AUTOMATIC
PRODUCT CLASSIFICATION

Project

place de marché

Automatically classify consumer goods based on product description and image

Challenges

- Avoid manual attribution of the category
- Improve user experience
- Prepare for the future scaling up

MISSION

Create a Proof of Concept of the unsupervised classification based on product description and image

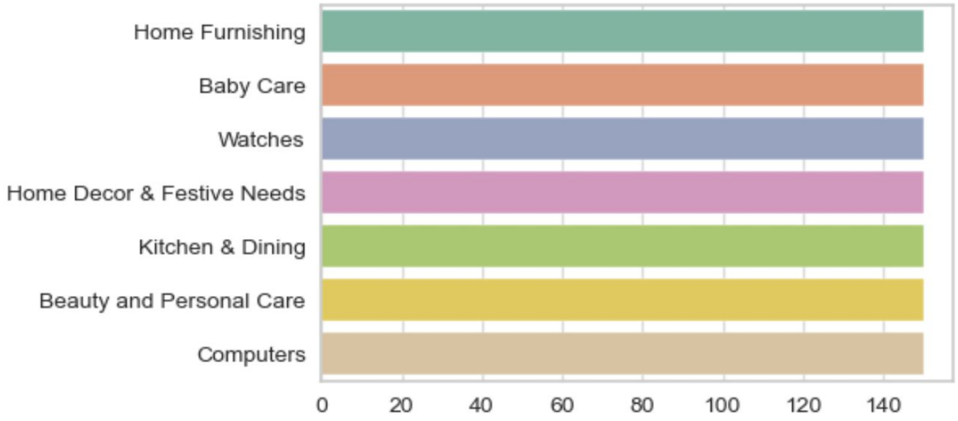
Perform a supervised product classification based on image

Test product extraction from an API

DATA

Quick overview

- 1050 products
- 15 features
- 1050 image files
- 2% null values
- 7 main categories
- 150 items per category
- free of intellectual property issues



	product_name	product_category_tree	image	description
175	Prithish I'm A Mom. What'S Your Superpower? Red with Blue Bubbles Double Coloured Ceramic Mug	["Kitchen & Dining >> Coffee Mugs >> Prithish Coffee Mugs"]	22ea3b69860bc9a5e0c69b87db2dcdab.jpg	Buy Prithish I'm A Mom. What'S Your Superpower? Red with Blue Bubbles Double Coloured Ceramic Mu...
804	Oxyglow Fruit Massage Cream With Vitamin-E & Lip Balm	["Beauty and Personal Care >> Combos and Kits >> Oxyglow Combos and Kits"]	648b52926285610670a756923b8a6f99.jpg	Buy Oxyglow Fruit Massage Cream With Vitamin-E & Lip Balm for Rs.362 online. Oxyglow Fruit Massa...
85	Smart Wall Guru Medium Adhesive Sticker	["Home Decor & Festive Needs >> Wall Decor & Clocks >> Wall Decals & Stickers >> Smart Wall Guru...	61c0139f3d5c97f1615797dbf046cc0c.jpg	Key Features of Smart Wall Guru Medium Adhesive Sticker We Showcase A Variety Of Wall Decal Cate...
366	Beverly Hills Polo Club RED Gift Set Combo Set	["Beauty and Personal Care >> Fragrances >> Deodorants >> Combos >> Beverly Hills Polo Club Comb...	47d7792e50e69b048b1f17176f170141.jpg	Flipkart.com: Buy Beverly Hills Polo Club RED Gift Set Combo Set online only for Rs. 399 from F...
783	Funcart Party Time - Food Grade Plastic	["Baby Care >> Feeding & Nursing >> Utensils >> Baby Spoons & Forks >> Funcart Baby Spoons & For...	9a0413117aa91c668ba5944770347cc5.jpg	Flipkart.com is offering Small Funcart Party Time - Food Grade Plastic for only for Rs 154 | Fl...
369	Tenda 3G300M	["Computers >> Network Components >> Routers >> Tenda Routers"]	53c4f1e5cb1767f1a6ba05d32dfaf107.jpg	Buy Tenda 3G300M only for Rs. 3000 from Flipkart.com. Only Genuine Products. 30 Day Replacement ...
392	Wild Stone Drift,Wave Combo Set	["Beauty and Personal Care >> Fragrances >> Deodorants >> Combos >> Wild Stone Combos"]	d376da2235096f74600d842b6f5f2d1a.jpg	Flipkart.com: Buy Wild Stone Drift,Wave Combo Set online only for Rs. 355 from Flipkart.com. Onl...
363	Aero Care Ice & Blue 4 Men Combo Set	["Beauty and Personal Care >> Fragrances >> Deodorants >> Combos >> Aero Care Combos"]	8932c9a89b487bf9f2b3fce06c7c6553.jpg	Flipkart.com: Buy Aero Care Ice & Blue 4 Men Combo Set online only for Rs. 165 from Flipkart.com...
820	Mxofere Combo Rose Soap And Aloevera Lemon Facewash Kit	["Beauty and Personal Care >> Combos and Kits >> Mxofere Combos and Kits"]	00d84a518e0550612fcfcba3b02b6255.jpg	Buy Mxofere Combo Rose Soap And Aloevera Lemon Facewash Kit for Rs.285 online. Mxofere Combo Ros...
146	Printland PMBA1044 Ceramic Mug	["Kitchen & Dining >> Coffee Mugs >> Printland Coffee Mugs"]	63bc43c0000e71eb11bfafd2f46214a4.jpg	Printland PMBA1044 Ceramic Mug (350 ml)\r\n Price: Rs. 299\r\n\t\t\t\t\r...

