Analyzing the Benefits and Limitations of Transfer Learning

By: Sherwin Amal, Nathan Cheng, Carl Song

Background

Transfer learning is the idea of reusing a pre-trained model for the purposes of solving a different task. It involves training a model on one dataset and using that trained model to train a separate dataset. It's meant to improve the performance for a model on a specific dataset, especially for datasets with limited data availability for training, validating, and testing. It can help with not only accuracy but efficiency as well (Hosna et al., 2022). Transfer learning is useful when there's a lack of quality data or resources to process large scale databases. Lots of research is poured into transfer learning because of its ability to enhance generalization and stability (Zhou et al., 2021). There's many types of transfer learning, but the most relevant type is model-based transfer learning (Zhou et al., 2021). This involves training a model on a dataset, and reusing it for a different dataset, to reach high testing accuracies. This paper will be a comparison analysis discussing transfer learning and understanding its effectiveness for various dataset sizes. This will require developing an elaborate model architecture to learn deep features of a dataset, and apply the model to classify two different datasets. The results will be used to compare the limitations of transfer learning as well.

Identifying Datasets

Quality datasets must be identified to evaluate model training and transfer learning. This became a focal point of the project. The goal was to identify a dataset with a small number of labels, and a large dataset. The second and third datasets will have the same number of labels, both more than the first dataset. The third database should be smaller compared to the second database. This is to showcase the benefits and potential limitations of transfer learning based on the dataset. A high accuracy model trained on the first dataset will then be used to transfer its learning to the other two datasets. The databases selected were based on flowers, Indonesian cuisine, and fruits. The flowers database had 5 labels with approximately 750-1000 images per label. The Indonesian cuisine database had 9 labels, but fewer images, with approximately 100 images per label. The fruit dataset had 9 labels as well but even fewer images per label, with approximately 40 images per label. This provides the opportunity to compare model accuracies for these databases and analyze the effectiveness of transfer learning on different types of datasets.

Dataset Preprocessing, Training, and Testing

Images in databases have to be properly processed for effective training. Image preprocessing and transformation has been found to improve training speeds and accuracy (Yousif & Balfaqih, 2024). Batches and images were outputted to gauge dataset quality (Saxena, 2021). The transformations applied to the selected databases included random cropping, normalizing, flipping on the horizontal axis, modifying brightness, applying contrast, saturation, and hue. Below is an example of the original image compared to the transformed image that will help the model train properly with high accuracy and confidence.

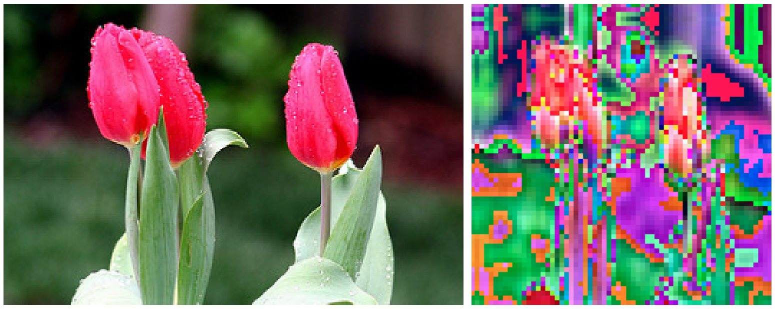

Figure 1: This shows the difference between an original image in the flowers dataset (left) and the same image after transformations (right).

The training and testing steps were based on the ImageClassificationBase class, which is commonly used (Purakkatt, 2020). The ImageClassificationBase class was modified to handle GPU acceleration for training. This testing and training environment optimized training speeds, which was beneficial for large-scale image datasets. Images were split into training, validation, and testing datasets to analyze model performance, reduce overfitting, and ensure the model can classify unseen images. 15% of the dataset was restricted to validation, another 15% was restricted to testing, and the remaining 70% of the dataset was meant for training. Additionally, a GPU device is used to further increase training rates, using Kaggle Notebooks, which offer an online environment to execute Python notebooks. CrossEntropyLoss criterion was used to calculate the loss based on model accuracy at every epoch. The model is trained on a training dataset, and evaluated further with a validation dataset. After a certain number of specified epochs, the model attempts to classify images from the testing dataset. Testing images are not used in training. These results can be used to evaluate the model's ability to generalize patterns and identify key classification traits in images not encountered in the training process.

