NATIONAL RESEARCH UNIVERSITY

Machine Learning and Data Mining

Project work

SHOPEE - PRICE MATCH GUARANTEE

Determine if two products are the same by their images

Moscow, 2021

CONTENT

1. TEAM INTRODUCTION
2. PROBLEM SOURCE
3. PROBLEM STATEMENT
4. DATASET
5. METHODS
6. CHALLENGES
7. RESULTS
8. CONCLUSION

TEAM MEMBERS

1. Ishitha Rajapakse
2. Garakhan
3. Sreenjay Sen

PROBLEM SOURCE

Featured Code Competition, KAGGLE

This competition was published on Kaggle in March 8, 2021.

Shopee is the leading e-commerce platform in Southeast Asia and Taiwan. Customers appreciate its easy, secure, and fast online shopping experience tailored to their region. The company also provides strong payment and logistical support along with a 'Lowest Price Guaranteed' feature on thousands of Shopee's listed products.

In this competition, we will apply your machine learning skills to build a model that predicts which items are the same products.

Problem Statement

- Companies use a variety of methods to assure customers that their products are the cheapest.
- Deep learning and traditional machine learning analyzes image and text information to compare similarity.
- In this problem, we try to find images, that are posted by different users, belonging to same products using machine learning algorithms and similarity metrics.

DATASET

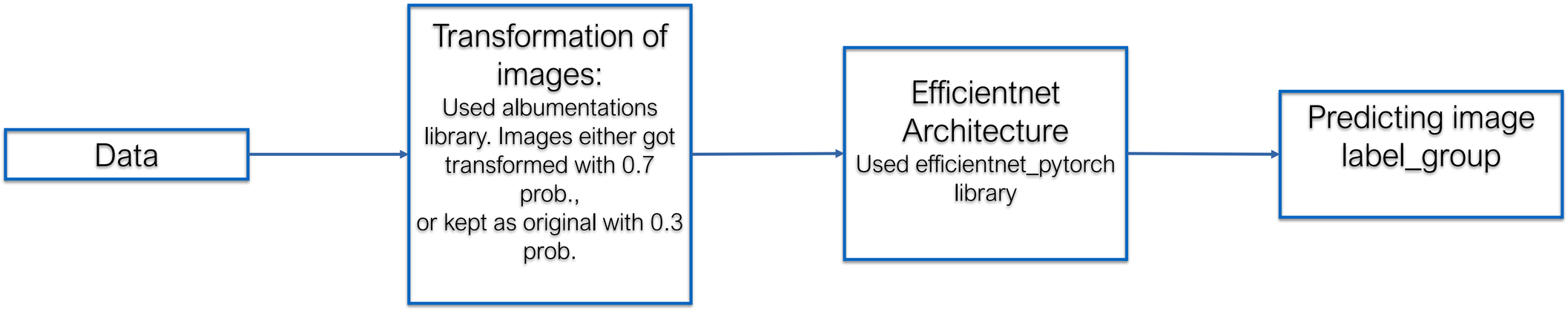
- **dataset size** : 1.92GB
- **trainset**: 34250 images
- **testset**: 3 rows available; 70000 images expected in hidden test set

- *posting_id* -the ID code for the posting.
- *image* - the image id/md5sum.
- *image_phash* - a perceptual hash of the image.
- *title* - the product description for the posting.
- *label_group* - ID code for all postings that map to the same product. Not provided for the test set.

	posting_id	image	image_phash	title	label_group
0	train_129225211	1.jpg	94974f937d4c2433	Paper Bag Victoria Secret	4185
1	train_3386243561	2.jpg	af3f9460c2838f0f	Double Tape 3M VHB 12 mm x 4,5 m ORIGINAL / DO...	2044
2	train_2288590299	3.jpg	b94cb00ed3e50f78	Maling TTS Canned Pork Luncheon Meat 397 gr	2368
3	train_2406599165	4.jpg	8514fc58eafea283	Daster Batik Lengan pendek - Motif Acak / Camp...	10170
4	train_3369186413	5.jpg	a6f319f924ad708c	Nescafe \xc3\x89clair Latte 220ml	5887

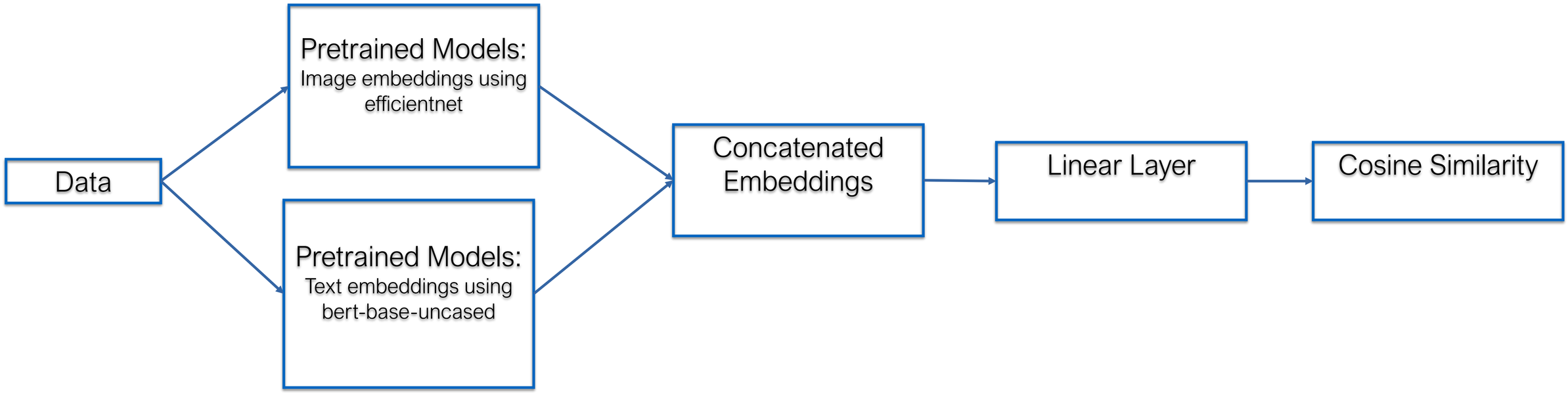
1ST METHOD: TRAINING OUR OWN MODEL

Classification: Classes are the label_group column on dataset



2ND METHOD: USING PRETRAINED MODEL

Similarity: Finding similar embeddings using cosine similarity



CHALLENGES

Problems with the first approach

- Number of dataset ~30K
- Number of classes ~11K
- Which means that we have data shortage.
- Tried data augmentation, but training session was too slow; because efficientnet model was too big.

RESULTS

1st Method	Still in progress*
2nd Method	0.5-0.6 F1 score**

* Training was too slow, and accuracy did not improve much < 0.1 %.
We will try to improve the model

** The submission file to competition is assessed using F1 score.

CONCLUSION

NATIONAL RESEARCH
UNIVERSITY

Thank you.