ViT에서 일반화와 Shift 등 특성에 관한 연구

* Introduction.

Transformer의 성능은 vision 에서 큰 성능을 보였다.

TF는 generalization 능력의 이유는 조사되지 않았다.

보통 train, test는 같은 분포에 있다고 가정하지만, 안그럴 때가 많다.

따라서 Out-of-Distribution의 성능은 중요하다.

∴ TF의 Distribution Shift (DS)에 대해 조사.

DS는 전경과 배경에 대한 정보가 포함된다.

전경은 pixel, texture, shape, part가 있다.

DS는 semantic이 다르므로 background, corruption, texture, style shift로 묶인다.

좋은 OOD를 위해선 인간의 비슷한 inductive bias를 가져야 함

↳ ∴ TF의 inductive bias가 있는지 조사.

↳ 1) ViT는 background, texture에 약한 bias 를 가지고. → 인간과 비슷
shape, style에 큰 bias를 가짐

∴ CNN보다 잘 일반화, OOD ↑

2) Model이 커짐에 따라, bias가 증가.

↳ corruption, background shift가 증가하여, IID 및 OOD의 정확이 좋음.

3) 큰 patch로 훈련은 texture shift 에만 좋고, 나머지 성능 ↓

이후 논문에서 generalization 이 강화된 GE-ViT 제시.

* Related Work.

* Distribution Shifts and Evaluation Protocols

- Taxonomy of Distribution Shifts

background, corruption, texture, style shift로 나누어 분석 수행

- Background Shift

auxiliary cue로 간주됨

하지만 배경이 지배적일수 있으며 이를 방지하기 위해서, background shift invariant 해야함.

- Corruption Shift

Image에 포함된 자연적인 불순물

↳ 이미지 촬영, 처리 단계에서 발생

Pixel 수준에서 발생하지만 성능이 현저히 저하될수 있음.

- Texture Shift

색과 intensive는 spacial information을 제공하고 classify에 중요.

- Style Shift.

Texture, shape 등을 포함하는 복잡한 Shift.

- Model Zoo

- Vision Transformer.
- Big Transfer

- Evaluation Protocols

classification은 encoder feature F와 classifier C로 나뉨

D_{train}, D_{iid}, D_{ood}로 dataset이 구성됨

- Accuracy on OOD Data

$$Acc(F, C; \mathscr{D}_{ood}) = \frac{1}{|\mathscr{D}_{ood}|} \sum_{(\mathrm{x}, y) \in \mathscr{D}_{ood}} \mathbb{1}(C(F(\mathrm{x})) = y),$$

- IID / OOD Generalize Gap

IID data와 비교하여 OOD에서 얼마나 잘 작동하는가.

∴ generalize Gap 사용

$$Gap(F, C; \mathscr{D}_{iid}, \mathscr{D}_{ood}) = Acc(F, C; \mathscr{D}_{iid}) - Acc(F, C; \mathscr{D}_{ood}).$$

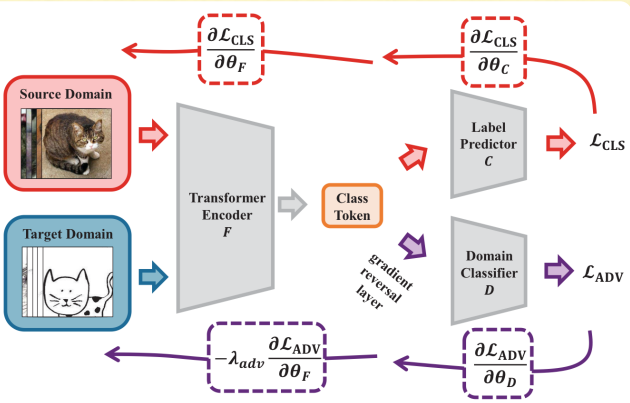
* Generalization - Enhanced ViT.

T-ADV, T-MME, T-SSL을 비교하여, GE-ViT 설계

- Adversarial Learning

Domain - invariant를 위해 domain discriminator 도입.