Developing The Model Architecture

The model architecture was a 2D convolutional neural network designed to learn deep features from a dataset. The architecture contained 4 repeated sections, followed by a 2D pooling and linear layer. The repeated blocks contained two convolutional, batch normalization, ReLU layers, and a single max pooling layer at the end of the section. This architecture is considered "state-of-art", with the usage of residual blocks (Sahoo, 2022). Deeper layers had more in-channels and out-channels for greater capacity to analyze deep patterns in image databases (Saxena, 2021). The shallow initial layers are meant to learn simple features. When having more or less than four repeated sections of the same set/order of layers, the model accuracies decreased for the initial computational experiment due to either underfitting or overfitting. The ReLU and Maxpool layers serve to introduce nonlinear activations, reduce dimensionality, overfitting potential, and computational complexity (Schilling, 2016). Additionally, the batch normalization layer accelerates the model's training time (Schilling, 2016). The model architecture is designed to adapt easily to varying numbers of input channels and output classes. This flexibility is particularly advantageous for transfer learning, especially when the target dataset has a different number of labels than the original training dataset.

Training Separate Models without Transfer Learning

The first computational experiment was to train separate models with the same architecture on the three databases mentioned. This was to understand how effective a model is on various datasets without transfer learning. The hypothesis is that the model will perform better on fewer labels and more images (flowers database) than the other two databases without transfer learning. The model trained on the flowers dataset had a testing accuracy of 78%. The validation loss fluctuated between 0.6 to 0.8. The model trained on Indonesian cuisine had a testing accuracy of 68%, with the validation loss fluctuating around 0.9 to 1.1. The model trained on fruits had a 60% testing accuracy, with the validation loss fluctuating around 0.9 to 1.2.

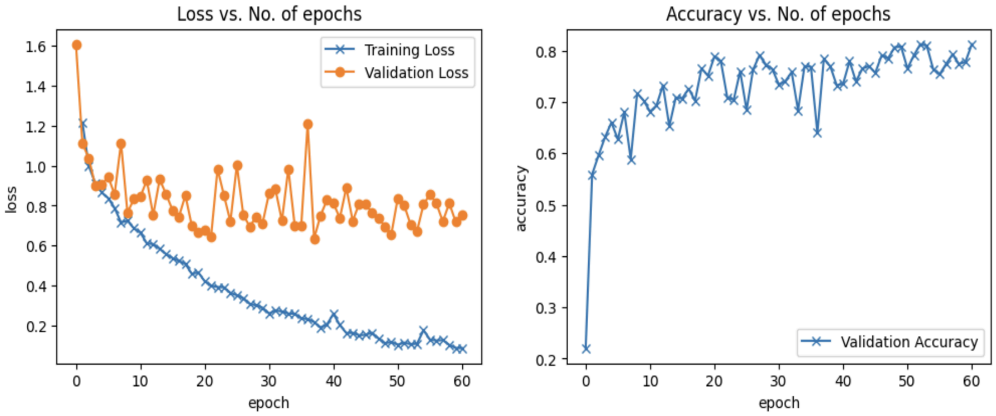

Figure 2: This shows the training loss and validation losses (left) and the validation accuracy across the 60 epochs run (right) for the Flowers dataset.

