MineralNet: Mineral Species Identification

Gawan Fiore, Guillermo Gomez, Jorge Cueto
Department of Computer Science
Stanford University
gfiore@stanford.edu, ggomez2@stanford.edu, jcueto@stanford.edu

Abstract

There are more than 5,300 mineral species in the world and most of the time identification is performed with scientific instruments or visually, by expert humans, both of which are costly and time consuming. Classification is an active area of computer vision research but it has not yet been applied to this specific domain. In this project, we explore traditional deep convolutional network approaches and also embedding-based approaches to solving the highly multiclass problem of mineral identification and design models that outperform a layman but do not yet match an expert.

1 Introduction

Despite a low public profile the extraction of mineral resources is becoming ever more important to sustain the growing consumption of certain natural materials by modern technology. A primary method for identification of potential ore deposits is by analysis of extracted mineral samples. However, the fields of mineralogy and geology have not yet leveraged technological advancements made available by machine learning to help solve these identification problems. While scientific tests for identification of mineral species exist, they are costlier and more difficult to access in the field compared to simple visual identification, which is almost always done manually by field researchers and exploration geologists. Similarly, mining companies spend significant resources separating debris from the specific minerals they are mining. Automated mineral species identification has the potential to be a valuable tool for geologists and mining companies, but also an educational one for the enjoyment of hobbyists and rockhounds.

A major challenge of visual identification in this domain lies in the fact that there are over 5,300 distinct mineral species, making this a massively-multi-class problem. Furthermore, geochemistry specific to localities causes significant visual variation within each species class. As a byproduct of the exponential frequency distribution often observed in nature, very few reference images exist for the majority of species. Given these factors that contribute to the high degree of variability, in this project we have chosen to examine only the most common species for which we also have the most data available.

2 Related work

Image classification is a very broad field with much active research, but there has not been any published research in this specific subdomain. Ever since the ImageNet paper in 2012 [5], the state of the art in the field has focused on ever deeper convolutional networks.

Popular general image classification models include ResNet [2] and Inception [6], both of which are convolutional neural network-based deep learning arhitectures that solve the problem of signal degradation in deep networks. ResNet employs residual connections, which are essentially straight-through copy connections between layers to propagate the signal. Inception involves multiple simultaneous convolution operations with filters of different sizes so the same layer can extract different features. Versions of both of these feature extractors that have been trained on the massive ImageNet dataset can be downloaded and fine-tuned, vastly improving train time for specific applications.

However, approaches such as these require many training examples for each category and are not especially suitable for very high numbers of classes. Since we eventually would want to classify all 5300 mineral species and perhaps also differentiate between localities (the same species in a different location and environment, which is visually distinct), FaceID-like algorithms for zero-shot [1] (no training examples for a class) or single-shot (a single or very few training examples for a class) learning could be appropriate. In this space, most systems work by creating an embedding space during train time which a new sample can be mapped into during test time. This could be as simple as Principle Component Analysis, a convolutional autoencoder [3], or more convolutional architectures built atop a typical feature extractor such as ResNet, Inception, or YOLO [4].

3 Dataset and Features

Our dataset comes from *Mindat.org*, a database containing images and taxonomic data for thousands of mineral species.

The dataset of mineral images was extracted from Mindat.org, a website database hosting images and taxonomic data for thousands of minerals organized by species and locality. From the 5300 species, we obtained verifiable images for 580 of them and from there we selected classes for which there were at least 2000 examples. This allowed us to construct a training set of 53 classes, each containing 1900 training examples, with 2.5% each held out for development and validation (50 each), for a total of 106,000 images.

We resized the images to be 360x360 pixels by 3 channels in size. During input to our models we performed random crops, horizontal flips, random pixel intensity variation, and conversions to grayscale. These modifications were intended to encourage the model to learn how to identify specimens under various light conditions and camera angles and to learn morphological features instead of focusing too much on color. In mineralogy, crystal form is often a stronger indicator of a mineral's identity than its color, though color is a more obvious feature.

4 Methods

We took two different approaches to solving the mineral classification problem.

First, we retrained the weights from Inception ResNet v2 and v3 [7] on our domain dataset and added a dense classification layer. Of the models available for retraining through Tensorflow, we selection this one because it has the largest receptive field, useful since in most of our images the mineral takes up most of the frame.