↳ Adv로 domain - confuse 학습



(a) T-ADV

$$\mathcal{L}_{\text{CLS}} = \sum_{(\text{x},y)\in\mathcal{D}_s} \mathcal{H}(\sigma(C(F(\text{x}))), y),$$

$$\mathcal{L}_{\text{ADV}} = \sum_{(\text{x},y_d)\in\mathcal{D}_s,\mathcal{D}_t} \mathcal{H}(\sigma(D(F(\text{x}))), y_d),$$

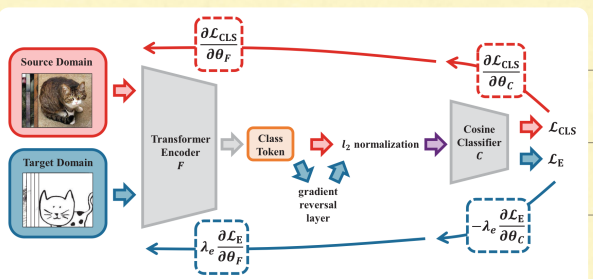
$$(\hat{\theta}_F, \hat{\theta}_C) = \arg\min_{\theta_F,\theta_C} \mathcal{L}_{\text{CLS}} + \lambda_{adv}\mathcal{L}_{\text{ADV}},$$

$$\hat{\theta}_D = \arg\max_{\theta_D} \mathcal{L}_{\text{ADV}},$$

→ Gradient Reversal Layer (GRL)

- Minimax Entropy

Target data의 conditional Entropy에 대한 minmax process를 이용하여, D를 학습하면서 distribution gap을 줄임



(b) T-MME

$$\mathcal{L}_{\text{CLS}} = \sum_{(\text{x},y)\in\mathcal{D}_s} \mathcal{H}(\sigma(C(F(\text{x}))), y),$$

$$\mathcal{L}_{\text{E}} = \sum_{\text{x}\in\mathcal{D}_t} \mathcal{H}(\sigma(C(F(\text{x})))),$$

$$\hat{\theta}_F = \arg\min_{\theta_F} \mathcal{L}_{\text{CLS}} + \lambda_e\mathcal{L}_{\text{E}},$$

$$\hat{\theta}_C = \arg\min_{\theta_C} \mathcal{L}_{\text{CLS}} - \lambda_e\mathcal{L}_{\text{E}},$$

→ cosine - similarity based C를 사용

→ C의 weight에 대해, C는 l_2 norm.

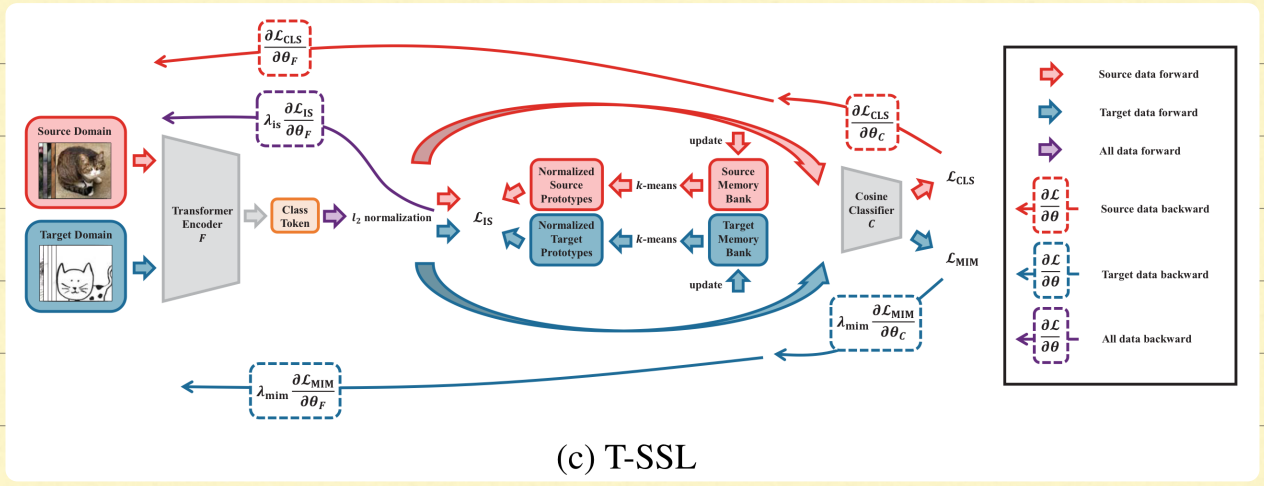
$\frac{F(x)}{\|F(x)\|}$, $\frac{1}{T}\frac{W^T F(x)}{\|F(x)\|}$

unlabeled에 대해 자기와 labeled data와의 거리를 최소화하여 feature 뽑기.

