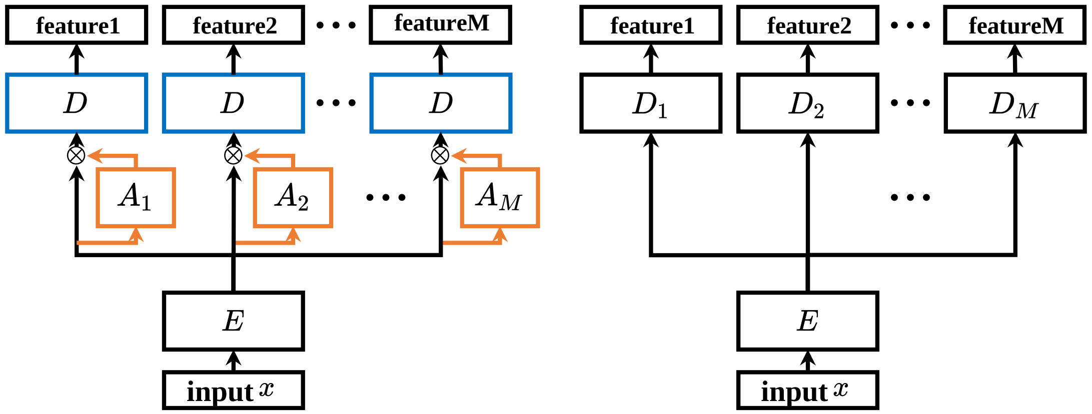


Contents

- **Background & Motivation**
- **Our Approach**
- **Results**
- **Summary**

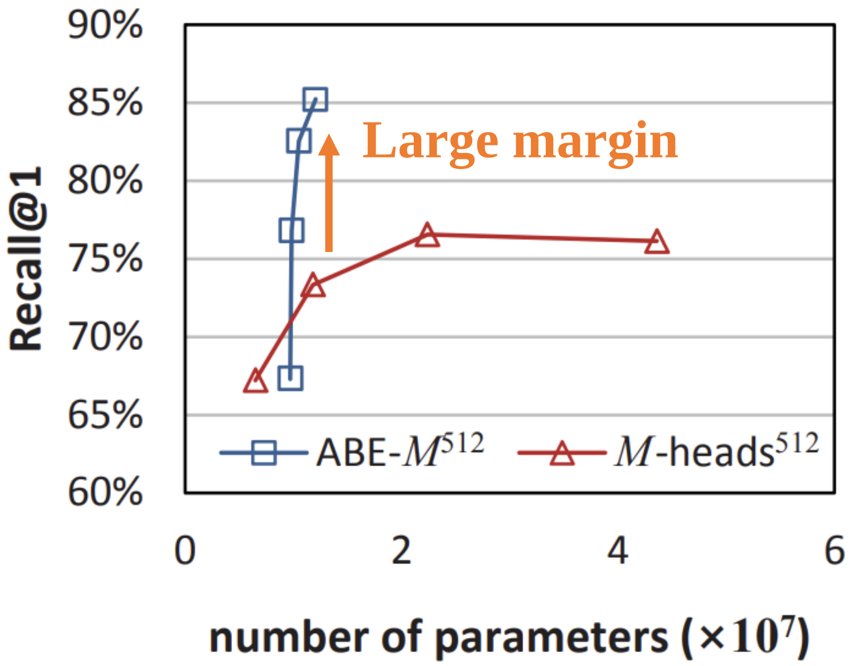
Recall@1: Ours vs Baseline

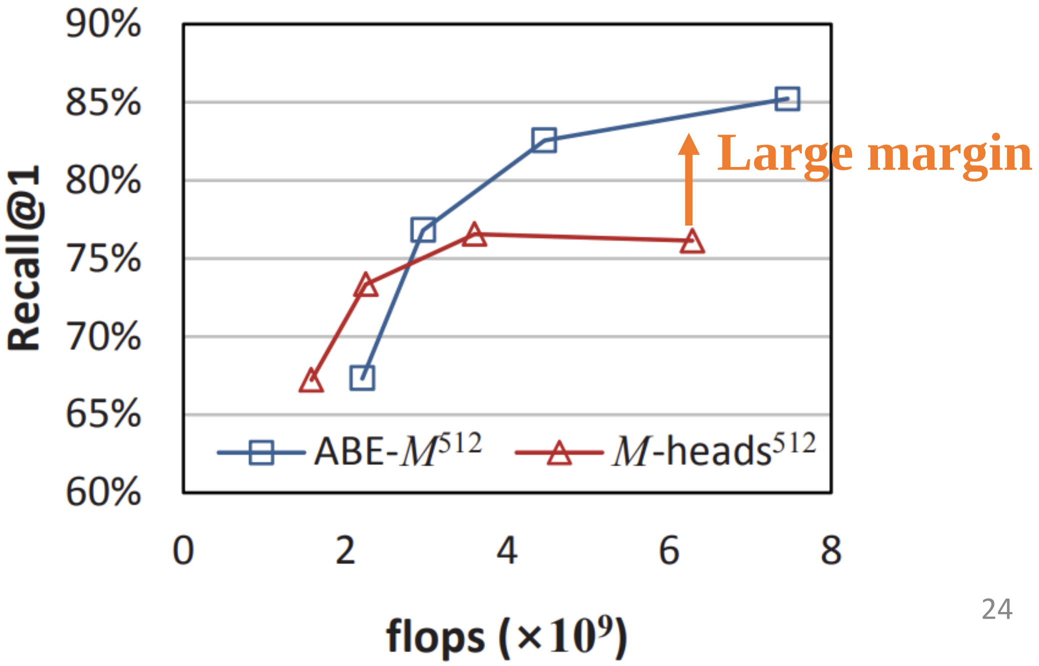
- **Dataset: CARS-196**
- **Embedding dimension: 512**



Ours (ABE-M)

Baseline (M-heads)





Recall@K: Ensemble & Individual Performance

- **M-heads: baseline** e.g., 8-heads^{512} : baseline with 8 models + 512 embedding dim
- **ABE-M: ours** e.g., ABE-8^{512} : ours with 8 models + 512 embedding dim

	K	Ensemble 1	Ensemble 2	Ensemble 4	Ensemble 8	Individual Learners 1	Individual Learners 2	Individual Learners 4	Individual Learners 8	params ($\times 10^7$)	flops ($\times 10^9$)
Baseline	1-head^{512}	67.2	77.4	85.3	90.7	-	-	-	–	0.65	1.58
	2-heads^{512}	73.3	82.5	88.6	93.0	70.2±.03	79.8±.52	86.7±.01	91.9±.37	1.18	2.25
