IMAGE

Classification
Recommendation

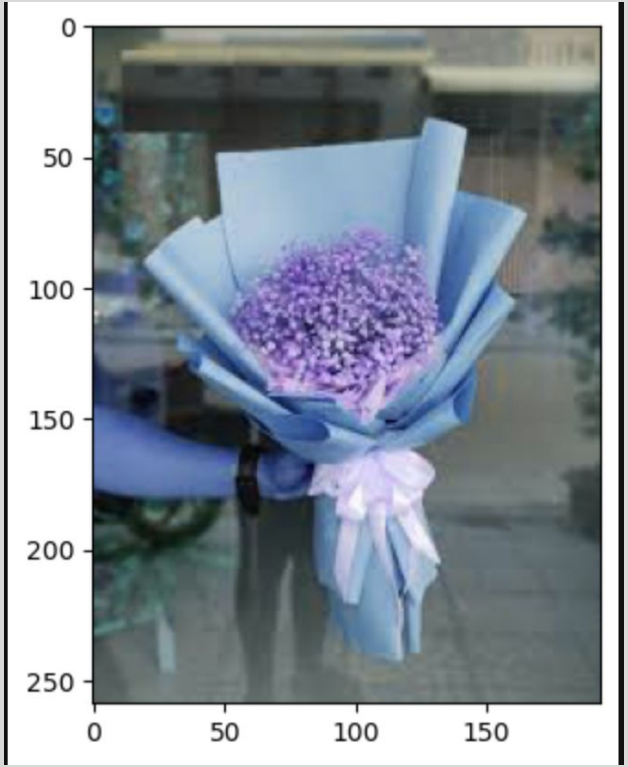
RMIT UNIVERSITY

Exploratory Data Analysis

Here are 8 types of flower and their quantities in the dataset:

- Calimero : 244 flowers
- Babi : 848 flowers
- Pingpong : 273 flowers
- Chrysanthemum : 649 flowers
- Rosy : 160 flowers
- Tana : 534 flowers
- Hydrangeas : 491 flowers
- Lisianthus : 849 flowers