↳ entropy도 높이 하여 source domain 학습에 잘 clustering되게끔

∴ F와 C 사이의 minmax

- Self-supervised learning



(c) T-SSL

$$(\hat{\theta}_F, \hat{\theta}_C) = \arg\min_{\theta_F, \theta_C} \mathcal{L}_{\text{CLS}} + \lambda_{\text{is}}\mathcal{L}_{\text{IS}} + \lambda_{\text{mim}}\mathcal{L}_{\text{MIM}},$$

$$\mathcal{L}_{\text{CLS}} = \sum_{(\text{x},y)\in\mathcal{D}_s} \mathcal{H}(\sigma(C(F(\text{x}))), y).$$

$$\mathcal{L}_{\text{IS}} = \sum_{i=1}^{|\mathcal{D}_s|} \mathcal{H}(P_i^s, c_s(i)) + \sum_{i=1}^{|\mathcal{D}_t|} \mathcal{H}(P_i^t, c_t(i)),$$

$$\mathcal{L}_{\text{MIM}} = \mathbb{E}_{\text{x}}[\mathcal{H}(p(y|\text{x};\theta)] - \mathcal{H}(\mathbb{E}_{\text{x}\in\mathcal{D}_s\cup\mathcal{D}_t}[p(y|\text{x};\theta]).$$

* Systematic Study on ViTs Generalization.

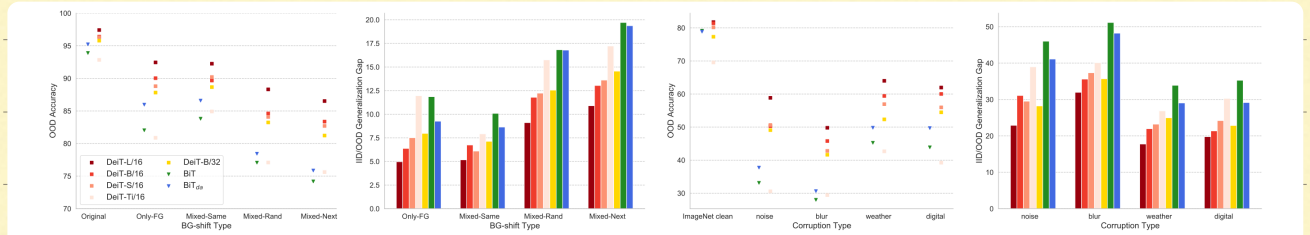
- In-Distribution Generalization.

1) Data 개수 scale이 증가함에 따라 이득 효율 ↓

2) DeiT-S/16이 BiT로 이길 수 있다.

- Background Shifts Generalization Analysis

Background Shifts 에 대한 조사는 ImageNet-9 활용



(a) (b) (c) (d)

Figure 2. **Results on ImageNet-9 and ImageNet-C.** (a)-(b) and (c)-(d) respectively illustrate the OOD Accuracy and IID/OOD Generalization Gap for different models on ImageNet-9 and ImageNet-C datasets. From (a) and (b), we conclude that **1)** ViTs perform with a weaker background-bias than CNNs, **2)** a larger ViT extracts a more background-irrelevant representation. From (c) and (d), we draw the conclusions that **1)** ViTs deal with corruption shifts better than CNNs and generalize better along with model size scaling up, **2)** ViTs do benefit from diverse augmentation in enhancing generalization towards vicinal impurities, but their architectural advantage cannot be overlooked as well, **3)** patch size for training has little influence on ViTs' generalization ability.

- ViTs perform with a weaker background-bias than CNNs

class와 관련이 없는 배경 mixed-same 와 neutral background의 mixed-rand의 정확도 차이로 배경 의존도 계산.

Fig 2. a)

↳ ViT는 CNN보다 낮은 배경 편향.

- A larger ViT extracts a more background-irrelevant representation.

ViT가 큰수록 foreground에 더 집중

- Corruption Shifts Generalization Analysis → Fig 2. c, d

- ViT deal with Corruption Shift better than CNNs

  and Generalize better along with model size scaling up

대부분 ViT가 BiT보다 성능이 좋음

나머지 OOD와 gap 성능

- ViTs benefit from diverse augmentations in enhancing generalization towards vicinal impurities,