Second, we built a convolutional autoencoder to learn an embedding for input images which we then used as the weights matrix for a classifier. We tried a variety of different architectures, with varying numbers of convolutional layers, pooling layers, and batch normalization in both the encoder and decoder. The best model had three convolutional and two max pooling layers for the encoder; three convolutional and three upsampling layers for the decoder; and two fully connected layers for the classifier. The embedding size was 45x45x8, a reduction to just 4% of the original images. We trained the autoencoder with the Adadelta optimizer and mean squared error, since we wanted to reconstruct the images and required a loss function that penalized all differences. The classifier portion had an Adam optimizer and categorical cross entropy loss, to allow multi-class learning.

5 Experiments/Results/Discussion

5.0.1 Baselines

As human baselines, we asked both a trained mineralogist and someone without experience to label 1200 images, termed the *expert* and *layman* in our figures. We computed two layman scores, one from someone casually familiar with mineralogy and another after that same person spent one hour

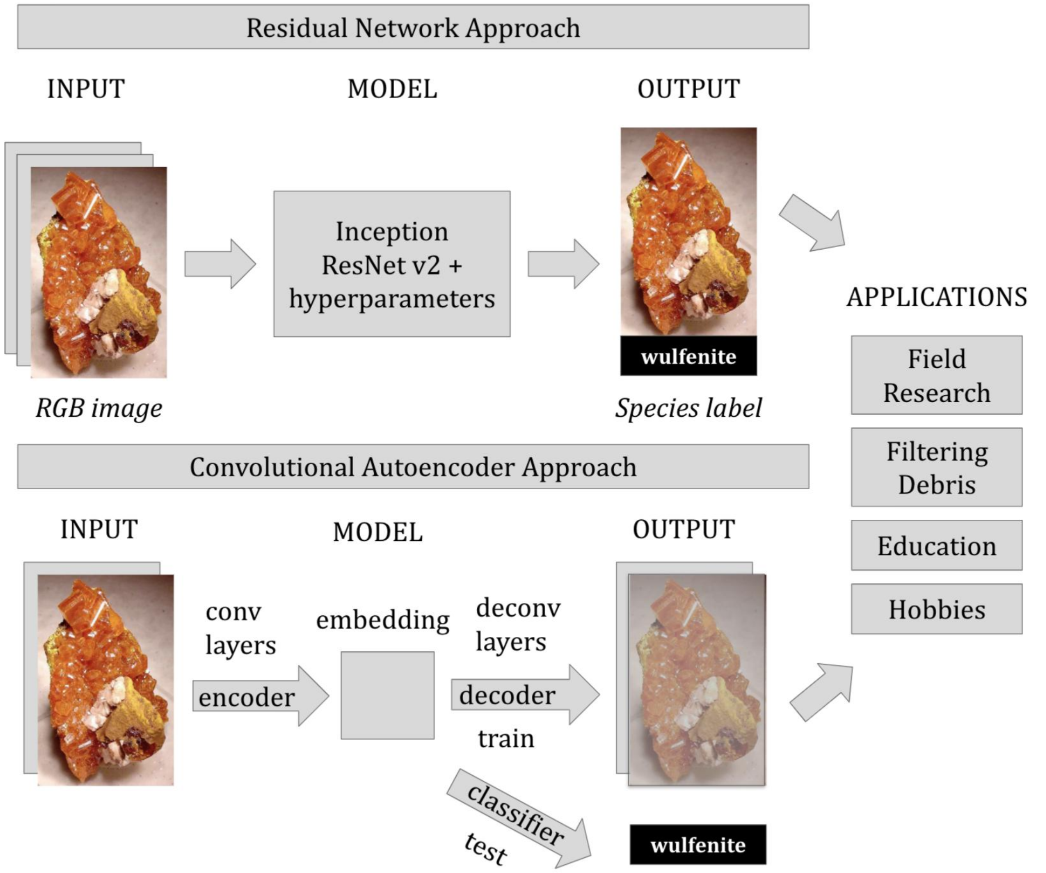
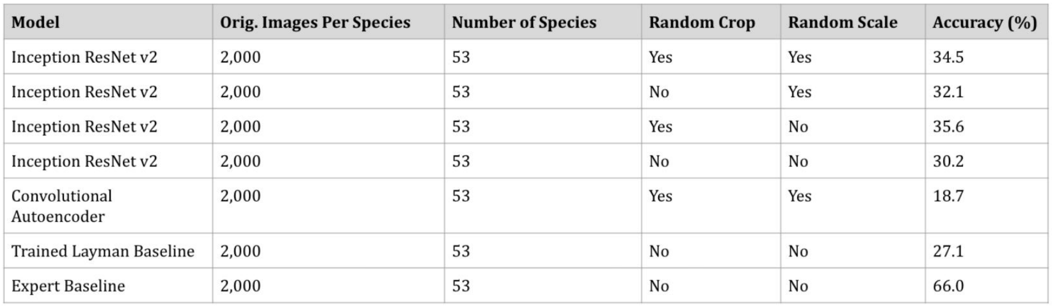


Figure 1: Effect of additional pieces of the model. Grey is the best model, with character embeddings and repurposed BiDaf self attention. Dark blue is the attempted novel full attention with character embeddings. Light blue is just self attention and orange is just character embeddings.