FEATURE ENGINEERING

FEATURE	INITIAL	TRANSFORMATION
CATEGORY	‘product_category_tree’	- Extract first level category name - Encode with **LabelEncoder**
DESCRIPTION	‘product_name’ ‘description’	Merge ‘product_name’ and ‘description’ fields
IMAGE	‘image’	Extract grayscale matrix from the image file

DESCRIPTION BASED

UNSUPERVISED CLASSIFICATION

TEXT PREPROCESSING

01

02

TECHNICAL

- lowercase
- remove stop-words
- remove punctuation
- keep alpha characters
- lemmatize
- remove 1 letter tokens

SEMANTIC

- remove the most frequent generic tokens
- remove rare tokens (< 3 times)

FEATURE EXTRACTION & CLASSIFICATION PROCESS

STEP 1

VECTORIZATION

Bag of Words and Word embeddings

STEP 2

DIMENSIONALITY REDUCTION

PCA & T-SNE to reduce to 2 features

STEP 3

GRAPHIC ANALYSIS

Plot with true categories

STEP 4

EVALUATION

K-Means clustering and ARI score

FEATURE EXTRACTION APPROACHES

BAG OF WORDS

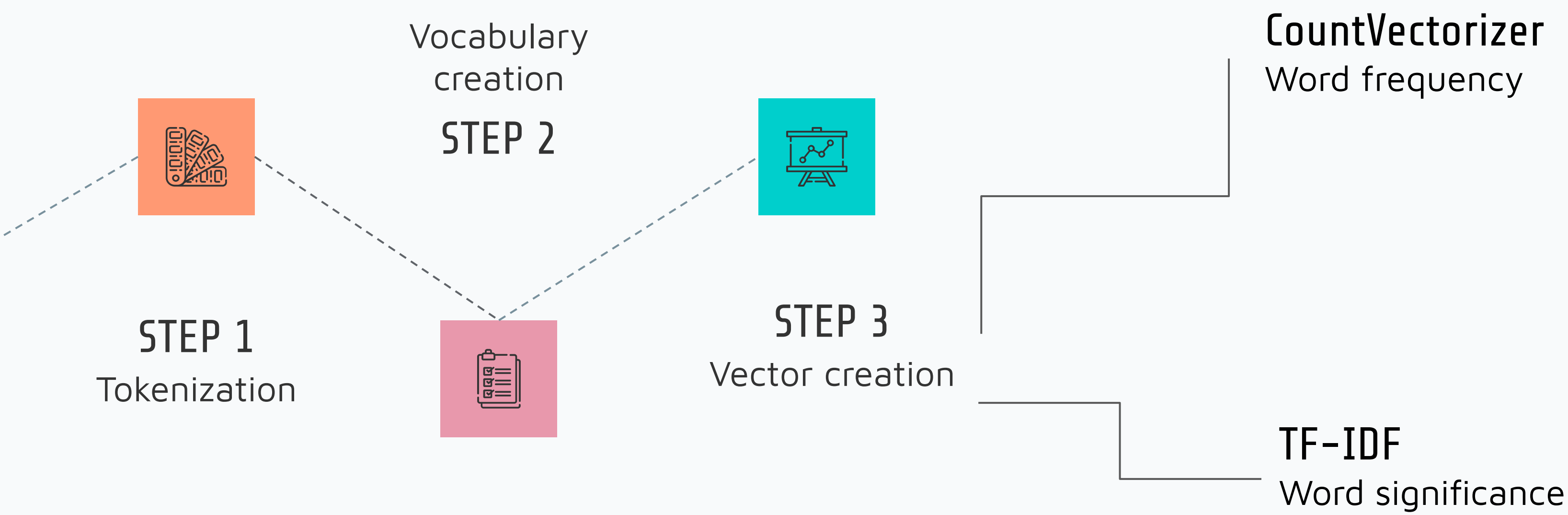
- CountVectorizer
- TF-IDF

WORD EMBEDDINGS

- Word2Vec
- BERT
- USE

Bag Of Words	Word Embeddings
Intuitive and easy to implement	Complex, heavy to implement
Based on word frequency	Based on neural networks
Dimensionality, sparse matrix issues	Control over dimensions
Do not consider context, order	Awareness of the context

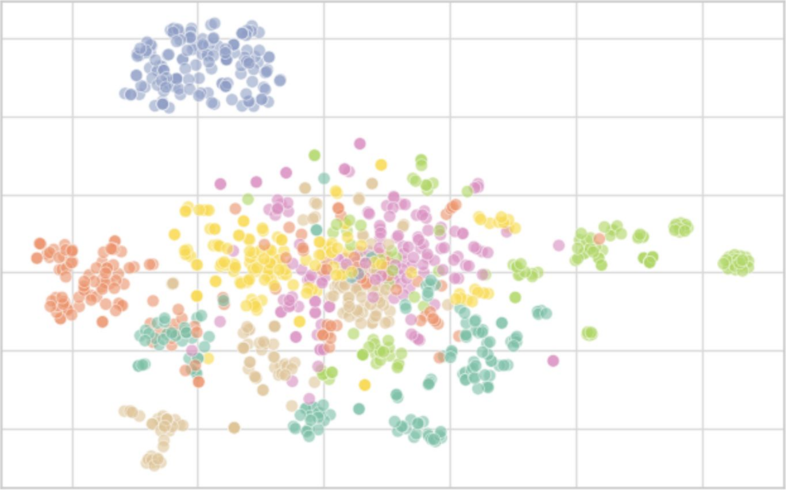
BAG OF WORDS



STEP 1
Tokenization
Vocabulary creation
STEP 2
STEP 3
Vector creation
CountVectorizer
Word frequency
TF-IDF
Word significance

