DSPT
DATA SCIENCE PORTUGAL

SHOPAI.CO

U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

VISUAL SIMILARITY WITH DEEP TRIPLET QUANTIZATION: APPLICATION TO THE FASHION INDUSTRY

PEDRO ESMERIZ| JOÃO GAMA

SCHOOL OF ECONOMICS AND MANAGEMENT OF THE UNIVERSITY OF PORTO

16TH MARCH 2023

ation: Confidential

Agenda

Classification: Confidential

Ecommerce platforms are targeting innovative search engines to enhance customer experience

Visual Search

Customer

- How to describe this dress?
- How to find similar products?

Brands

- High bounce rate
- Low conversion rate
- Human error on recommendations

Classification: Confidential

Deep learning based visual search engines have achieved high performance but at a high infrastructure cost

Research Questions

1. What are the most appropriate CNN architectures for image feature extraction?

2. What is the most effective strategy for similarity search?

3. What is the most effective dimensionality reduction process to apply?

Classification: Confidential

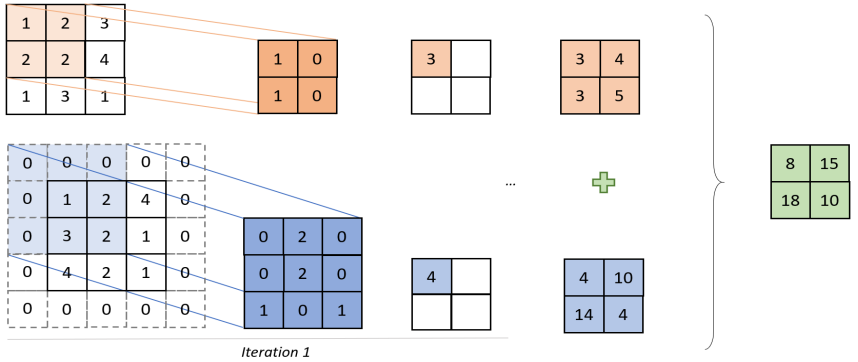
Agenda

Classification: Confidential

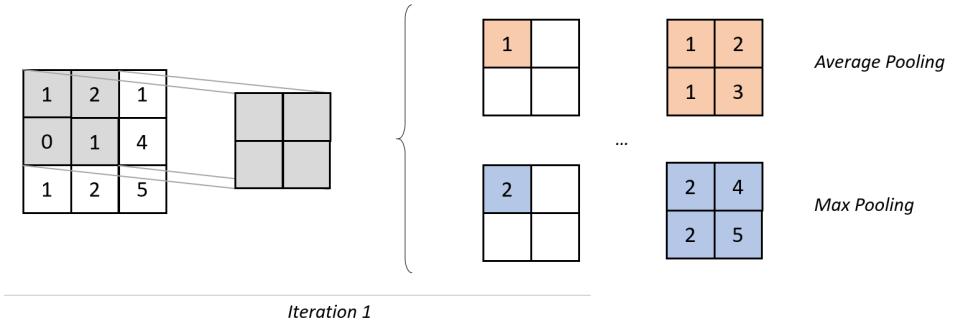
Convolutional neural networks make use of additional modules specifically targeted towards image processing

