Unsupervised Representation Learning by Predicting Image Rotation

Yongwon Jo

School of Industrial and Management Engineering, Korea University

KOREA UNIVERSITY DMQA

Contents

- ❖ Research Purpose
- ❖ RotNet
- ❖ Experiments
- ❖ Conclusion

Research Purpose

❖ **Unsupervised Representation Learning by Predicting Image Rotation**

- 2018년 International Conference on Learning Representations 에서 발표된 논문
- 2022년 5월 6일 기준 1660회 인용
- 논문 내 제안 방법론은 Self-supervised Learning 방법론 중 Pretext task 방식에 속하는 방법론
- 'The simple is, the best'

Published as a conference paper at ICLR 2018

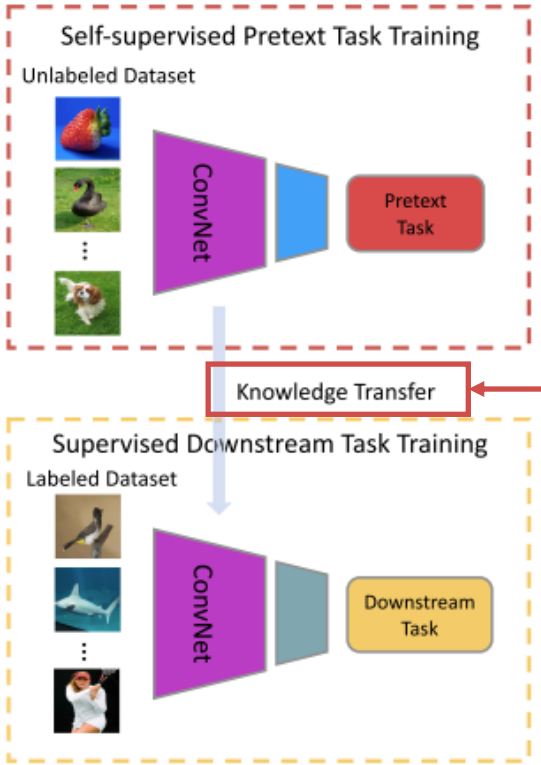
UNSUPERVISED REPRESENTATION LEARNING BY PREDICTING IMAGE ROTATIONS

Spyros Gidaris, Praveer Singh, Nikos Komodakis
University Paris-Est, LIGM
Ecole des Ponts ParisTech
{spyros.gidaris,praveer.singh,nikos.komodakis}@enpc.fr

Research Purpose

❖ Self-supervised Representation Learning(SSL) 필요성

- Computer vision 최신 방법론 학습을 위해서는 방대한 (입력데이터-출력데이터) 쌍 필요
- 하지만 (입력-출력) 쌍으로 구성된 데이터셋을 구축하기 위해선 시간 및 비용 필요
- 데이터 셋을 구축하기보다 새로운 레이블을 정의해서 학습하자는 것이 SSL의 목표



- 입력 데이터만 사용(Unlabeled)
- 새롭게 정의한 출력 데이터(Self-supervision)
- 학습된 모델 가중치 전이
- 소수의 (입력-출력) 데이터 사용(Labeled)
- 원래 해결하고 했던 문제(Downstream task)

Jing, L, & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence, 43(11), 4037-4058.