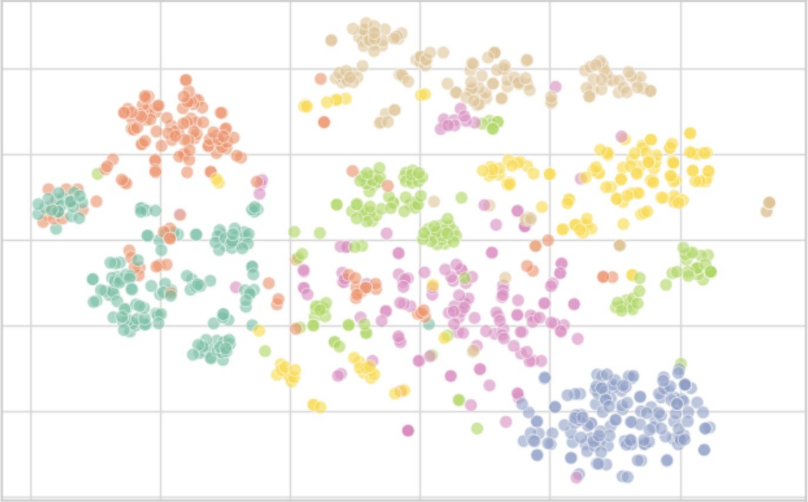
BAG OF WORDS results

CountVectorizer



ARI: **0.39** execution time: **0.03 sec**

TF-IDF



ARI: **0.51** execution time: **0.03 sec**

WORD EMBEDDINGS

Word2Vec	BERT	USE
• Simple single layered neural network • Word embedding model • Single vector per word • Trained on the given corpus	• 12 layer deep neural network • Transformer-based architecture • Language representation model (word + sentence) • Context-based different vectors per word • Different pre-trained models available	• Transformer-based or Deep Averaging Network • Sentence embedding model • Trained on Wikipedia and web content
~ 8 sec	~ 233 sec	< 1 sec
0.45 score	0.31 score	0.66 score

KEY POINTS

TF-IDF (0.51)

- Simple, efficient, performant
- Can be improved by using n-grams, tuning parameters
- Sparse matrix that will get larger

USE (0.66)

- Straightforward to implement
- Pre-trained model that is well suited for our dataset

IMAGE BASED

UNSUPERVISED CLASSIFICATION

IMAGES

Home Furnishing





Kitchen & Dining




Watches




FEATURE EXTRACTION & CLASSIFICATION PROCESS

STEP 1

FEATURE EXTRACTION

SIFT, ORB, CNN

STEP 2

DIMENSIONALITY REDUCTION

PCA & T-SNE to reduce to 2 features

STEP 3

GRAPHIC ANALYSIS

Plot with true categories

STEP 4

EVALUATION

K-Means clustering and ARI score

FEATURE EXTRACTION APPROACHES

- ORB
- SIFT

Convolutional Neural Networks

- VGG16
- ResNet50

ORB & SIFT FEATURE EXTRACTION

STEP 1

Preprocess image (apply equalizer on grayscale images)

STEP 2

Calculate **descriptors** for all images

STEP 3

Perform descriptors clustering

STEP 4

Build descriptors histogram for each image

KEYPOINTS & DESCRIPTORS DETECTION

ORB	SIFT
	
~ 24 sec	~ 177 sec
0.035 score	0.045 score

Image features should be:

- Repeatable (robust to scale, rotation, noise, etc)
- Distinctive
- Local

TRANSFER LEARNING: BENEFITS

TIME
Saving training time

PERFORMANCE
Better performance of neural networks

SMALL DATA
Not needing a lot of data

ML XP
Easy to implement and customize

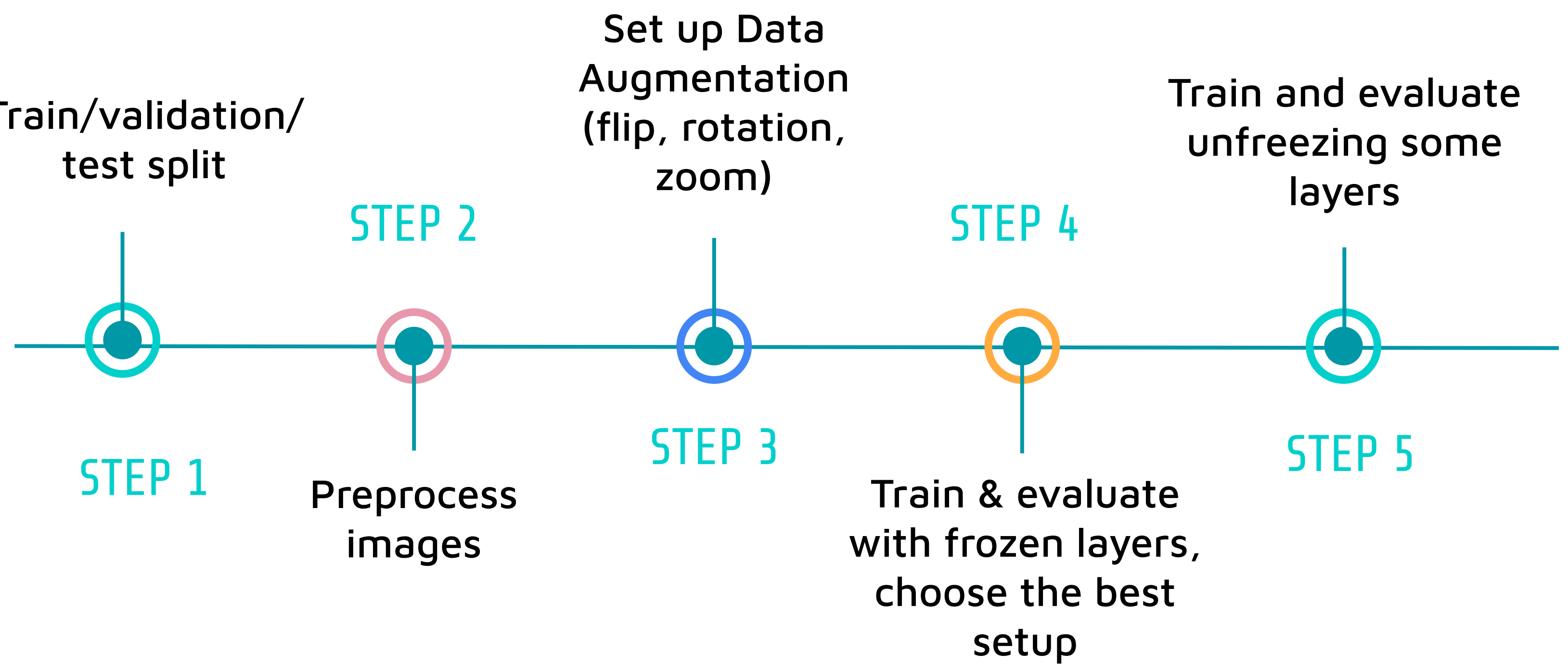
FEATURE EXTRACTION WITH CNN

VGG16	ResNet50
• 16 layers • trained on Imagenet • big model size due to its depth and the number of fully-connected nodes	• 50 layers • trained on Imagenet • smaller model size due to the usage of global average pooling
~ 99 sec	~ 65 sec
0.43 score	0.55 score

