캡스톤 디자인

‘딥페이크 탐지’

#15. random 모델 생성 및 다른 네트워크 perturbation 적용

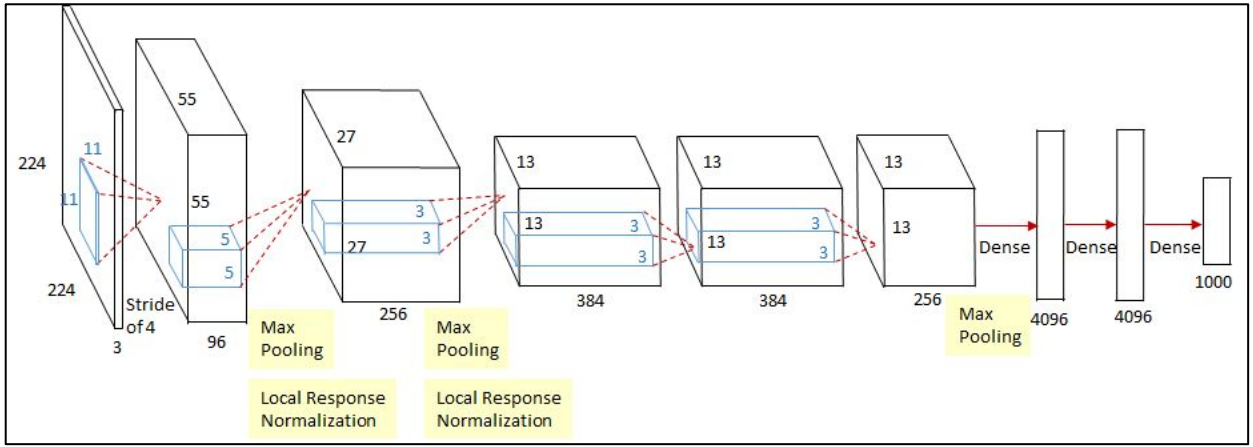
김지수, 김민지, 민지민

지난 캡스톤 회의 내용

- 다른 네트워크 perturbation 을 적용한 이미지 생성 후 성능 확인
- random model 생성 후 inference

CaffeNet 조사

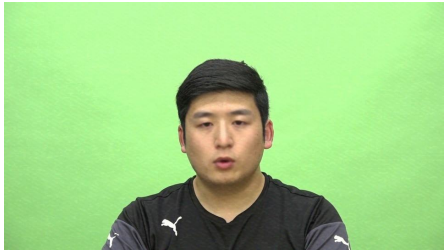
- ILSVR2012 우승한 AlexNet의 1-GPU 버전
- AlexNet과 pooling & normalization layer 순서 다름



perturbation 추론 - caffeNet

생성 이미지

Level1



Level4

Level7

Level10

(기본 모델)추론 결과

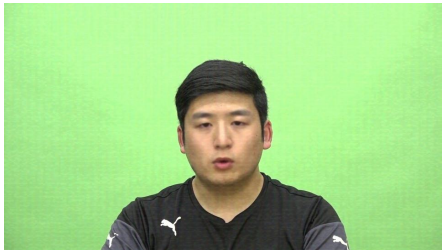
basic -> caffe 추론

Level	Loss	Accuracy
1	0.0045	1
2	0.0308	0.998
3	0.1506	0.988
4	0.3234	0.878
5	0.4269	0.772
6	0.535	0.684
7	0.6238	0.63
8	0.6567	0.594
9	0.6652	0.58
10	0.6704	0.548

perturbation 추론 - ResNet

생성 이미지

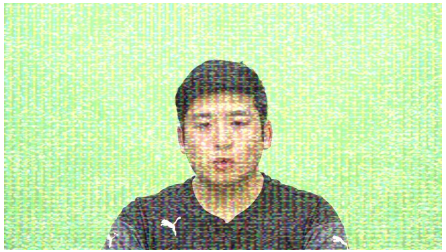
Level1



Level4

Level7

Level10



(기본 모델)추론 결과

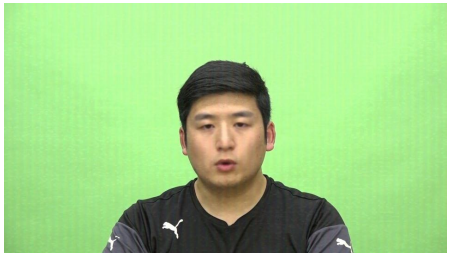
basic -> resnet 추론

Level	Loss	Accuracy
1	0.0046	0.999
2	0.0323	0.998
3	0.1369	0.987
4	0.3127	0.951
5	0.465	0.781
6	0.5818	0.657
7	0.6478	0.595
8	0.668	0.576
9	0.6754	0.543
10	0.6787	0.544