Research Purpose

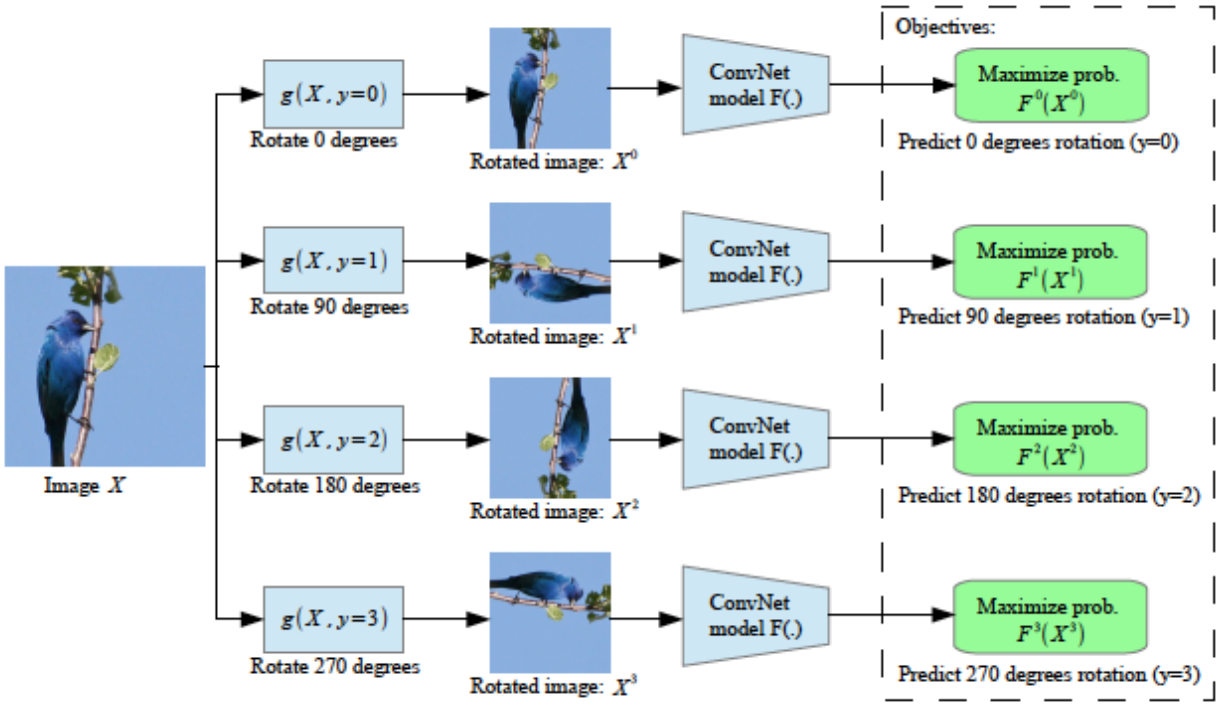
❖ **Geometric transformation(Rotation)을 사용한 Self-supervision 정의**

- 이미지 분류 문제에 대한 Self-supervision 정의 방식을 제안(RotNet)
- 입력 이미지를 90도, 180도, 270도 회전시키고 회전 시킨 각도를 Self-supervision으로 정의
- 간단하지만 강력한 성능을 보임을 실험적으로 증명

RotNet

❖ Geometric transformation (Image Rotation)

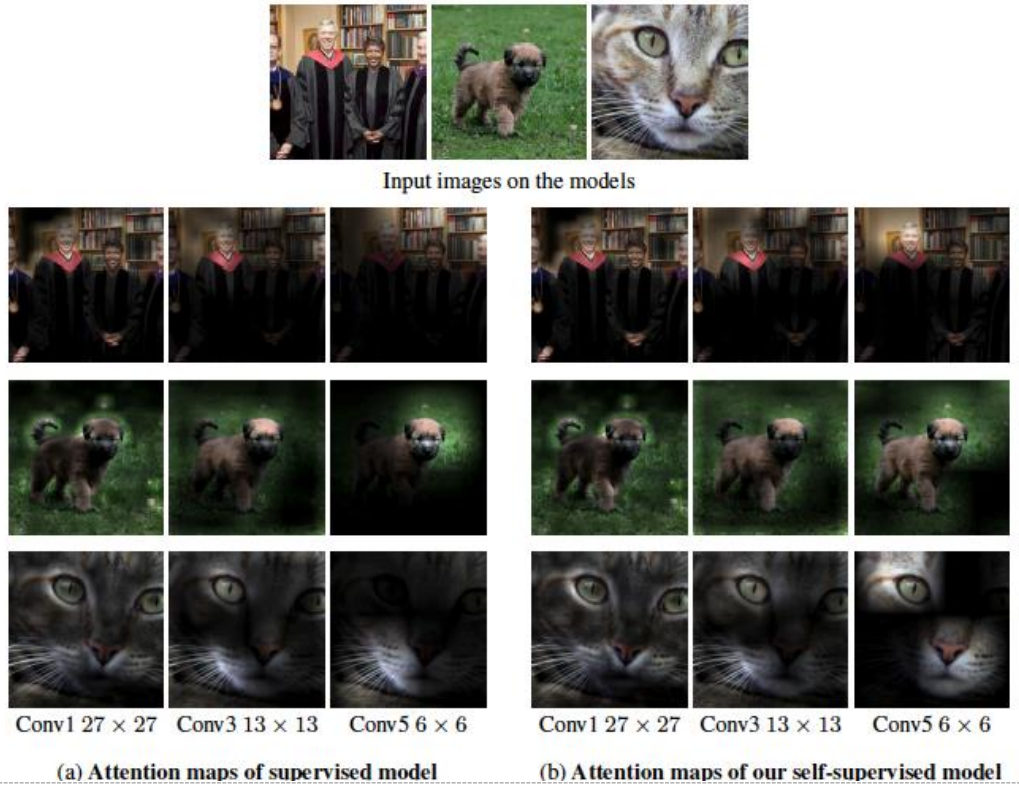
- 입력 이미지를 90도, 180도, 270도를 회전시키고 회전 각도를 출력 변수로 지정
- 회전한 이미지에 Random Cropping을 진행 후 모델에 입력
- 단일 이미지와 변환된 세 장 이미지를 동시에 입력한다는 것이 특징
  - ➢ Batch size가 16일 경우 모델 학습 시 Batch size는 64(=16*4)로 연산이 진행됨



