Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning

herbarium digitization

✓ Digitized Herbarium Specimens

- 디지털화: rich environment for the application and development of machine learning techniques을 제공
- 디지털화의 한계: only capturing information from specimen labels, leaving a wealth of data in hard-to-browse specimen images
- 딥러닝: 딥러닝과 같은 기술로 metadata를 최대한 활용하려고 함. 하지만, 딥러닝은 충분한 데이터가 필요함. 방대한 양의 데이터가 있지만, 그 데이터들이 불완전한 경우가 많음. (GBIF의 경우, 전체 데이터 중 2/5만이 이미지를 이용 가능 & iDigBio의 경우 1/2만이 이미지 이용 가능)

✓ Deep Learning

- feature를 수동으로 추출하는 것이 아니라 deep learning을 통해 추출하고 싶음.
- 딥러닝은 데이터가 매우 중요함.
  - data-hungry 특성: 좋은 성과를 내기 위해선 대량의 라벨링된 이미지가 요구됨.
  - Herbarium 데이터의 경우 long-tailed distribution인 경우가 많음.

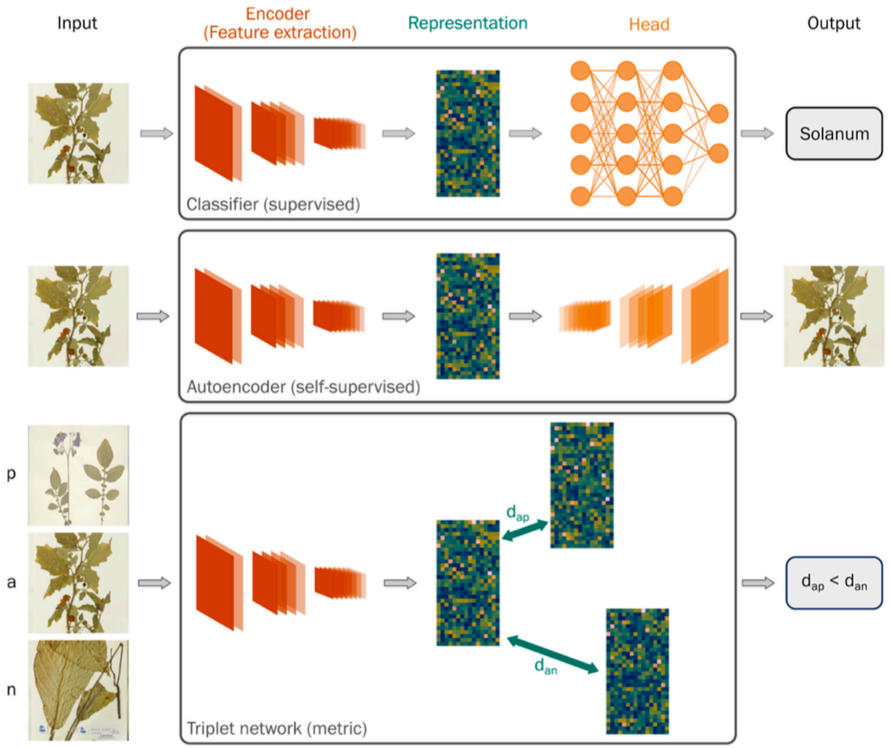
Representation Learning

✓ **Representation Learning**

- 논문에서 제안하는 방식으로 feature을 자동으로 추출하는 네트워크(encoder network)를 학습하고 난 후 different task를 수행함.
- Learning generalized representations: 일반화된 표현을 학습

✓ **샘플 이미지의 representation을 학습한 세 개의 딥러닝 네트워크를 비교해서 성능 체크**

- supervised learning: Classfier
- self-supervised learning: Autoencoder
- metric learning: Triplet network