perturbation 추론 - VGG16

생성 이미지

Level1



Level4

Level7

Level10

(기본 모델)추론 결과

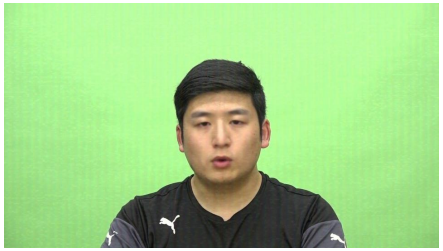
basic -> vgg16 추론

Level	Loss	Accuracy
1	0.0069	0.997
2	0.0258	0.995
3	0.0658	0.994
4	0.1614	0.989
5	0.3093	0.933
6	0.4457	0.799
7	0.5265	0.681
8	0.5819	0.637
9	0.6187	0.613
10	0.6434	0.599

perturbation 추론 - VGG19

생성 이미지

Level1



Level4

Level7

Level10

(기본 모델)추론 결과

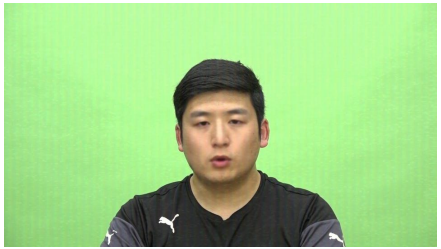
basic -> vgg19 추론

Level	Loss	Accuracy
1	0.004	0.999
2	0.0164	0.997
3	0.0545	0.993
4	0.1497	0.993
5	0.2872	0.967
6	0.4238	0.89
7	0.5213	0.724
8	0.5881	0.638
9	0.6329	0.582
10	0.6578	0.567

perturbation 추론 - VGGf

생성 이미지

Level1



Level4

Level7

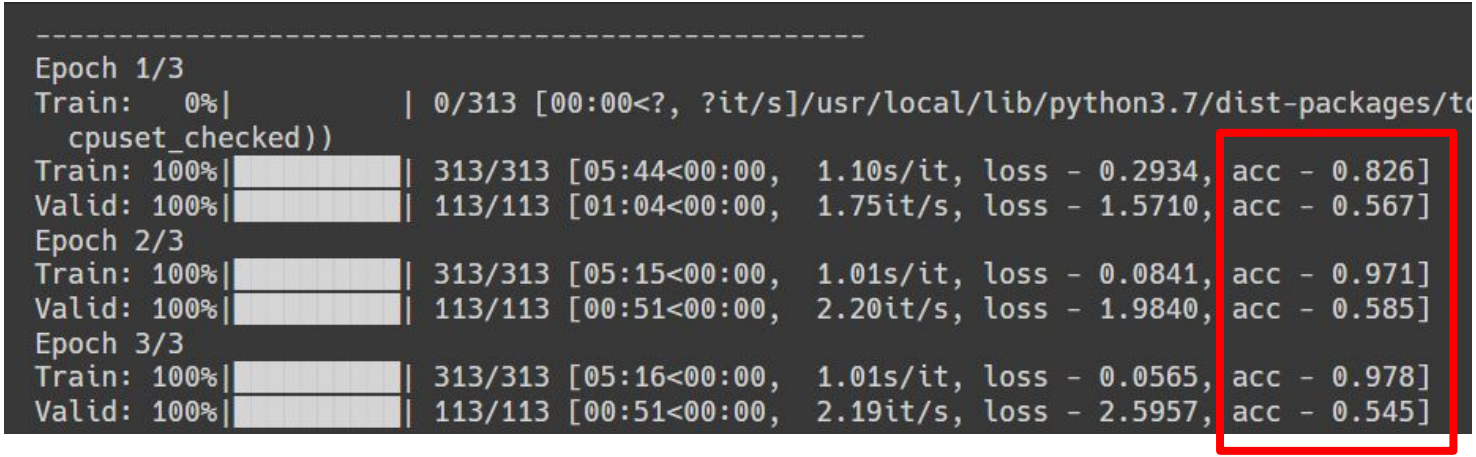
Level10

(기본 모델)추론 결과

basic -> vggf 추론

Level	Loss	Accuracy
1	0.0036	1
2	0.0162	0.997
3	0.0575	0.996
4	0.1742	0.988
5	0.3215	0.928
6	0.4292	0.802
7	0.4991	0.754
8	0.56	0.715
9	0.6067	0.663
10	0.6365	0.631

blur를 반영한 random model



대부분 **fake**로 판별하는 경향

학습에 반영한 기법 :