RotNet

❖ **Attention map visualization (Supervised VS Self-supervised)**

- 지도학습 방식으로 학습된 모델과 RotNet 방식으로 학습된 모델의 Attention map 비교
- 아래 그림 내 Conv1, Conv3 등은 어떠한 Layer 이후의 Attention map인지를 의미
- 지도 학습 방식으로 학습된 모델과 RotNet 방식으로 학습된 모델의 Attention map이 유사함



Experiments

❖ **RotNet과 Convolution block 개수 차이에 따른 성능 비교(CIFAR-10)**

- Convolution block으로 구성된 Encoder를 사용하고 k 번째 block에 대한 연산 후 특징 지도 추출
- 특징 지도를 Flatten 하여 3개 Linear layer 통과 후 클래스별 logit 산출
- 특징 지도를 산출하는 Encoder 부분과 Linear layer 학습 (Downstream task)

Table 1: Evaluation of the unsupervised learned features by measuring the classification accuracy that they achieve when we train a non-linear object classifier on top of them. The reported results are from CIFAR-10. The size of the ConvB1 feature maps is $96 \times 16 \times 16$ and the size of the rest feature maps is $192 \times 8 \times 8$.

Model	ConvB1	ConvB2	ConvB3	ConvB4	ConvB5
RotNet with 3 conv. blocks	85.45	88.26	62.09	-	-
RotNet with 4 conv. blocks	85.07	89.06	86.21	61.73	-
RotNet with 5 conv. blocks	85.04	**89.76**	86.82	74.50	50.37

Experiments

❖ **RotNet 학습 시 각도 개수에 따른 성능 비교(CIFAR-10)**

- RotNet 학습 시 앞에서는 0도, 90도, 180도, 270도만 출력 변수로 사용
- Ablation study로 각도를 다양하게 변화시키며 실험
- 각도를 다양하게 함에도 불구하고 큰 성능 변화는 없는 것으로 확인

Table 2: Exploring the quality of the self-supervised learned features w.r.t. the number of recognized rotations. For all the entries we trained a non-linear classifier with 3 fully connected layers (similar to Table 1) on top of the feature maps generated by the 2nd conv. block of a RotNet model with 4 conv. blocks in total. The reported results are from CIFAR-10.

# Rotations	Rotations	CIFAR-10 Classification Accuracy
4	0°, 90°, 180°, 270°	**89.06**
8	0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°	88.51
2	0°, 180°	87.46
2	90°, 270°	85.52

Experiments

❖ RotNet 의 Classifier 구조 변화에 따른 성능 비교(CIFAR-10)

- (None-linear) 3개 Linear layers를 가진 Classifier이며 (Conv) 은 Conv+Non-linear
- 서로 다른 구조를 가진 모델들과 사전학습여부에 따른 성능 비교
- 또한, 과거 Self-supervised representation learning 방법론과 비교

