Big Transfer (BiT): General Visual Representation Learning (2020)

Alexander Kolesnikov*, Lucas Beyer*, Xiaohua Zhai*, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Housby
Google Research, Brain Team

IIE8557-01 동적계획법과 강화학습

경영과학연구실 김윤석

- 특정 작업에 심층 학습을 사용하기 위해서는 많은 양의 데이터와 연산을 필요로 함
- 이러한 요구 사항은 심층 학습의 비용을 매우 비싸게 만들 수 있음
- Transfer learning은 비용 문제를 해결하기 위한 방법으로 제안 됨

BigTransfer는 새로운 구성 요소나 복잡성을 도입하지 않고 특정 작업에 비용이 발생하지 않는 심층 신경망 훈련에 대해 연구함

- Big Transfer는 큰 training 소스 데이터에서 사전훈련을 수행하고 특정 작업에 대해 fine tuning을 함
- BiT-HyperRule은 Downstream task에 대해 미세 조정을 할 때 중요한 하이퍼파라미터를 결정하기 위한 휴리스틱 규칙임

BiT-HyperRule의 주요 특징

- 고정된 하이퍼파라미터: 대부분의 하이퍼파라미터는 모든 데이터셋에 걸쳐 고정됨
- 데이터 전처리: 모든 작업에 대해 랜덤 크롭과 수평 뒤집기를 적용함 (단, 크롭이나 뒤집기가 레이블의 의미를 손상시키는 경우 해당 전처리를 생략함)
- 학습률 조정: 미세 조정 중에는 학습률을 여러 시점에서 감소시킴
- MixUp: 중간 및 큰 작업에 대해 특정 α 값으로 MixUp을 적용함

- BiT은 기본적으로 ResNet-v2 모델을 사용함
- BiT의 사이즈는 Pre-train에 사용된 dataset의 크기로 구분됨
- 학습은 SGD와 모멘텀을 사용함
- 해상도는 224x224로 전처리 되며, 예외적으로 128x128을 사용하는 경우와 384x384로 사용하는 경우가 있음

BiT-Size

- BiT-S: ILSVRC-2012(ImageNet-1k)로 학습한 모델로 약 1.3M개의 이미지를 사용함
- BiT-M: ImageNet-21k로 학습한 모델로 약 14M개의 이미지를 사용함
- BiT-L: JFT-300M으로 학습한 모델로 약 300M개의 이미지를 사용함

- 실험은 BiT-S, M, L을 많은 downstream 작업에서 평가함
- BiT은 데이터의 양이 많거나 적은 경우 모두 뛰어난 성능을 보임

실험 종류

- Standard Computer Vision Benchmarks: ILSVRC-2012, CIFAR-10, CIFAR-100, Pets, Flowers, VTAB등 성능 평가를 많이 하는 dataset에 대해 BiT의 성능을 평가함
- Tasks with Few Datapoints: 클래스당 데이터를 1, 5, 10, 25, 100개 등으로 fine tuning하며 성능을 평가함
- ObjectNet(Recognition on a “Real-World” Test set): ObjecNet 데이터셋에서 BiT을 평가함
- Object Detection: Object detection task에서 BiT을 평가함

Standard Computer Vision Benchmarks

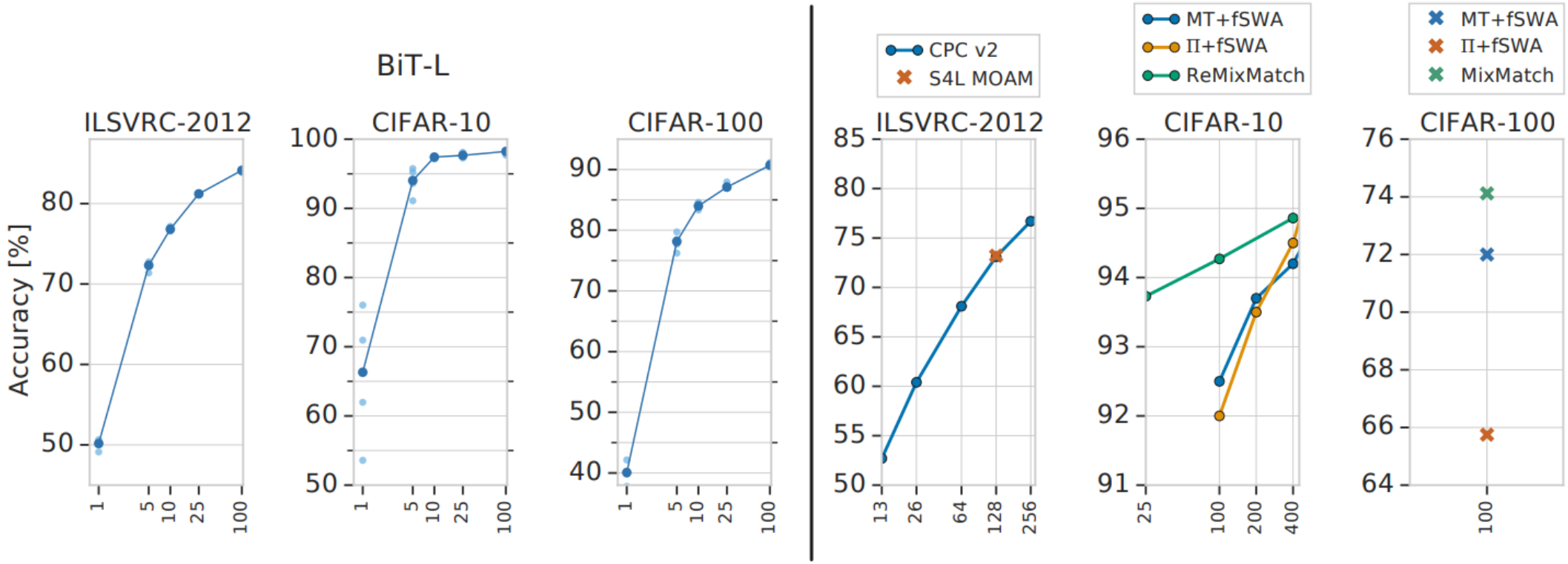
- BiT-L은 SOTA보다 일반적인 Generalist SOTA보다 더 좋은 성능을 보이며, Specialist SOTA보다 더 좋은 경우도 보임
- BiT-S와 BiT-M의 결과를 통해 BiT-M이 더 향상된 시각적 표현을 가짐을 보여줌