IMAGE-BASED

SUPERVISED CLASSIFICATION

CLASSIFICATION WITH RESNET50 MODEL WORKFLOW

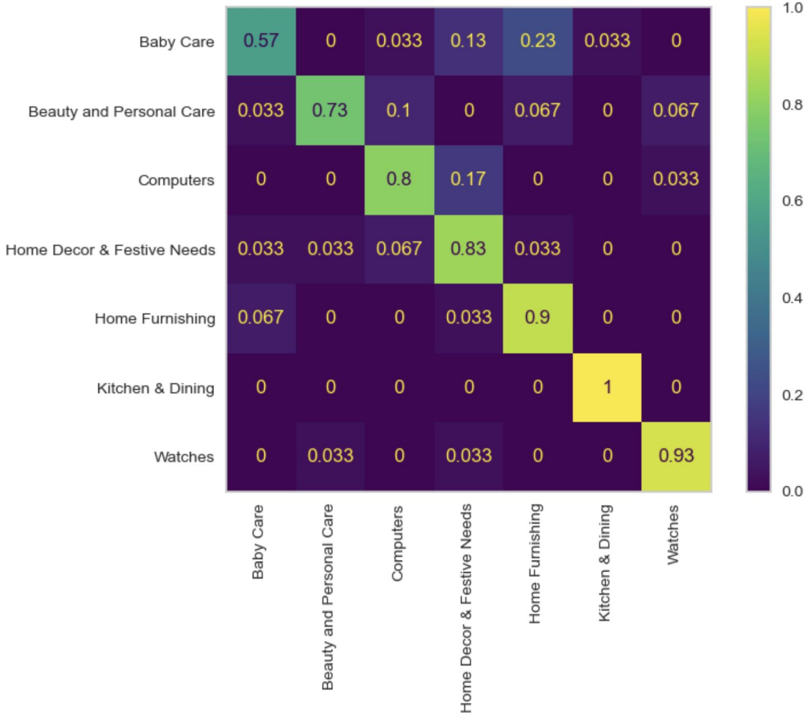


TESTED MODEL SETUPS

Dense layer 256 nodes	2 dense layers	Dense layer 512 nodes
Data augmentation -> Base ResNet50 model -> Global Average Pooling		
Dense layer 256 nodes (‘relu’) Dropout (0.5)	**Dense layer 256 nodes (‘relu’)** Dropout (0.2) **Dense layer 128 nodes (‘relu’)** Dropout (0.5)	**Dense layer 512 nodes (‘relu’)** Dropout (0.5)
Dense layer 7 nodes (‘softmax’)		
val accuracy: 0.84 val loss: 0.57 test accuracy: 0.81 test loss: 0.6	val accuracy: 0.85 val loss: 0.51 test accuracy: 0.82 test loss: 0.6	val accuracy: 0.83 val loss: 0.57 test accuracy: 0.81 test loss: 0.59