Flower Images



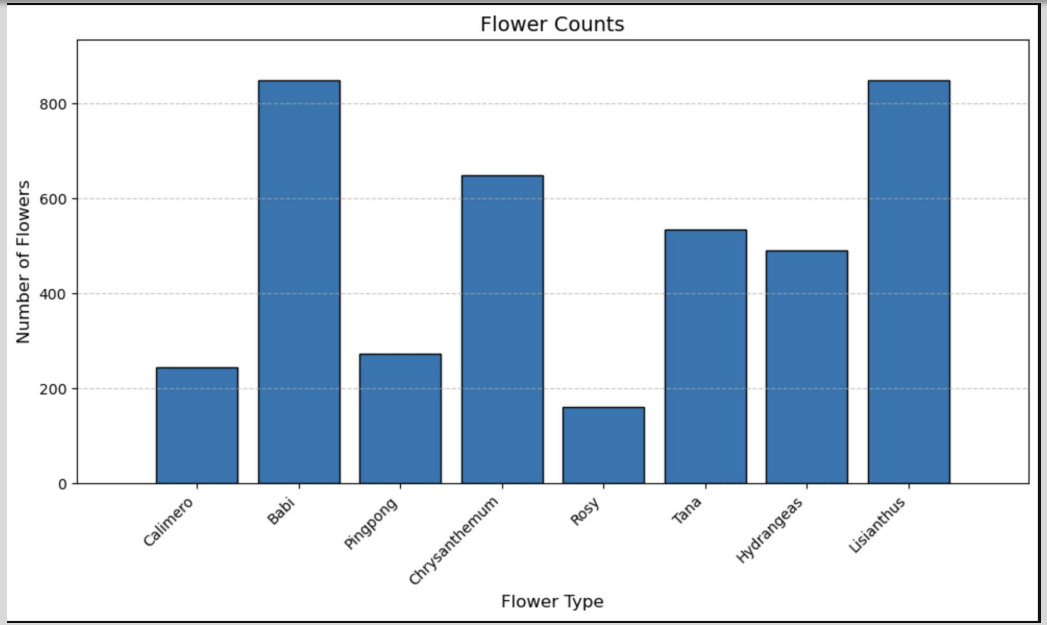
Dirty Dataset

Mixed Different Flower Together

Image Includes Irrelevant Things

Good Image with Noise

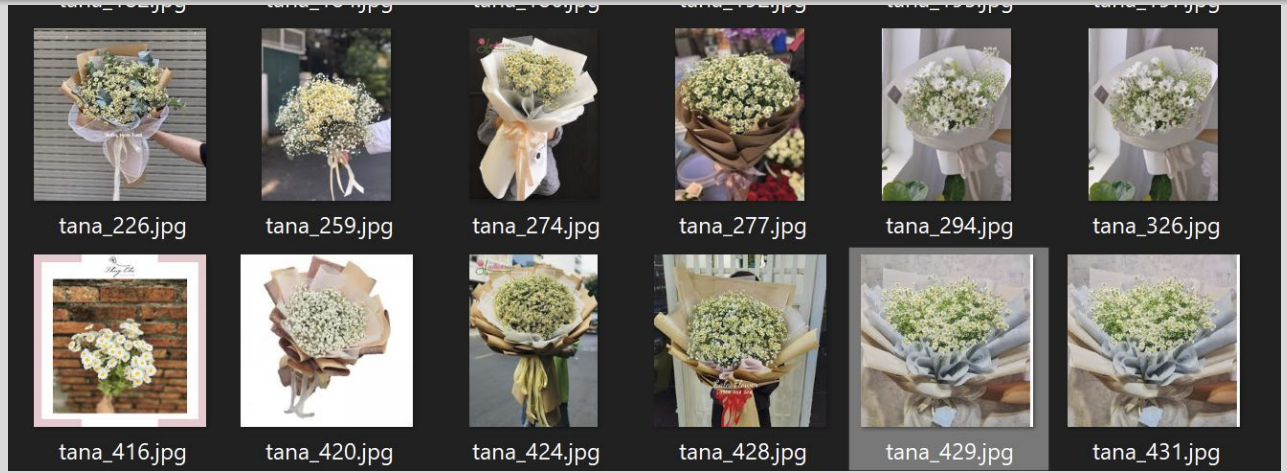
Imbalanced Dataset



Saini *et al.** studied the effect of imbalanced image dataset on image classification and found that “results in the biased classification towards the majority class”.

*Reference: Saini, M., & Susan, S. (2019). Comparison of Deep Learning, Data Augmentation and Bag of-Visual-Words for Classification of Imbalanced Image Datasets. Recent Trends in Image Processing and Pattern Recognition, 561–571. doi:10.1007/978-981-13-9181-1_49

Data Cleansing - Duplicate Images



In a study on phishing detection CNN , Zhu et al.[1] found that duplicate data inside a dataset “will trap the training of the neural networks into the problem of over-fitting”. Thus, we believe that duplication handling is needed to solve this problem

Data Cleansing - Dirty Images

Even though, there are several frameworks such as ImageDC[2] for automated data cleaning, we think manual cleansing was needed:

- First is that the relatively small dataset makes it a possibility.
- Second is because we felt that training a neural network to account for all the data irregularities would be challenging and time consuming.
- During our manual cleansing, we iterate through all the images in each flower species and delete any images that are either wrongly labelled or are ambiguous.

Additional: Image Scraping

It was used to collect:

- nearly 1000 images
- 200 images of each flower types across several websites (mostly from flower shops)

The benefits of image scraping:

1. Can automatically scrape a large amount of images for each types of flowers
2. Can be run passively in the background
3. Can download a large amount of flowers quickly when the internet is fast.

Data Preprocessing : Re-Weighting

```
The generated class_weight:
---------------------------
('Babi', 0.5941629955947136)
('Calimero', 2.2860169491525424)
('Chrysanthemum', 0.7856796116504854)
('Hydrangeas', 1.0065298507462686)
('Lisianthus', 0.5872641509433962)
('Pingpong', 1.8732638888888888)
('Rosy', 3.1611328125)
('Tana', 0.9431818181818182)
```

Generating class weight as a priority for each class.

Making the CNN models less bias.

Data Augmentation

Poojary et al.[3] studied the effect of data augmentation on CNN, and found that data augmentation significantly increase test accuracy

Table 3. Effect of data augmentation on performance measures for fine-tuned models

Model	VGG16 fine-tuned model		ResNet50 fine-tuned model	
	With data augmentation	Without data augmentation	With data augmentation	Without data augmentation
Final training accuracy	93.5%	88%	95%	95%
Final validation accuracy	91.5%	83%	93%	84%
Test accuracy	88%	80%	90%	82%

Data Augmentation

Applying random flip as horizontal and random rotation for 20%.

IMAGE Classification

This dataset carries some of the most stressful challenges ranging from dirty data to imbalanced dataset so that using a simple CNN will not be helpful.

Leveraging the great architecture of the strong models like **ResNet50, Xception and DenseNet121** to use them as the base models then fine tuning it to make it work well with the given dataset is a best option.

The use of models like ResNet101 or DenseNet169 will make the train set learn faster than the test set which will cause overfitting. Besides that, it will also be a waste of computing power.