	4-heads^{512}	76.6	84.2	89.3	93.2	70.4±.80	79.9±.38	86.5±.43	91.4±.42	2.24	3.60
	8-heads^{512}	76.1	84.3	90.3	93.9	68.3±.39	78.5±.39	86.0±.37	91.3±.31	4.36	6.28
Ours	ABE-1^{512}	67.3	77.3	85.3	90.9	-	-	-	-	0.97	2.21
	ABE-2^{512}	76.8	84.9	90.2	94.0	70.9±.58	80.3±.04	87.1±.07	92.2±.20	0.98	2.96
	ABE-4^{512}	82.5	89.1	93.0	95.5	74.4±.51	83.1±.47	89.1±.34	93.2±.36	1.05	4.46
	ABE-8^{512}	**85.2**	**90.5**	**93.9**	**96.1**	75.0±.39	83.4±.24	89.2±.31	93.2±.24	1.20	7.46
	ABE-1^{64}	65.9	76.5	83.7	89.3	-	-	-	-	0.92	2.21
	ABE-2^{128}	75.5	84.0	89.4	93.6	68.6±.38	78.8±.38	85.7±.43	91.3±.16	0.96	2.96
	ABE-4^{256}	81.8	88.5	92.4	95.1	72.3±.68	81.4±.45	87.9±.23	92.3±.13	1.04	4.46

Effect of Divergence Loss

- **Recall@K on CARS-196 dataset**

- **Individual learners:** L_{div} **leads to increase performance slightly.**
- **Ensemble :** L_{div} **leads to increase performance significantly.**