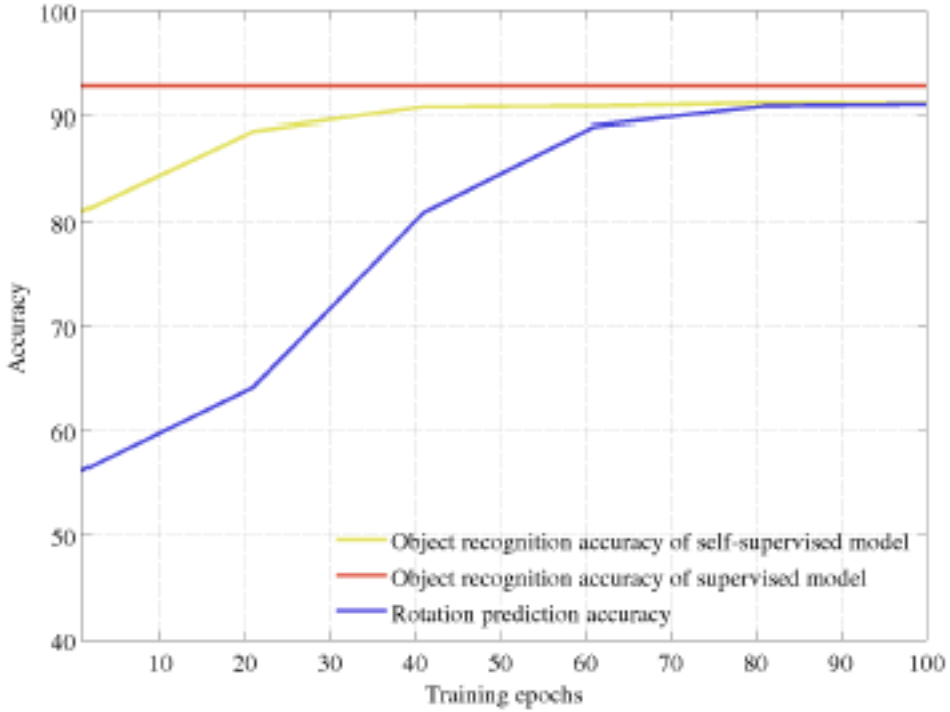
Table 3: Evaluation of unsupervised feature learning methods on CIFAR-10. The *Supervised NIN* and the *(Ours) RotNet + conv* entries have exactly the same architecture but the first was trained fully supervised while on the second the first 2 conv. blocks were trained unsupervised with our rotation prediction task and the 3rd block only was trained in a supervised manner. In the *Random Init. + conv* entry a conv. classifier (similar to that of *(Ours) RotNet + conv*) is trained on top of two NIN conv. blocks that are randomly initialized and stay frozen. Note that each of the prior approaches has a different ConvNet architecture and thus the comparison with them is just indicative.

Method	Accuracy
Supervised NIN	92.80
Random Init. + conv	72.50
(Ours) RotNet + non-linear	89.06
(Ours) RotNet + conv	**91.16**
(Ours) RotNet + non-linear (fine-tuned)	91.73
(Ours) RotNet + conv (fine-tuned)	92.17
Roto-Scat + SVM Oyallon & Mallat (2015)	82.3
ExemplarCNN Dosovitskiy et al. (2014)	84.3
DCGAN Radford et al. (2015)	82.8
Scattering Oyallon et al. (2017)	84.7

Experiments

❖ 학습 과정에 따른 성능 비교(CIFAR-10)

- 특정 Epoch 종료 후 성능을 리포트
- Fully supervised **VS** Self-supervised(Full fine-tune) **VS** Self-supervised(Only classifier)
- Full fine-tuning 시 빠르게 Fully-supervised 와 유사한 성능에 도달하는 것을 확인 가능