Literature Review and Comparison

To compare our model from task 1 to those from existing literature, we picked a CNN transfer learning model from Narvekar and Rao[4] and a CNN model with stochastic pooling strategy from Prasad *et al.*[5] . Both of these models perform image classification on flower species, similar to ours. As shown above, the models from existing literature vastly outperforms ours in testing accuracy.

Model	Chosen ResNet50 Based CNN	Narvekar and Rao[4]	Prasad *et al.*[5]
Test Accuracy	77%	91%	93.98%

Model Evaluation

	ResNet50 Base Model	DenseNet121 Base Model	Xception Base Model
Accuracy	0.7590	0.7441	0.7305
Precision	0.7778	0.76736	0.7186
Recall	0.7639	0.75401	0.7156
F1-Score	0.76124	0.7472	0.70605

The results above achieved from running 2 epochs with the batch size of 64.

IMAGE Recommendation

The 2 more popular approach for recommendation are content-based and collaborative filtering recommendation.

However, the lack of the detailed information and description about each photo in the dataset prevents us from doing those ways. Therefore the approach that we chose to do is similarity-base recommendation.

Acknowledging that challenge, our approach are using the CNN model to extract features from the images and then using those feature vectors to clustering them in to 5 groups and use that do to recommendation.

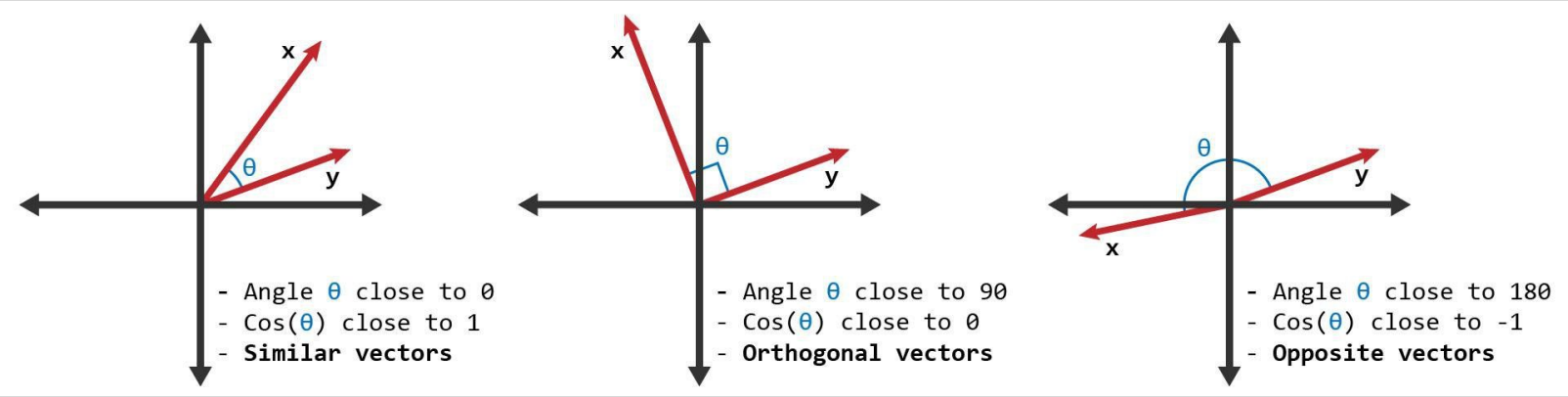
Literature Review

ANOVA cosine similarity recommendation from Sejal *et al.*

1. Users queries through text. Recommendation works using text and visual features integration
2. Compute images' feature vectors -> Compute ANOVA of text terms in description -> Compute cosine similarity and recommend images based on similarity score

Reference: D. Sejal, T. Ganeshsingh, K. R. Venugopal, S. S. Iyengar, and L. M. Patnaik, "Image Recommendation Based on ANOVA Cosine Similarity," Procedia Computer Science, vol. 89, pp. 562–567, Jan. 2016, doi: 10.1016/j.procs.2016.06.091.