Model	Orig. Images Per Species	Number of Species	Random Crop	Random Scale	Accuracy (%)
Inception ResNet v2	2,000	53	Yes	Yes	34.5
Inception ResNet v2	2,000	53	No	Yes	32.1
Inception ResNet v2	2,000	53	Yes	No	35.6
Inception ResNet v2	2,000	53	No	No	30.2
Convolutional Autoencoder	2,000	53	Yes	Yes	18.7
Trained Layman Baseline	2,000	53	No	No	27.1
Expert Baseline	2,000	53	No	No	66.0

Figure 2: Performance of various models on 53 classes.

with a list of the possible classes in which to research the various species. They accurately labeled 18% and 27% of images in each of these conditions. Even the expert had trouble correctly identifying many of the images, correctly labeling 66% of images, demonstrating just how difficult this problem is.

A random classifier, selecting labels for each image uniformly, would correctly label only 1.8% of images.

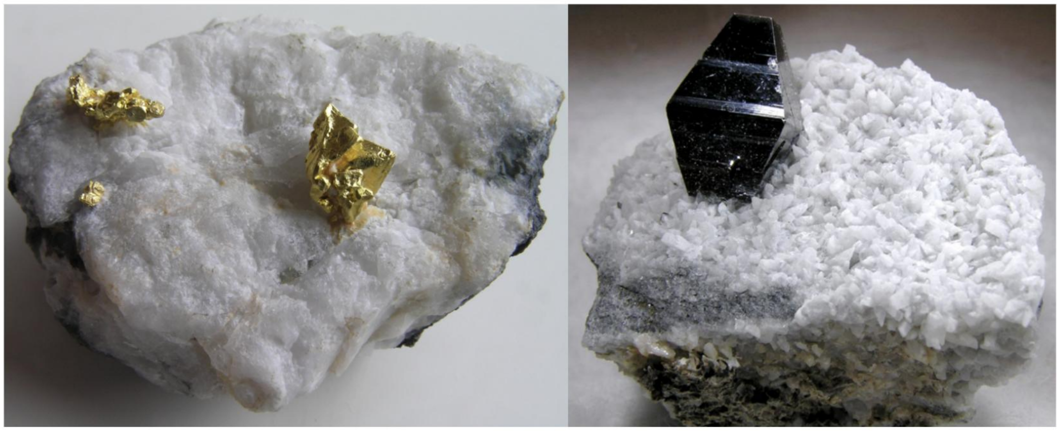


Figure 3: Gold (on the left) and Anatase (on the right), with similar quartz-based matrices, but clearly different crystal form and color.

5.0.2 Models

Using the retrained Inception ResNet classifier, we ran several experiments to test the effects of retraining amount, number of classes, and image preprocessing.

With the autoencoder, we experimented with architecture, embedding size, and number of classes. We found that a larger hidden size in the classifier improved performance more than modifications to the dimensions or number of layers in the autoencoder. From the literature, it seems that the number of layers we used is fairly standard.

For both models, as expected, including fewer classes improves results. It is interesting to consider, however, how the choice of which classes to include effects performance. When run over five classes for which the models both generally performed well (above 40% accuracy), Aquamarine, Dioptase, Galena, Pyrite, Wulfenite, they each continued to perform well, with 87% accuracy for the autoencoder and 95% accuracy for the retrained ResNetv2 model. When classes were selected entirely as random, performance declined to just 49% and 68%, still higher than for more classes but much lower.

Interestingly, the layman and expert did not in general make the same mistakes as each other or as the models, although there are exceptions. Humans seem to be attending to different details in the images than the models are, as discussed further below.

For both humans, there were come classes that they had apparently learned entirely and for which they had very high precision, and others where they seemed to chronically misclassify. Computer models do not show this same trend and have much more uniform accuracy profiles. However, the best model did much better on colorful species than less colorful. For example, it scored just 12% on Hematite (dark and metallic) versus 36% on Pyromorphite (bright and colorful).