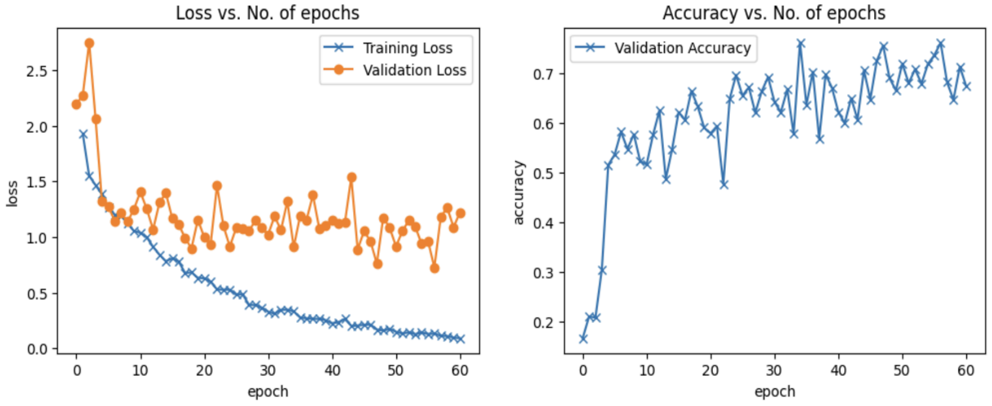

Figure 3: This shows the training loss and validation losses (left) and the validation accuracy across the 60 epochs run (right) for the Indonesian cuisine dataset.

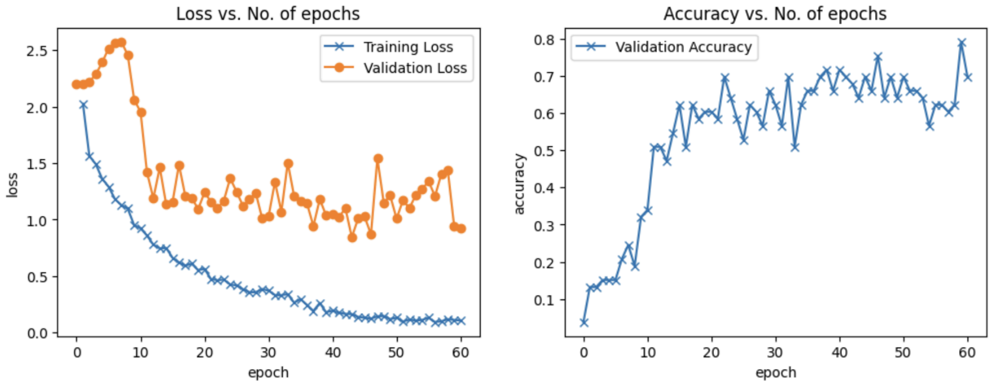

Figure 4: This shows the training loss and validation losses (left) and the validation accuracy across the 60 epochs run (right) for the fruits dataset.

When a model was trained on fruits and Indonesian cuisine datasets, there was heavy variation in validation loss. It seemed as if the model didn’t fully understand the datasets. One explanation is that these datasets had more labels but less images to train from. As a result, the model had few opportunities to distinguish features between various labels and categories. For all three datasets, the training loss gradually decreased. But by the end of the training process, the validation loss ended up stabilizing. The flowers model did an effective job at generalizing features of various flowers. It was learned that the model most likely benefited from less labels and more training images, hence the predicted hypothesis was correct. Because of the flowers model’s ability to confidently classify flower images, it will be used to transfer its learning to the Indonesian cuisine and fruits datasets separately. Its lower layers can be reused to classify common characteristics in these two datasets.

Improving Model Accuracy with Transfer Learning

The second experiment involved transferring the learning from a high precision model to other datasets, and comparing results. This experiment was meant to showcase model improvements from transfer learning. The hypothesis is that the fruits dataset will benefit more from transfer learning compared to the Indonesian cuisine dataset because there appears to be more similar characteristics to the flowers database which a model is initially trained on. A model replica was made and trained on flowers. To utilize this model for transfer learning, the model's final linear layer was modified to increase the number of labels to 9 for the remaining two datasets, Indonesian cuisine and fruits. These models were immediately set to train on the Indonesian cuisine and fruits datasets. The Indonesian cuisine dataset benefited greatly from the transfer learning process, with a testing accuracy of 81%. The training loss decreases gradually to values of .06 by the end of the simulation. Because of the small learning rate, there were fewer aggressive jumps in learning accuracy, but steadily the model improved over time. Even with the high testing accuracy and increasing validation accuracies, the validation loss stabilized between 0.6 to 0.8. This could be considered a good indicator of the model with stable and effective generalization, and no signs of overfitting.