Cosine Similarity



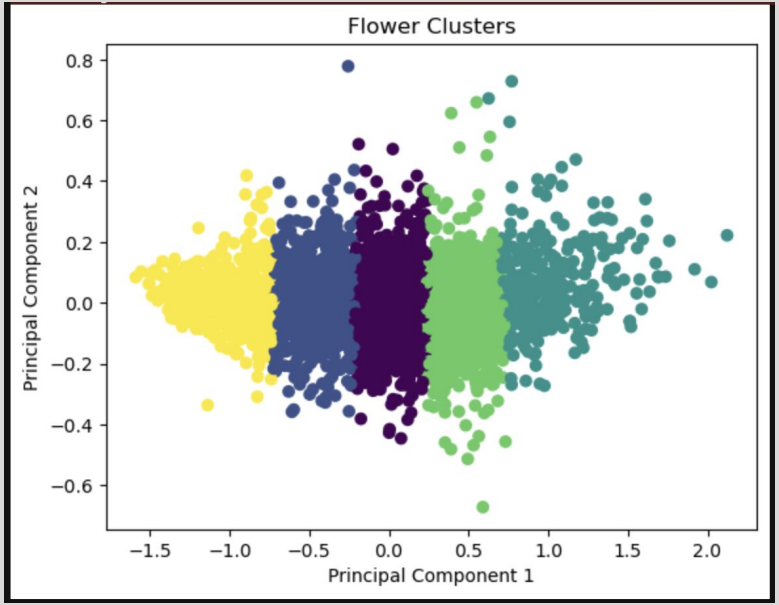
- Close to 1 means highly similar
- Close to -1 means highly dissimilar

Approach

1. Extract feature vectors of image using CNN based on ResNet50
2. Use Principal Component Analysis (PCA) to reduce dimensionality of feature vectors to 2D(Visualization & Computational needs)
3. Apply k-means clustering to group similar feature vectors into clusters for visualization
4. Computer cosine similarity of input images then recommend based on top 10 similarity scores

K-means clustering visualization

K-value = 5



Observations:

1. Clusters are grouped quite close to each other
2. No isolated cluster

=> Data is quite general

Recommendation Result

Input:

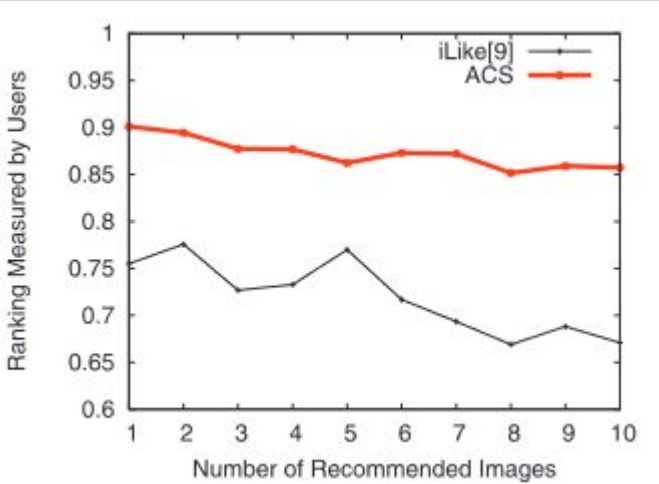
Top 10 recommended images

Model Evaluation

Our self-evaluation & observation

- Recommended images are similar in color
- Recommended images are usually similar in shape, size & orientation

Sejal *et al.* metrics



=> User relevance score of 85%

Summary & Further Improvement

At the end, even though our models for each tasks are not perfect but our group are pretty satisfied with the results. We learned and enjoyed the learning journey we completed after finishing this project.

To further improve this project, here are our thoughts:

- Handle imbalanced dataset using advanced data augmentation or using ensemble Learning methods for deep learning.

References

[1] E. Zhu, Y. Ju, Z. Chen, F. Liu, and X. Fang, “DTOF-ANN: An Artificial Neural Network phishing detection model based on Decision Tree and Optimal Features,” Appl. Soft Comput., vol. 95, p. 106505, Oct. 2020, doi: 10.1016/j.asoc.2020.106505.

[2] Y. Zhang, Z. Jin, F. Liu, W. Zhu, W. Mu, and W. Wang, “ImageDC: Image Data Cleaning Framework Based on Deep Learning,” in 2020 IEEE International Conference on Artificial Intelligence and Information Systems (ICAIIS), Mar. 2020, pp. 748–752. doi: 10.1109/ICAIIS49377.2020.9194803.

[3] R. Poojary, R. Raina, and A. Kumar Mondal, “Effect of data-augmentation on fine-tuned CNN model performance,” IAES Int. J. Artif. Intell. IJ-AI, vol. 10, no. 1, p. 84, Mar. 2021, doi: 10.11591/ijai.v10.i1.pp84-92.

[4] C. Narvekar and M. Rao, “Flower classification using CNN and transfer learning in CNN- Agriculture Perspective,” in 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Dec. 2020, pp. 660–664. doi: 10.1109/ICISS49785.2020.9316030.

[5] M. V.D. Prasad et al., “An efficient classification of flower images with convolutional neural networks,” Int. J. Eng. Technol., vol. 7, no. 1.1, p. 384, Dec. 2017, doi: 10.14419/ijet.v7i1.1.9857.