EVALUATION

Recall



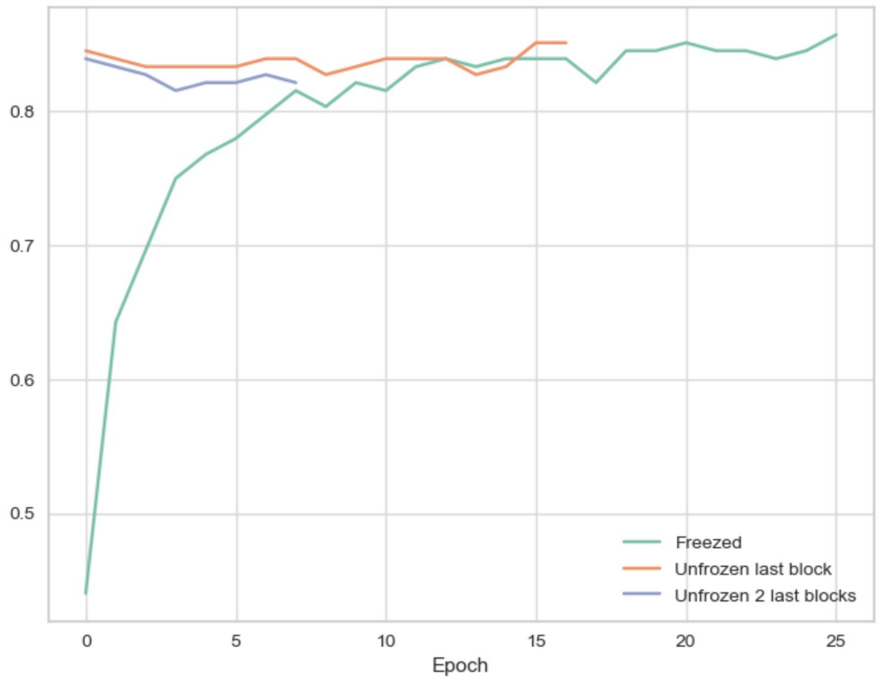
Precision

true labels

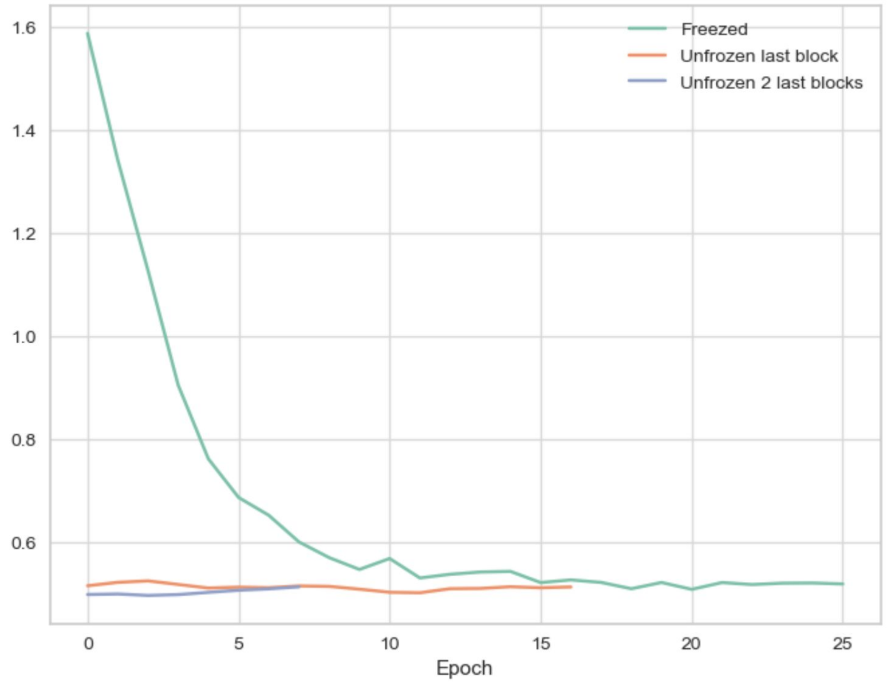
predicted

FINE-TUNING

Validation accuracy



Validation loss



DATA EXTRACTION FROM AN API

API INFORMATION

DATA SOURCE	Edamam
DATA PROVIDER	RapidApi
ENDPOINT	/api/food-database/v2/parser
QUERY	{"ingr": "champagne"}

EXTRACTED DATA

	foodId	label	category	foodContentsLabel	image
0	food_a656mk2a5dmqb2adiamu6beihduu	Champagne	Generic foods	NaN	https://www.edamam.com/food-img/a71/a718cf3c52...
1	food_b753ithamdb8psbt0w2k9aquo06c	Champagne Vinaigrette, Champagne	Packaged foods	OLIVE OIL; BALSAMIC VINEGAR; CHAMPAGNE VINEGAR...	NaN
2	food_b3dyababjo54xobm6r8jzbghjgqe	Champagne Vinaigrette, Champagne	Packaged foods	INGREDIENTS: WATER; CANOLA OIL; CHAMPAGNE VINE...	https://www.edamam.com/food-img/d88/d88b64d973...
3	food_a9e0ghsamvoc45bwa2ybsa3gken9	Champagne Vinaigrette, Champagne	Packaged foods	CANOLA AND SOYBEAN OIL; WHITE WINE (CONTAINS S...	NaN
4	food_an4jjueaucpus2a3u1ni8auhe7q9	Champagne Vinaigrette, Champagne	Packaged foods	WATER; CANOLA AND SOYBEAN OIL; WHITE WINE (CON...	NaN
5	food_bmu5dmkazwuvpaa5prh1daa8jxs0	Champagne Dressing, Champagne	Packaged foods	SOYBEAN OIL; WHITE WINE (PRESERVED WITH SULFIT...	https://www.edamam.com/food-img/ab2/ab2459fc2a...