	BiT-L	Generalist SOTA	Specialist SOTA
ILSVRC-2012	**87.54 ± 0.02**	86.4 [57]	88.4 [61]*
CIFAR-10	**99.37 ± 0.06**	99.0 [19]	-
CIFAR-100	**93.51 ± 0.08**	91.7 [55]	-
Pets	**96.62 ± 0.23**	95.9 [19]	97.1 [38]
Flowers	**99.63 ± 0.03**	98.8 [55]	97.7 [38]
VTAB (19 tasks)	**76.29 ± 1.70**	70.5 [58]	-

	ILSVRC-2012	CIFAR-10	CIFAR-100	Pets	Flowers	VTAB-1k (19 tasks)
BiT-S (ILSVRC-2012)	81.30	97.51	86.21	93.97	89.89	66.87
BiT-M (ImageNet-21k)	85.39	98.91	92.17	94.46	99.30	70.64
Improvement	+4.09	+1.40	+5.96	+0.49	+9.41	+3.77

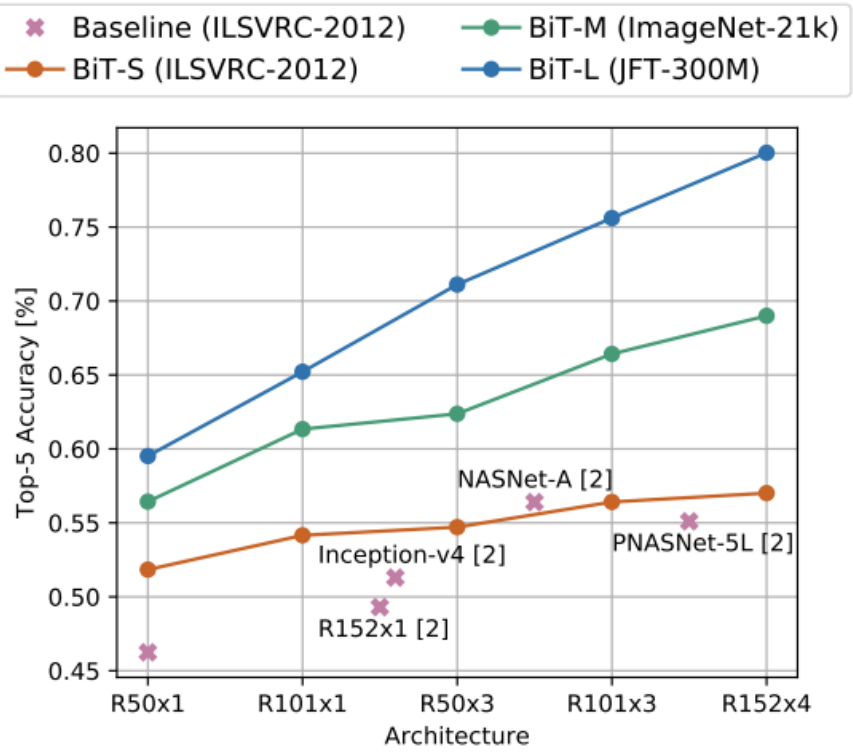
Tasks with Few Datapoints

- BiT-L의 transfer learning을 위해 필요한 샘플 수에 대해 평가한 실험임
- BiT-L은 클래스당 매우 적은 샘플로도 강력한 성능을 보임



ObjectNet: Recognition on a "Real-World" Test Set

- ObjectNet 데이터셋은 실생활 시나리오와 매우 유사하게 객체 카테고리가 비정형적인 맥락, 시점, 회전 등에서 나타날 수 있음
- 더 큰 구조와 더 많은 데이터에 대한 사전 훈련이 풍부한 시각 표현을 통해 좋은 성능을 보임



Object Detection

- 객체 검출에서 BiT을 평가한 실험임
- 본 실험은 COCO-2017 데이터셋을 사용하고 사전훈련된 BiT 모델을 백본으로 사용하여 실험을 진행함

Model	Upstream data	AP
RetinaNet [33]	ILSVRC-2012	40.8
RetinaNet (BiT-S)	ILSVRC-2012	41.7
RetinaNet (BiT-M)	ImageNet-21k	43.2
RetinaNet (BiT-L)	JFT-300M	**43.8**

- BiT은 심층 학습에서 특정 작업에서 발생하는 비용을 줄이기 위한 Transfer learning에 대해 연구함
- BiT은 더 크고 다양한 데이터셋을 활용한 사전 학습이 다양한 시각적 표현을 통해 여러 작업에서 좋은 정확도를 갖는 것을 보임
- 저자는 BiT을 활용한 downstream task에서 hyperparmater를 효율적으로 tuning할 수 있도록 BiT-HyperRule을 제시함

Q & A