However, the fruits dataset still seemed to have some struggles with model accuracy. There certainly was an improvement with transfer learning but there could've been several reasons as to why there wasn't a greater benefit specifically for the fruits dataset. The biggest potential reason was because this dataset had much less images compared to the Indonesian cuisine dataset, which is three times larger than the fruits dataset. Additionally, it could be that there weren't enough layers in the model to notice deeper features found in the fruits dataset for higher model accuracy. The model's validation and testing accuracy plateaued around the upper 60% and lower 70%. The final testing accuracy was at 70%. By the end of the simulation for the fruits dataset, the training loss leveled to around .02, while the validation loss stabilized to 0.6 to 0.8, with an extremely high outlier in the first few epochs, as the model adjusts to a different dataset outside of flowers. This was very interesting because flowers and fruits share more visual characteristics compared to flowers and Indonesian cuisine, so the expectation was that the fruits model would perform significantly better.

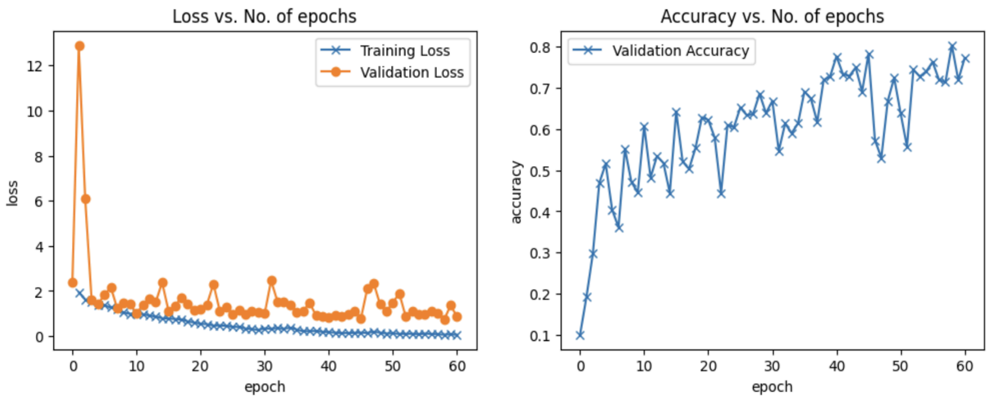

Figure 5: This shows the training loss and validation losses (left) and the validation accuracy across the 60 epochs run (right) for the Indonesian cuisine dataset after transferring learning from the initial training that occurred on the flowers database.

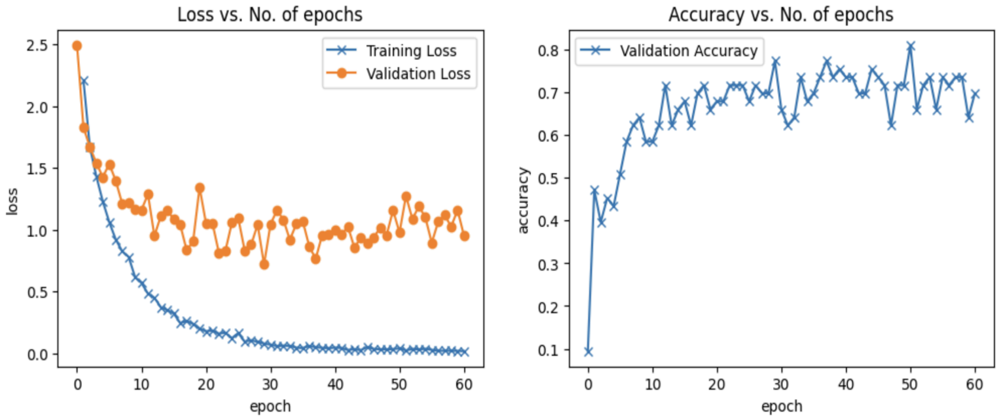

Figure 6: This shows the training loss and validation losses (left) and the validation accuracy across the 60 epochs run (right) for the fruits dataset after transferring learning from the initial training that occurred on the flowers database.