Gaussian blur, Gaussian noise, Sharpening, Universal Perturbation(google net), Jpeg

blur 제외한 random 모델

```
----------------------------------------------------
Epoch 1/3
Train:   0%|          | 0/313 [00:00<?, ?it/s]/usr/local/lib/python3.7/dist-packages/
  cpuset_checked))
Train: 100%|██████████| 313/313 [05:44<00:00,  1.10s/it, loss - 0.3029, acc - 0.821]
Valid: 100%|██████████| 113/113 [00:55<00:00,  2.02it/s, loss - 1.1036, acc - 0.789]
Epoch 2/3
Train: 100%|██████████| 313/313 [05:49<00:00,  1.12s/it, loss - 0.0668, acc - 0.977]
Valid: 100%|██████████| 113/113 [00:56<00:00,  2.00it/s, loss - 0.2215, acc - 0.902]
Epoch 3/3
Train: 100%|██████████| 313/313 [05:50<00:00,  1.12s/it, loss - 0.0307, acc - 0.989]
Valid: 100%|██████████| 113/113 [00:57<00:00,  1.95it/s, loss - 0.2183, acc - 0.916]
```

학습에 반영한 기법 :

Gaussian noise, Sharpening, Universal Perturbation(google net), Jpeg

random model -> unsharp 데이터셋에 대한 추론 성능

random -> unsharp 추론		
Level	**Loss**	**Accuracy**
1	0.0899	0.967
2	0.6052	0.907
3	0.8274	0.871
4	1.0762	0.819
5	0.3009	0.777
6	1.4899	0.753
7	1.624	0.739
8	1.7225	0.726
9	1.79	0.722
10	1.8256	0.718

impulse noise - 원본 데이터셋만 학습한 모델로 추론

	loss	acc
level 1 (1)	**11.2096**	**0.5**
level 2 (3)	**9.7332**	**0.5**
level 3 (5)	**7.2340**	**0.5**
level 4 (6)	**5.8425**	**0.5**
level 5 (7)	**4.6297**	**0.5**
level 6 (9)	**3.1389**	**0.5**
level 7 (11)	**2.5344**	**0.5**
level 8 (13)	**2.3761**	**0.5**
level 9 (15)	**2.3796**	**0.5**
level 10 (17)	**2.4415**	**0.5**

impulse noise - random 모델로 추론

	loss	acc
level 1 (1)	**0.1026**	**0.959**
level 2 (3)	**0.2270**	**0.917**
level 3 (5)	**0.4765**	**0.847**
level 4 (6)	**0.5636**	**0.809**
level 5 (7)	**0.6685**	**0.763**
level 6 (9)	**0.8769**	**0.706**
level 7 (11)	**1.1300**	**0.629**
level 8 (13)	**1.2799**	**0.597**
level 9 (15)	**1.4561**	**0.566**
level 10 (17)	**1.5629**	**0.552**

poisson noise - 원본 데이터셋만 학습한 모델로 추론

	loss	acc
level 1	6.9061	**0.5**
level 2	7.2971	**0.5**
level 3	7.0077	**0.5**
level 4	6.5752	**0.5**
level 5	6.0992	**0.5**
level 6	5.5676	**0.5**
level 7	5.0500	**0.5**
level 8	4.5922	**0.5**
level 9	4.1349	**0.5**
level 10	3.7825	**0.5**

poisson noise - random 모델로 추론

	loss	acc
level 1	0.0850	0.967
level 2	0.0707	0.973
level 3	0.0690	0.975
level 4	0.0737	0.977
level 5	0.0804	0.975
level 6	0.0982	0.962
level 7	0.1182	0.952
level 8	0.1333	0.946
level 9	0.1528	0.931
level 10	0.1978	0.912