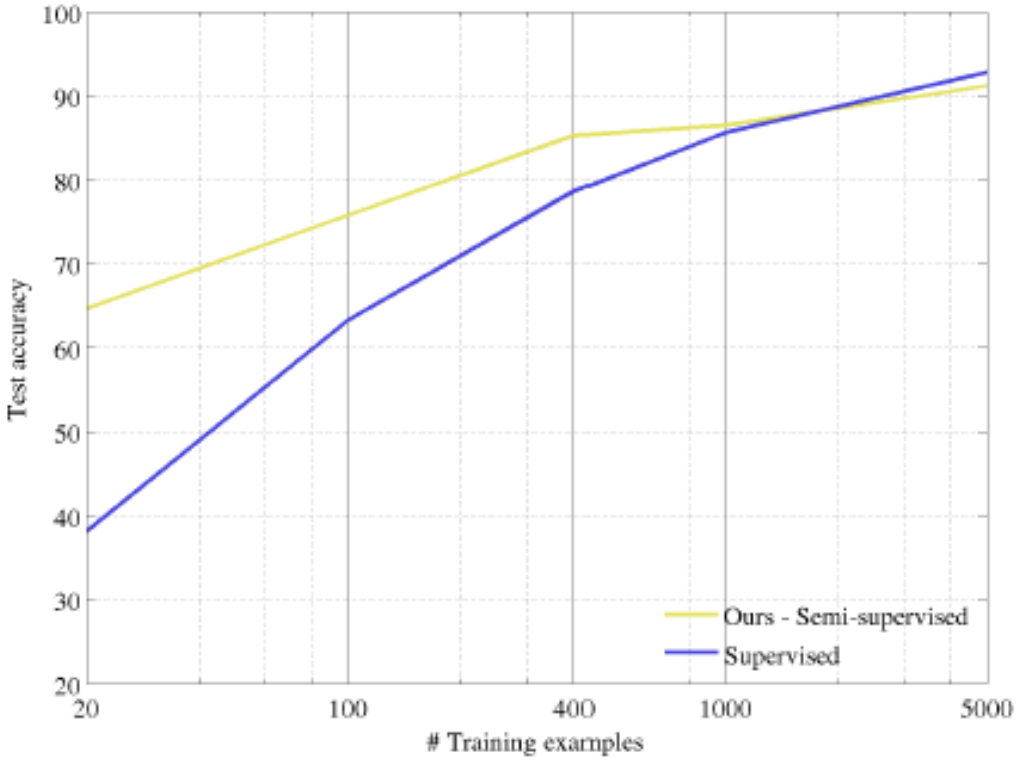


(a)

Experiments

❖ **Semi-supervised learning 형태로 모델 학습(CIFAR-10)**

- 범주별 학습 데이터 수를 20, 100, 400, 1000, 5000개 사용하여 모델 학습 진행
- 학습 데이터 수가 작을 때 RotNet 형태 모델 학습이 성능 향상에 도움을 주는 것을 확인



(b)

Experiments

❖ **다른 Self-supervised learning & Fully-supervised learning과 비교(ImageNet)**

- 아래 수치들은 ImageNet dataset 중 Test dataset에 대한 Top 1 정확도
- **Conv 4, Conv 5 에서 나온 특징 지도를 Flatten하여 Non-linear layer 를 사용해 Logit 산출**
- Fully-supervised learning 모델 성능과의 Margin을 매우 줄인 것을 확인 가능

Table 4: **Task Generalization: ImageNet top-1 classification with non-linear layers**. We compare our unsupervised feature learning approach with other unsupervised approaches by training non-linear classifiers on top of the feature maps of each layer to perform the 1000-way ImageNet classification task, as proposed by Noroozi & Favaro (2016). For instance, for the conv5 feature map we train the layers that follow the conv5 layer in the AlexNet architecture (i.e., fc6, fc7, and fc8). Similarly for the conv4 feature maps. We implemented those non-linear classifiers with batch normalization units after each linear layer (fully connected or convolutional) and without employing drop out units. All approaches use AlexNet variants and were pre-trained on ImageNet without labels except the ImageNet labels and Random entries. During testing we use a single crop and do not perform flipping augmentation. We report top-1 classification accuracy.

Method	Conv4	Conv5
ImageNet labels from (Bojanowski & Joulin, 2017)	59.7	59.7
Random from (Noroozi & Favaro, 2016)	27.1	12.0
Tracking Wang & Gupta (2015)	38.8	29.8
Context (Doersch et al., 2015)	45.6	30.4
Colorization (Zhang et al., 2016a)	40.7	35.2
Jigsaw Puzzles (Noroozi & Favaro, 2016)	45.3	34.6
BIGAN (Donahue et al., 2016)	41.9	32.2
NAT (Bojanowski & Joulin, 2017)	-	36.0
(Ours) RotNet	50.0	43.8