Representation Learning

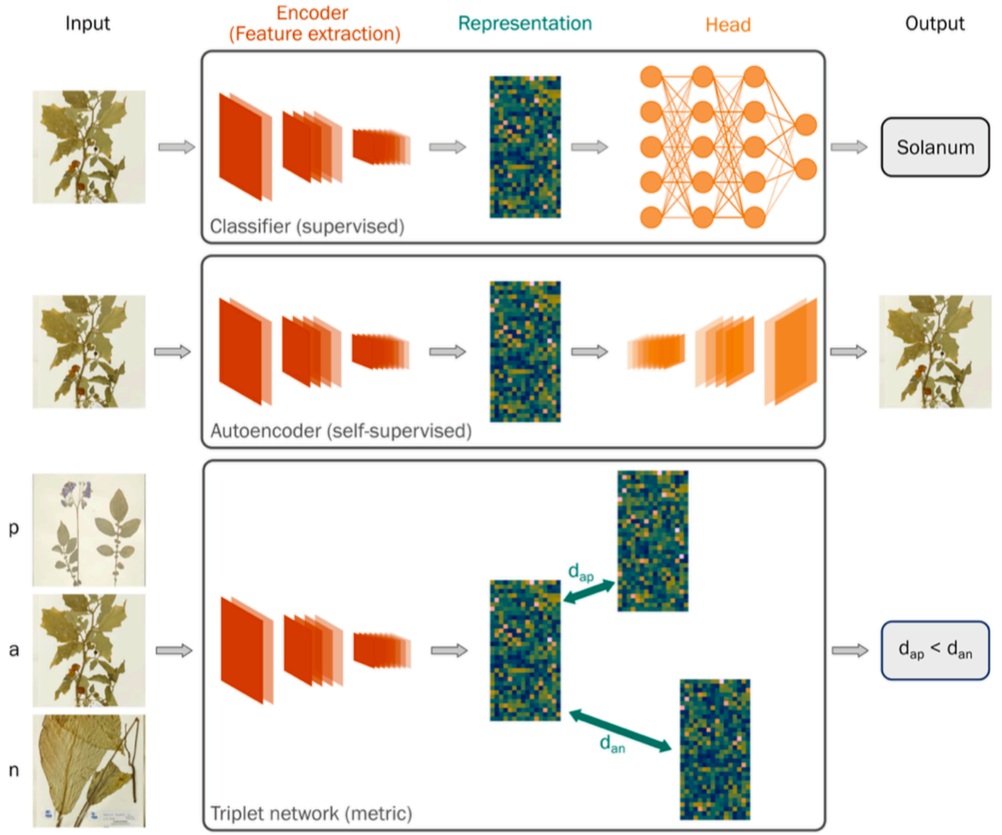
✓ **supervised learning**
- Classfier를 통해 예측

✓ **self-supervised learning**
- AutoEncoder를 통해 원본 이미지 재구성

✓ **metric learning**
- metric learning

  데이터의 유사도를 수치화해서 feature extractor를 구성
- Triplet network

  세 개의 이미지가 입력 (anchor(a)와 anchor와 같은 클래스인 positive(p), 다른 클래스인 negative(n)). 네트워크는 동일한 클래스(dap)의 샘플 간 거리를 최소화하고 다른 클래스의 샘플 간 거리(dan)를 최대화하는 feature를 학습



Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning

result

✓ Data

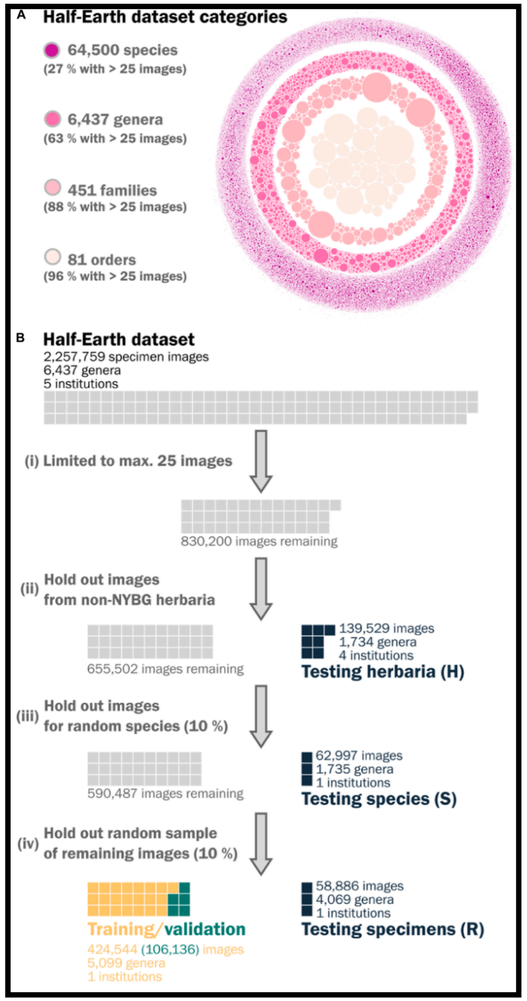
- labeled dataset (2,257,759 images covering 64,500종 in 6,437속 across 451과 and 81목 of vascular plants)
- train : validation = 8 : 2
- 최종 이미지 사이즈 = 256 x 256

✓ Neural Networks

- pre-trained ResNet-18 architecture as an encoder, producing feature representations of 512 units
- 25 epochs on a Tesla V100 GPU
- 목표: good balance between reducing the total number of classes and minimizing the variation within each class

TABLE 1 | Description of the three neural networks used for representation learning.