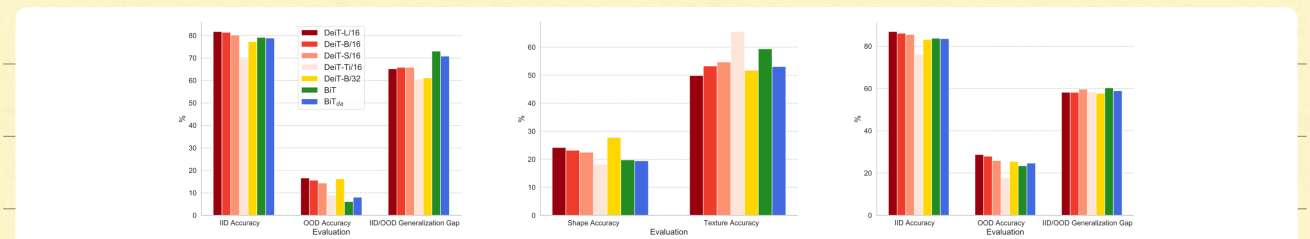
  but their architectural advantage cannot be overlooked.

BiT에서 augmentation은 sensitivity가 줄어들었지만, ViT가 성능이 좋은데, architecture 덕분으로 볼 수 있다.

- Patch size for training has little influence on ViT's generalization ability.

Patch가 크면 일반화가 더 좋지만, model 내에서만 적용됨

- Texture Shifts Generalization Analysis → Fig 3. a, b



(a) Stylized-ImageNet (b) Cue Conflict Stimuli (c) ImageNet-R

Figure 3. **Results on Stylized-ImageNet, Cue Conflict Stimuli and ImageNet-R.** (a), (b) and (c) respectively illustrate the OOD Accuracy and IID/OOD Generalization Gap for different models on Stylized-ImageNet, Cue Conflict Stimuli, and ImageNet-R data sets. From (a) and (b) we could draw the following conclusions that **1)** ViTs' stronger bias towards shape enables them to generalize better under texture shifts and their shape biases have a positive correlation with their sizes, **3)** ViTs with larger patch size exhibit a stronger bias towards the shape. From (c) we observe that most ViTs beat BiTs in OOD accuracy while having little difference in the IID/OOD generalization gap.

- ViTs' Stronger bias towards shape enables them to generalize better under texture shifts

  and their shape biases have a positive correlation with their sizes

ViT는 shape에 강한 편향 때문에 texture는 일반화 잘됨

나쁜부정인 영향은 덜 받음

- ViTs with larger patch size exhibit a stronger bias towards the shape.

큰 patch 단위로 global하게 집중

- Style Shifts Generalization Analysis → Fig 3. c.

- ViTs have a diverse performance on IID/OOD generalization gap under Style Shift.

(a) DeiT-B/16 (b) DeiT-S/16 (c) BiT (d) BiT$_{da}$

Figure 4. **Results on DomainNet.** From the results, we can conclude that **1)** DeiT-S/16 performs better on the small-scale datasets in IID conditions. Thus, the model easily outperforms BiTs in OOD accuracy, **2)** when inspecting the IID/OOD generalization gap, the results differ a lot. When models are trained on clipart and painting, there is no obvious difference of gap between DeiT-S/16 and BiTs.

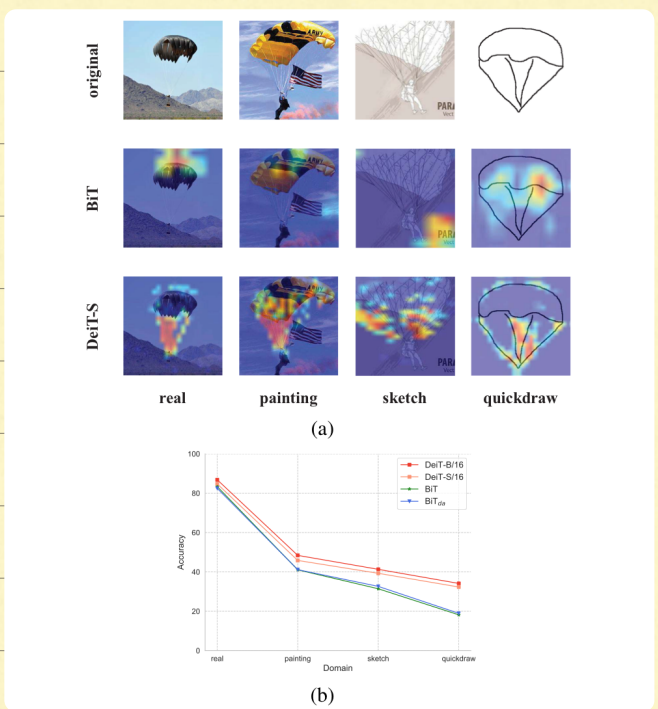
대부분의 ViT의 OOD가 BiT의 OOD를 능가.