Experiments

❖ **다른 Self-supervised learning & Fully-supervised learning과 비교(ImageNet)**

- 아래 수치들은 ImageNet dataset 중 Test dataset에 대한 Top 1 정확도
- **특징 지도를 Flatten하여 Logistic regression 학습 후 성능 도출**
- Fully-supervised learning 모델 성능과의 Margin을 매우 줄인 것을 확인 가능

Table 5: **Task Generalization: ImageNet top-1 classification with linear layers.** We compare our unsupervised feature learning approach with other unsupervised approaches by training logistic regression classifiers on top of the feature maps of each layer to perform the 1000-way ImageNet classification task, as proposed by Zhang et al. (2016a). All weights are frozen and feature maps are spatially resized (with adaptive max pooling) so as to have around 9000 elements. All approaches use AlexNet variants and were pre-trained on ImageNet without labels except the ImageNet labels and Random entries.

Method	Conv1	Conv2	Conv3	Conv4	Conv5
ImageNet labels	19.3	36.3	44.2	48.3	50.5
Random	11.6	17.1	16.9	16.3	14.1
Random rescaled Krähenbühl et al. (2015)	17.5	23.0	24.5	23.2	20.6
Context (Doersch et al., 2015)	16.2	23.3	30.2	31.7	29.6
Context Encoders (Pathak et al., 2016b)	14.1	20.7	21.0	19.8	15.5
Colorization (Zhang et al., 2016a)	12.5	24.5	30.4	31.5	30.3
Jigsaw Puzzles (Noroozi & Favaro, 2016)	18.2	28.8	34.0	33.9	27.1
BIGAN (Donahue et al., 2016)	17.7	24.5	31.0	29.9	28.0
Split-Brain (Zhang et al., 2016b)	17.7	29.3	35.4	35.2	32.8
Counting (Noroozi et al., 2017)	18.0	30.6	34.3	32.5	25.7
(Ours) RotNet	**18.8**	**31.7**	**38.7**	**38.2**	**36.5**

Experiments

❖ **다른 Self-supervised learning & Fully-supervised learning과 비교(Pascal VOC 2007)**

- Pascal VOC 20**는 Computer vision 중 객체 인식 관련 대표적인 벤치마크
- Image classification이외 Computer vision 문제에 대해서는 성능 개선이 필요한 것으로 보임

	Classification (%mAP)		Detection (%mAP)	Segmentation (%mIoU)
Trained layers	fc6-8	all	all	all
ImageNet labels	78.9	79.9	56.8	48.0
Random		53.3	43.4	19.8
Random rescaled Krähenbühl et al. (2015)	39.2	56.6	45.6	32.6
Egomotion (Agrawal et al., 2015)	31.0	54.2	43.9	
Context Encoders (Pathak et al., 2016b)	34.6	56.5	44.5	29.7
Tracking (Wang & Gupta, 2015)	55.6	63.1	47.4	
Context (Doersch et al., 2015)	55.1	65.3	51.1	
Colorization (Zhang et al., 2016a)	61.5	65.6	46.9	35.6
BIGAN (Donahue et al., 2016)	52.3	60.1	46.9	34.9
Jigsaw Puzzles (Noroozi & Favaro, 2016)	-	67.6	53.2	37.6
NAT (Bojanowski & Joulin, 2017)	56.7	65.3	49.4	
Split-Brain (Zhang et al., 2016b)	63.0	67.1	46.7	36.0
ColorProxy (Larsson et al., 2017)		65.9		38.4
Counting (Noroozi et al., 2017)	-	67.7	51.4	36.6
(Ours) RotNet	**70.87**	**72.97**	**54.4**	**39.1**

Conclusion

❖ Conclusion

- 간단하게 Rotation 각도를 인식하게 하는 Self-supervision을 정의하여 Encoder 사전 학습
- CIFAR-10, ImageNet, Pascal VOC 2007 등에 대해서 State-of-the-art 달성

❖ 제안 방법론에 대한 나의 의견

- 'The simple is, The best'라는 말이 생각나는 제안 방법론
- 단순하게 Self-supervision을 정의했지만 쉽게 적용하다는 특징
- 왜 Network-In-Network 모델을 사용했는지, 이에 대한 Ablation study도 있었다면 좋을 듯
- AlexNet만 사용한 것 같은데 VGGNet, ResNet에 대한 실험 결과 역시 궁금

Thank you