Model	Description	Pre-training dataset	Loss function
Autoencoder	A symmetric autoencoder with a ResNet-18 encoder, a latent space of 256 units, and a ResNet-18 decoder where convolutions have been replaced by resizing convolutions.	CIFAR-10	Mean squared error (MSE) of the reconstructed image.
Triplet network	A ResNet-18 encoder, through which three images are passed—an anchor, an example from the same class (positive), and an example from a different class (negative).	ImageNet	Triplet loss
Classifier	A ResNet-18 encoder with a classification head comprising two densely connected layers, each preceded by batch normalization and dropout layers.	ImageNet	Cross-entropy loss



Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning

적용 - 비교

✓ Applying representations

- 3개의 application을 적용함으로써 network(generalized representations)의 성능을 측정

1. **Taxonomic Identification at Different Scales**

   일반적인 분류 문제

2. **Discrimination of Similar and Distinct Genre**

   모든 가능성이 아닌 두 개의 가능한 분류군 중 어디에 속하는지 분류

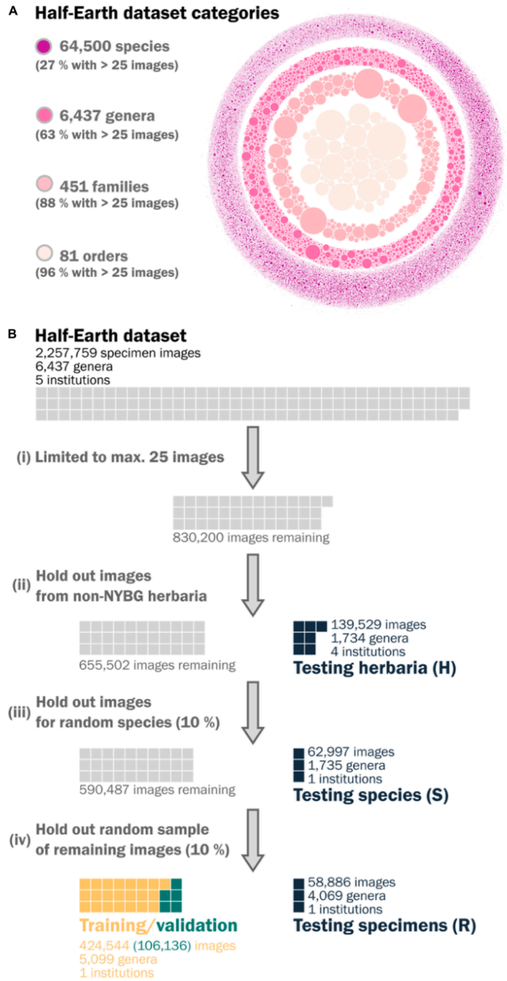
3. **Identification of Mislabeled Specimens**

   잘못 분류된 표본을 식별 (디지털화 중 실수 및 업데이트에 의한)

✓ **각 적용 케이스마다 Application Data와 Application Method 이 다름**

TABLE 2 | Description of image datasets used.

Stage	Task	Name	Source	Number of specimens
Learning representations	-	Training set	Half-earth dataset	424,544
Learning representations	-	Validation set	Half-earth dataset	106,136
Applying representations	Taxonomic identification	Herbarium test set (H)	Half-earth dataset	139,529
Applying representations	Taxonomic identification	Species test set (S)	Half-earth dataset	62,997
Applying representations	Taxonomic identification	Random specimen test set (R)	Half-earth dataset	58,886
Applying representations	Genus discrimination/Identifying mislabels	*Syzygium*	RBG, Kew	1,996
Applying representations	Genus discrimination/Identifying mislabels	*Eugenia*	RBG, Kew	8,358
Applying representations	Genus discrimination/Identifying mislabels	*Dendrobium*	RBG, Kew	1,004



Harnessing Large-Scale Herbarium Image Datasets Through Representation Learning

result

1. **Taxonomic Identification at Different Scales**

   autoencoder (most sensitive)

2. **Discrimination of Similar and Distinct Genre**

   autoencoder & triplet representations > classfier

3. **Identification of Mislabeled Specimens**

   triplet representations > autoencoder > classifier

✓ Conclusion: Triplet network is best

(metric learning through a triplet network: providing the best balance between fully and self-supervised representation learning)

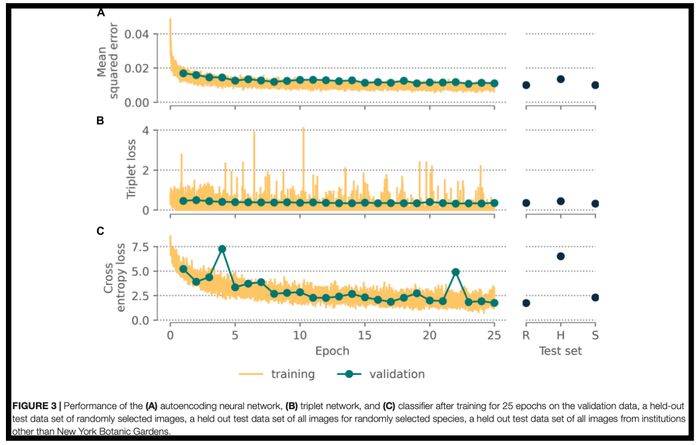


FIGURE 3 | Performance of the **(A)** autoencoding neural network, **(B)** triplet network, and **(C)** classifier after training for 25 epochs on the validation data, a held-out test data set of randomly selected images, a held out test data set of all images for randomly selected species, a held out test data set of all images from institutions other than New York Botanic Gardens.