6	food_alpl44taoyv11ra0lic1qa8xculi	Champagne Buttercream	Generic meals	sugar; butter; shortening; vanilla; champagne;...	NaN
7	food_byap67hab6evc3a0f9w1oag3s0qf	Champagne Sorbet	Generic meals	Sugar; Lemon juice; brandy; Champagne; Peach	NaN
8	food_am5egz6aq3fpjlaf8xpkdbc2asis	Champagne Truffles	Generic meals	butter; cocoa; sweetened condensed milk; vanil...	NaN
9	food_bcz8rhiajk1fuva0vkfmeakbouc0	Champagne Vinaigrette	Generic meals	champagne vinegar; olive oil; Dijon mustard; s...	NaN

GDPR COMPLIANCE

1	PURPOSE LIMITATION	The data must be processed only for the given purpose
2	DATA RELEVANCE	Gather only the necessary data
3	STORAGE LIMITATIONS	The data must not be stored longer than intended
4	INTEGRITY, CONFIDENTIALITY	The data must be processed only by the people who have access to it
5	LAWFULNESS, FAIRNESS, TRANSPARENCY	Valid legal basis, best person's interest, clear communication

TO GO FURTHER

- Continue customizing Data Augmentation techniques
- Train on text and image features together
- Try out other network architectures or different models
- Change the target from the main category to a subcategory

THANKS!