K	Ensemble 1	Ensemble 2	Ensemble 4	Ensemble 8	Individual Learners 1	Individual Learners 2	Individual Learners 4	Individual Learners 8
ABE-8^{512}	85.2	90.5	93.9	96.1	75.0±0.39	83.4±0.24	89.2±0.31	93.2±0.24
ABE-8^{512} without L_{div}	69.7	78.8	86.2	91.5	69.5±0.11	78.8±0.14	86.1±0.15	91.5±0.09

significantly increase

slightly increase

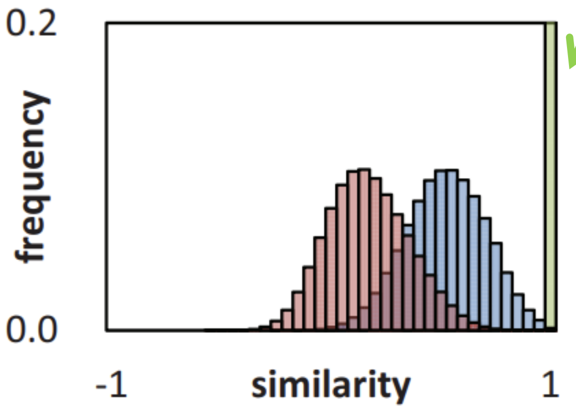
Cat/Dog pictures ack. https://medium.com/@crimy/

Effect of Divergence Loss

- **Divergence loss leads to decrease in cosine similarity of same pair.(i.e., diversity↑)**
- **same pair:** $(f_1(\quad), f_2(\quad))$
- **positive pair:** $(f_1(\quad), f_2(\quad))$
- **negative pair:** $(f_1(\quad), f_2($  $))$

Not diverse

Diverse



→

ABE-8 w/o L_{div}

ABE-8 w/ L_{div}

Qualitative Result: Attention

- **Different learners attend different parts of the car.**
  i.e., Ours satisfies diversity of individual models for ensemble.

Model 1 (upper part)

Model 2 (bottom part)

Model 3 (entire part)

Masked input image by 27th channel of attention mask

Qualitative Result: Attention

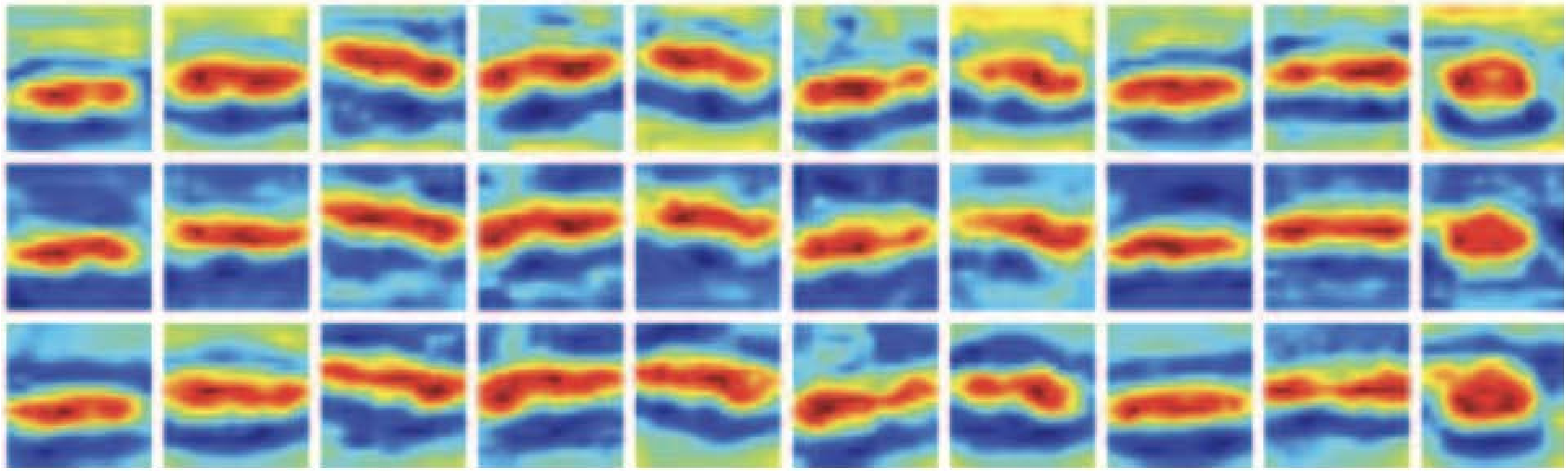
- **All models focus on entire part of the car.**

  i.e., Our attention module can help the model focus on important region.