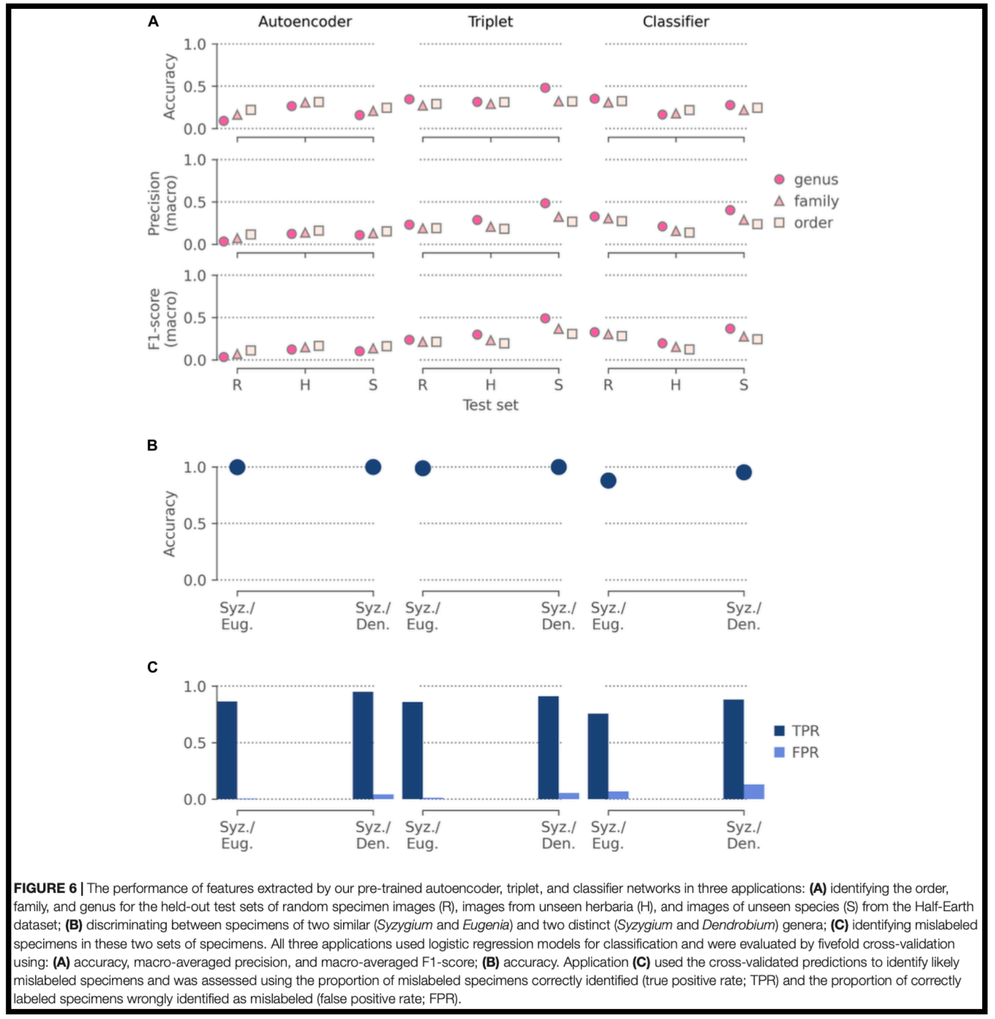


FIGURE 6 | The performance of features extracted by our pre-trained autoencoder, triplet, and classifier networks in three applications: **(A)** identifying the order, family, and genus for the held-out test sets of random specimen images (R), images from unseen herbaria (H), and images of unseen species (S) from the Half-Earth dataset; **(B)** discriminating between specimens of two similar (*Syzygium* and *Eugenia*) and two distinct (*Syzygium* and *Dendrobium*) genera; **(C)** identifying mislabeled specimens in these two sets of specimens. All three applications used logistic regression models for classification and were evaluated by fivefold cross-validation using: **(A)** accuracy, macro-averaged precision, and macro-averaged F1-score; **(B)** accuracy. Application **(C)** used the cross-validated predictions to identify likely mislabeled specimens and was assessed using the proportion of mislabeled specimens correctly identified (true positive rate; TPR) and the proportion of correctly labeled specimens wrongly identified as mislabeled (false positive rate; FPR).