IID/OOD의 gap은 비슷

∴ 예술 작품과 같이 task에선 ViT가 큰 효과는 없음

그리고, ViT의 경우 pre-trained 를 더 잘 활용하고,

downstream에 좋다.



- ViTs shows stronger bias towards object structure.

Structure만 남은 degrade image에 대해서 강함.

ViT는 주요 구조 정보만 남아도 잘 catch하여, BiT는 실패.

ViT는 구조에 더 집중

- ViTs will eliminate different levels of DS in different layers.

ViT는 여러 layer에서 DS 처리

* Studies on Generalization-Enhanced ViTs

- Setting

DomainNet 사용

BiT와 그에의 ViT에 대해서도 실험.

- Performance Analysis

1) GE-ViT는 4% 향상

2) 3개의 GE-ViT 모두 그렇.

CNN은 self-supervised의 이점.

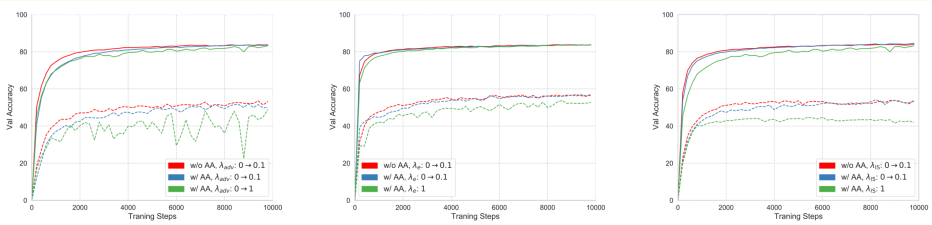
CNN과 ViT 모두 장점이 있지만 보통

ViT가 더 큰 이점

Table 2. **Results of Generalization-enhanced methods.** Specifically, we compare three types of GE-ViTs with their corresponding CNNs. From the results we could conclude that 1) equipped with GE-ViTs, we achieve significant performance boosts towards out-of-distribution data by 4% from vanilla ViTs. 2) three GE-ViTs have almost the same improvement from vanilla models on OOD accuracy. 3) for the enhanced transformer models, larger ViTs still benefit more for the out-of-distribution generalization.

Model	Method	R to C	R to P	P to C	C to S	S to P	R to S	P to R	Avg.
DeiT-B/16	-	54.64	48.40	40.37	45.69	36.75	41.31	55.33	46.07
	T-ADV	58.19	50.85	41.91	51.18	46.12	47.47	55.65	50.20
	T-MME	**60.59**	**51.98**	42.30	50.32	45.79	**47.92**	54.87	50.54
	T-SSL	56.80	49.06	**45.96**	**51.79**	**46.95**	45.95	**60.98**	**51.07**
DeiT-S/16	-	50.60	45.82	36.09	43.39	35.24	39.29	52.08	43.22
	T-ADV	53.60	47.84	37.99	47.10	41.61	41.94	52.82	46.13
	T-MME	**56.86**	**49.15**	38.97	46.48	42.95	**42.07**	52.49	47.00
	T-SSL	53.86	46.71	**42.79**	**47.25**	**43.01**	40.94	**57.07**	**47.37**
BiT	-	42.18	41.14	30.72	37.01	28.23	32.64	48.54	36.78
	DANN [10]	45.20	42.86	32.96	40.44	36.63	35.26	49.25	40.37
	MME [24]	50.21	**44.61**	34.75	40.27	38.41	37.83	47.58	41.95
	SSL [34]	**52.55**	42.80	**39.03**	**45.72**	**39.08**	**39.65**	**56.07**	**44.98**
VGG-16	-	39.39	37.32	26.36	32.96	25.55	27.79	45.70	33.58
	DANN [10]	43.26	40.09	28.68	36.22	31.63	**35.45**	44.73	37.15
	MME [24]	42.65	**42.46**	27.41	**36.93**	33.94	32.58	**45.87**	37.41
	SSL [34]	**43.79**	41.88	**32.19**	35.73	**36.99**	31.05	55.18	**39.54**

- Smooth Feature Alignment.



(a) T-ADV (b) T-MME (c) T-SSL

Figure 7. **Investigation of Generalization-enhanced methods with different training strategies.** (a)-(c) show training curves on both source domain and target domain. From the results, we can conclude that classical training strategies (the green lines) on CNNs are not suitable for ViTs, which need smoother strategies (the red lines) to align features in both domains.