Model 1

Model 2

Model 3



Mean activation of attention masks across all channels

Recall@K: Comparison with SOTA(1)

- **Dataset: Stanford online products(SOP)**

	K	1	10	100	1000
Other SOTA	Contrastive^{128} [29]	42.0	58.2	73.8	89.1
	LiftedStruct^{512} [29]	62.1	79.8	91.3	97.4
	N-Pairs^{512} [27]	67.7	83.8	93.0	97.8
	Clustering^{64} [28]	67.0	83.7	93.2	-
	Proxy NCA†^{64} [20]	73.7	-	-	-
	Margin†^{128} [38]	72.7	86.2	93.8	98.0
	HDC^{384} [39]	69.5	84.4	92.8	97.7
	A-Bier^{512} [23]	74.2	86.9	94.0	97.8
Ours	ABE-2^{512}	75.4	88.0	94.7	**98.2**
	ABE-4^{512}	75.9	88.3	94.8	**98.2**
	ABE-8^{512}	**76.3**	**88.4**	**94.8**	**98.2**

Recall@K: Comparison with SOTA(2)

- **Dataset: In-shop Clothes Retrieval Benchmark**

	K	1	10	20	30	40	50
Other SOTA	FasionNet+Joints^{4096} [18]	41.0	64.0	68.0	71.0	73.0	73.5
	FasionNet+Poselets^{4096} [18]	42.0	65.0	70.0	72.0	72.0	75.0
	FasionNet^{4096} [18]	53.0	73.0	76.0	77.0	79.0	80.0
	HDC^{384} [39]	62.1	84.9	89.0	91.2	92.3	93.1
	A-BIER^{512} [23]	83.1	95.1	96.9	97.5	97.8	98.0
Ours	ABE-2^{512}	85.2	96.0	97.2	97.8	98.2	98.4
	ABE-4^{512}	86.7	96.4	97.6	98.0	98.4	98.6
	ABE-8^{512}	**87.3**	**96.7**	**97.9**	**98.2**	**98.5**	**98.7**

Contents

- **Background & Motivation**
- **Our Approach**
- **Results**
- **Summary**

Contributions

- **Satisfy required properties for Ensemble**

  (1) Individual models should have high-performance.

  (2) Individual models should be diverse.

- **by**

  (1) Apply **Attention** for high-performance.

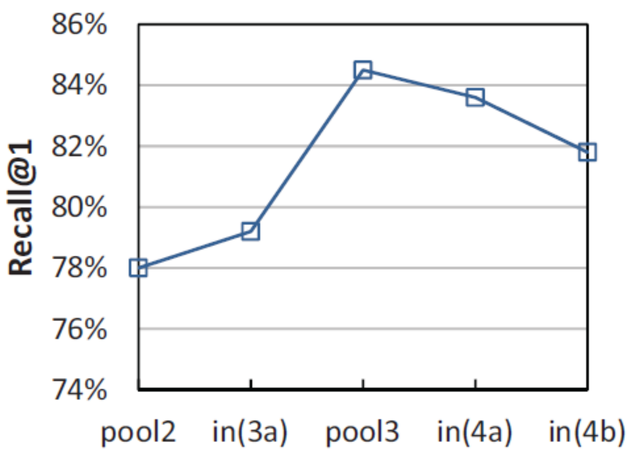
  (2) Propose **Divergence Loss** for model diversity.

- **As a results, Achieve SOTA performance in image retrieval task**

Strengths & Weaknesses

- **Strengths**
  - **Attention increases the performance of individual models.**
  - **Proposed divergence loss encourages individual models to extract features keeping diversity.**
  - **Hence, proposed method improves the performance of individual model and the diversity, thereby increasing the performance of the ensemble model.**

- **Weaknesses**
  - **The proposed method should experiment to find the best place to insert the attention module for given backbone network.**
  - **Performance changes a lot depending on the place of attention module.**



place of attention module

THANK YOU

Quiz

- **Please submit this google form.**

Link will be posted in the zoom session.