Regardless, transfer learning for both datasets benefitted them, taking less epochs to reach high validation accuracies without overfitting. Here, the predicted hypothesis is incorrect. Both models equally increased in accuracy by approximately 10-13%, which is still a significant improvement nonetheless. It was learned that there are limitations of transfer learning, being still dependent on a dataset’s quality and quantity for performance and accuracy.

Conclusion and Future Steps

These two computational experiments indicated the capabilities of transfer learning, along with their limitations by the datasets themselves. The first computational experiment showcased the initial training results, while the second computational experiment showed the generally positive effects transfer learning had. The model architecture allowed for deep pattern recognition in all the datasets, to various extents. However, if given more time, there could have been further analyses to increase image classification capabilities. Research has shown that methods like grayscaling and further image transformation can improve classification, and that can be something worth looking into (Kanan & Cottrell, 2012). Additionally, there can be various regularization methods incorporated into these computational experiments, such as weight decay, dropout, early stopping, CutMix, Mixup, and etc (Santos & Papa, 2018). These techniques are involved in the input and training periods for a model. These regularization techniques might improve initial training processes, along with improving the effect of transfer learning.

Works Cited

Afrinanto, F. F., (2022). Padang Cuisine (Indonesian Food Image Dataset) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/4053613

Hosna, A., Merry, E., Gyalmo, J., Alom, Z., Aung, Z., & Azim, M. A. (2022). Transfer learning: A friendly introduction. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00652-w

Kanan, C., & Cottrell, G. W. (2012, January 10). *Color-to-grayscale: Does the method matter in image recognition?*. PLOS ONE. https://doi.org/10.1371%2Fjournal.pone.0029740

Maher, S. (2023). Fruits Dataset (Images) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/5514079

Mamaev, A (2021), Flowers Recognition [Data set]. Kaggle. https://www.kaggle.com/datasets/alxmamaev/flowers-recognition/data

Purakkatt, A. (2020, August 8). *Image Classification with Pytorch*. Medium. https://medium.com/analytics-vidhya/image-classification-with-pytorch-184e76c2cf3b

Santos, C. F. G. D., & Papa, J. P. (2018, January). *A survey on regularization methods for convolutional neural ...* ACM Digital Library. https://arxiv.org/pdf/2201.03299.pdf

Sahoo, S. (2022, September 26). *Residual blocks - building blocks of Resnet*. Medium. https://towardsdatascience.com/residual-blocks-building-blocks-of-resnet-fd90ca15d6ec

Schilling, F. (2016, August 26). *The effect of batch normalization on deep convolutional Neural Networks*. DIVA. https://kth.diva-portal.org/smash/record.jsf?pid=diva2%3A955562&dswid=9852

Saxena, S. (2021b, June 19). *Classifying cifar-10 using resnets - pytorch*. Mr-siddy Experimental Lab. https://mr-siddy.github.io/ML-blog/deep_learning/2021/06/19/cifar10-resnets.html

Saxena, S. (2021, June 9). *Image Classification using Convolutional Neural Networks - Pytorch*. Mr-siddy Experimental Lab. https://mr-siddy.github.io/ML-blog/deep_learning/2021/06/09/Image-Classification-CNN-pytorch.html

Yousif, M. J., & Balfaqih, M. (2024). Enhancing the accuracy of image classification using Deep Learning and preprocessing methods. *Artificial Intelligence &amp; Robotics Development Journal*, *3*(4), 269–281. https://doi.org/10.52098/airdj.2023348