However, the ResNetv2 retrained model often made mistakes that make sense to the human eye. For example, when the correct label was Hematite, the majority of the mistaken labels were other minerals that are also dark and metallic, like Stibnite, Anatase, and Magnetite. When the correct label was Anatase, however, the model seemed to pick up more on the shape than the color, as it selected other minerals with similar crystal form such as Zircon. Like humans, the model frequently mistook various gemmy minerals for each other, like Aquamarine, Elbaite, and Topaz. These were mistakes that the layman, but not the expert, also made. For the mineral Pyromorphite, the majority of mislabelings were Mimetite and Vanadinite, both of which are similar in shape and color. Humans also often mistake Mimetite and Pyromorphite, and the visual similarities arise because they are chemically related.

Interestingly, there were a number of mistakes that a human would never make in which the model cued off of the matrix on which the crystal was situated, rather than the crystal itself. For example, Anatase and Gold (seen in Figure 3, above) were frequently mistaken for each other and have similar matrices but no similarity in crystal structure or color, so the model must be focusing on the matrix. Neither person ever made this mistake.

Notably, if we allowed the model to output its top two guesses rather than just one, the accuracy almost doubled, from 35.6% to 58.1%. This implies that more sophisticated ranking or final answer selection could improve performance significantly.

6 Conclusion/Future Work

Since even an expert human only correctly labeled 66% of images, this is a problem space in which there is room for computers to exceed humans in accuracy. However, the high number of classes and similarity in appearance of different mineral species means this is a very difficult problem. Training a model that begins to understand elements of crystallography, transparency, and association with other species will be essential to surpassing expert identification.

Due to limited time, we retrained Inception ResNetv2 off of its already learned weights, which gave a huge initial boost due to low-level features already baked into the model, but limited the specificity of what we were able to detect. Training a similar model end-to-end would allow us to tune the receptive field and filter sizes to pick up features particular to minerals. Our convolutional autoencoder model showed promise, but since it did not start from pretrained weights our dataset is too small for the model to be generalizable. With a larger dataset and more time, we hypothesize this would be the best architecture to pursue.

Taking a different approach to post-encoding classification also shows promise. Instead of relying on a neural classifier architecture, moving forward we may try a nearest neighbor search in a tuned embedding space as a method of classification of the autoencoder output. This and other embedding-based approaches likely hold the most promise as they have been shown in the literature to work well on tasks where there are few examples of each class and the classes may be similar to each other (ie. in facial recognition). We simply have not have enough time or resources to explore all of the embedding approaches yet.

People who are very good at identifying minerals also use features that our models did not have available, like the total size of the specimen and, at times, where or how long ago it was found. Incorporating such environmental features, for example using bucketized one-hot vectors for time, country of origin, and physical size, could help the model's performance improve.

Finally, minerals in the field look very different than they do in a museum or gallery setting. Acquiring a broader set of training data that represents more of these modalities would make results for any model more generalizable.

7 Contributions

Gawan Fiore scraped the data initially from Mindat, performed data sanitizaiton/augmentation, set up the GPU cluster, helped set up the ResNet models, wrote and trained the autoencoder, analyzed classification performance for both models, and wrote most of the paper and proposal.

Guillermo Gomez set up and finetuned the ResNet model, including acquiring the pre-trained weights, ran hyperparameter experiments, and helped with the paper and poster.

Jorgue Cueto helped write the proposal, wrote the milestone, and designed the poster.

8 Code

GitHub Repo: https://github.com/gawanfiore/min-id

References

[1] Bansal, et al. *Zero-Shot Object Detection*. 2018. https://arxiv.org/pdf/1804.04340.pdf

[2] He, et al. *Identity Mappings in Deep Residual Networks*. 2016. https://arxiv.org/pdf/1603.05027.pdf

[3] Mao, et al. *Feature Representation Using Deep Autoencoder for Lung Nodule Image Classification*. 2018. https://www.hindawi.com/journals/complexity/2018/3078374/

[4] Redmon, Joseph and Farhadi, Ali. *YOLO9000: Better, Faster, Stronger*. 2016. https://arxiv.org/pdf/1612.08242.pdf

[5] Sutskever, et al. *ImageNet Classification with Deep Convolutional Neural Networks*. 2012. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[6] Szegedy, et al. *Going deeper with convolutions*. 2014. https://arxiv.org/pdf/1409.4842.pdf

[7] Tensorflow ResNet v2. https://github.com/tensorflow/models/blob/master/research/slim/nets/inception_resnet_v2.py