Zhou, Y., Zhang, X., Wang, Y., & Zhang, B. (2021). Transfer learning and its application research. *Journal of Physics: Conference Series*, *1920*(1), 012058. https://doi.org/10.1088/1742-6596/1920/1/012058

List of Contributions and Roles of each Group member to the Project

- Sherwin Amal
    - Identifying datasets and preprocessing
    - Identifying and utilizing training and testing architectures
    - Designed presentation
    - Helped with running computational experiment #1
    - Worked on drafting ideas and helping with writing of paper
- Nathan Cheng
    - Responsible for training the model for both transfer learning training on the Fruits and Indonesian food datasets
    - Designed model architecture for datasets to run on
    - Improved training and testing by utilizing GPUs
    - Wrote and formatted paper to be more readable
    - Screen recorded presentation for the video
- Carl Song
    - Revision work on final paper and presentation
    - Fine-tuned procedures and set up both computational experiments
    - Helped with running computational experiment #2
    - Scheduled discord calls for meetings on next tasks

Modelcard

Model Description

- Developed by: Sherwin Amal, Carl Song, Nathan Cheng
- Model type: Convolutional Neural Network
- Languages: Python
- License: N/A

Model Details

- The model consists of four sets of convolutional layers followed by max-pooling layers, batch normalization, and ReLU layers to extract hierarchical features from the input images.
- The flattened output from the convolutional layers is passed through one fully connected layer, which acts as a classifier.
- ReLU activation functions are used throughout the model to introduce non-linearity.
- The model ends with a final linear layer producing logits for each of the output classes.

Uses

- The model is used to distinguish different types of items such as fruits, indonesian cuisine, and flowers

- This model is also designed to be flexible to handle different datasets with different amounts of classes. This helps for the model to be able to train on different data and is the basis for transfer learning

Biases, Risk, and Limitations

- N/A

How to Get Started with the Model

- Initialize the model: to_device(base_model(3, number_of_labels), device)
- The to_device() method(moves tensor to chosen device) and device must already be defined.

Training Details

- Before being used by the model, the data was subjected to random cropping, normalizing, flipping on the horizontal axis, modifying brightness, and applying contrast, saturation, and hue. Doing this ensures that the model is able more accurately predict a given image.
- The training and testing steps were based on the ImageClassificationBase class (Purakkatt, 2020)
- The model went through 60 epochs and an Adam optimizer for all the datasets
- Images were split into training (70%), validation(15%), and testing(15%) datasets to analyze model performance, reduce overfitting, and ensure the model can classify unseen images

Evaluation

- Overall, the model did a great job in distinguishing images. This model was tested on its capabilities to learn datasets, and generalize patterns to be transferable to other datasets.
- The testing data was Indonesian cuisine, fruits, and flowers for the two computational experiments.
- While there was a high rate of change in accuracy in all the datasets, the model showed quantifiable improvements in training and validation accuracies through transfer learning.

Datasheet for Indonesian Food Dataset

For what purpose was the dataset created?

This dataset was created to identify common foods from the Indonesian/Padang cuisine, with 9 of the most popular Padang cuisine foods.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

This dataset was created by Faldo Fajri Afrinanto, on behalf of no entity.

Who funded the creation of the dataset?

There was no funding for the creation of the dataset, as the image dataset was collected with Bing Image Downloader, a python library.

Any other comments?

No

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?

The dataset has many images of the 9 most popular Padang cuisine foods.

How many instances are there in total (of each type, if appropriate)?

There are a total of 993 images in this dataset. There are 9 different types of Padang cuisine. Ayam goreng has 107 images, ayam pop has 113 images, daging rendang has 104 images, dendeng batokok has 109 images, gulai ikan has 111 images, gulai tambusu has 103 images, gulai tunjang has 119 images, telur balado has 111 images, and telur dadar has 116 images.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

These images are only a sample, as there are more images for each food type besides the images collected for the dataset.

What data does each instance consist of?

These images are just the different types of food.

Is there a label or target associated with each instance?

With each instance there is a label associated, and the name of the image file will show the appropriate matching label.

Is any information missing from individual instances?

There is not any information missing from individual instances.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?

There are no relationships between individual instances made explicit.

Are there recommended data splits (for example, training, development/validation, testing)?

Are there any errors, sources of noise, or redundancies in the dataset?

No, there are not recommended data splits, and there are no errors, sources of noise, or redundancies from the dataset. Out of preference, the datasets were split into 70% training, 15% validation, and 15% was stored for testing.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?

The datasets do not link or rely on external resources, although the images were retrieved from Bing.

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?

There is no data in this dataset considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

There is no offensive, insulting, threatening, or anxiety inducing data in this dataset.

Datasheet for Flowers Dataset

For what purpose was the dataset created?

This dataset was created to identify different types of flowers, with 5 total types of flowers.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

This dataset was created by Alexander Mamaev, on behalf of no entity.

Who funded the creation of the dataset?

There was no funding for the creation of the dataset, as the image dataset was collected through Flickr, Google Images, and Yandex Images.

Any other comments?

No

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?

The dataset has many images of the 5 types of flowers.

How many instances are there in total (of each type, if appropriate)?

There are a total of 4317 images in this dataset. There are 5 different types of flowers. Daisies have 764 images, dandelions have 1052 images, roses have 784 images, sunflowers have 733 images, and tulips have 984 images.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

These images are only a sample, as there are more images for each flower type besides the images collected for the dataset.

What data does each instance consist of?

These images are just the different types of flowers.

Is there a label or target associated with each instance?

With each instance there is a label associated, and the name of the image file will show the appropriate matching label.

Is any information missing from individual instances?

There is not any information missing from individual instances.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?

There are no relationships between individual instances made explicit.

Are there recommended data splits (for example, training, development/validation, testing)?

Are there any errors, sources of noise, or redundancies in the dataset?

No, there are not recommended data splits, and there are no errors, sources of noise, or redundancies from the dataset. Out of preference, the datasets were split into 70% training, 15% validation, and 15% was stored for testing.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?

The datasets do not link or rely on external resources, although the images were retrieved from Flickr, Google Images, and Yandex Images.

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?

There is no data in this dataset considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

There is no offensive, insulting, threatening, or anxiety inducing data in this dataset.

Datasheet for Fruits Dataset

For what purpose was the dataset created?

This dataset was created to identify common fruits, with 9 common types of fruit.

Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

This dataset was created by Shreya Maher, on behalf of no entity. This person is a student at Northeastern University.

Who funded the creation of the dataset?

There was no funding for the creation of the dataset, as the image dataset was collected from the internet.

Any other comments?

No

What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?

The dataset has images of 9 common fruits.

How many instances are there in total (of each type, if appropriate)?

There are a total of 360 images in this dataset. There are 9 different types of fruits in this dataset. Apples have 40 images, bananas have 40 images, cherries have 40 images, chikoos have 40 images, grapes have 40 images, kiwis have 40 images, mangos have 40 images, oranges have 40 images, and strawberries have 40 images.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?

These images are only a sample, as there are more images for each food type besides the images collected for the dataset.

What data does each instance consist of?

These images are just the different types of fruits.

Is there a label or target associated with each instance?

With each instance there is a label associated, and the name of the image file will show the appropriate matching label.

Is any information missing from individual instances?

There is not any information missing from individual instances.

Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)?

There are no relationships between individual instances made explicit.

Are there recommended data splits (for example, training, development/validation, testing)?

Are there any errors, sources of noise, or redundancies in the dataset?

No, there are not recommended data splits, and there are no errors, sources of noise, or redundancies from the dataset. Out of preference, the datasets were split into 70% training, 15% validation, and 15% was stored for testing.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?

The datasets do not link or rely on external resources, although the images were retrieved from the internet.

Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?

There is no data in this dataset considered confidential.

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?

There is no offensive, insulting, threatening, or anxiety inducing data in this dataset.

-