CNN Components

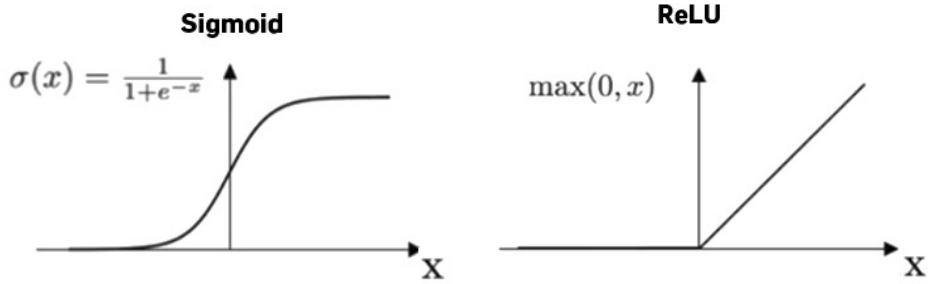
Convolution



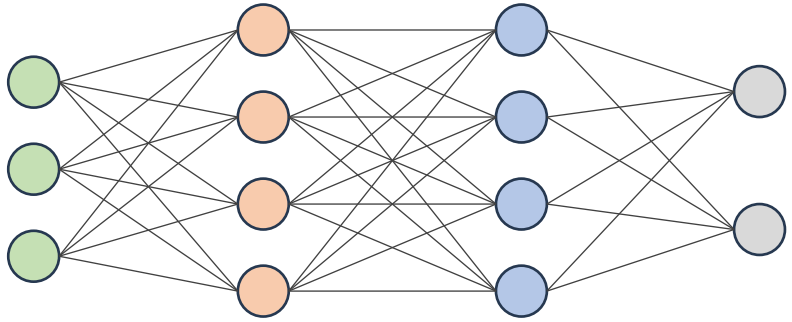
Pooling



Activation Functions



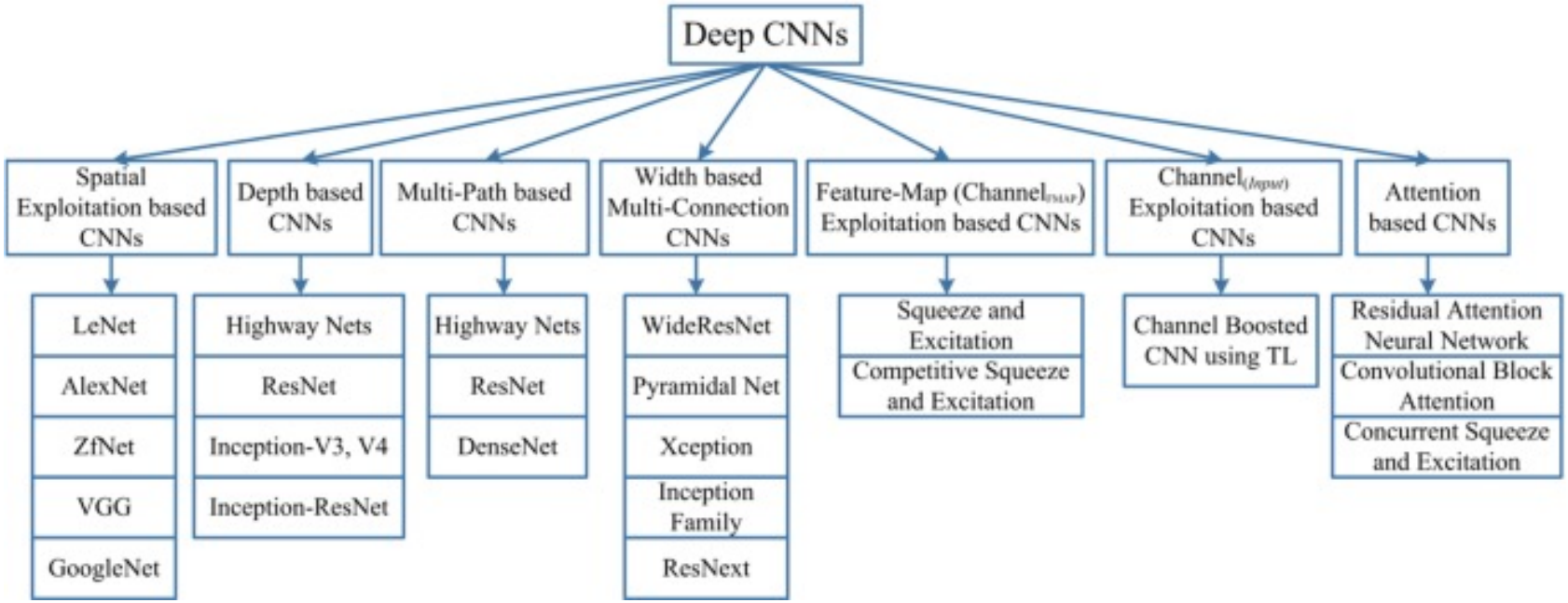
Fully Connected Layers



Classification: Confidential

Over the years, several architectures have been developed to improve the analysis of both broad and fine image details

CNN Architectures

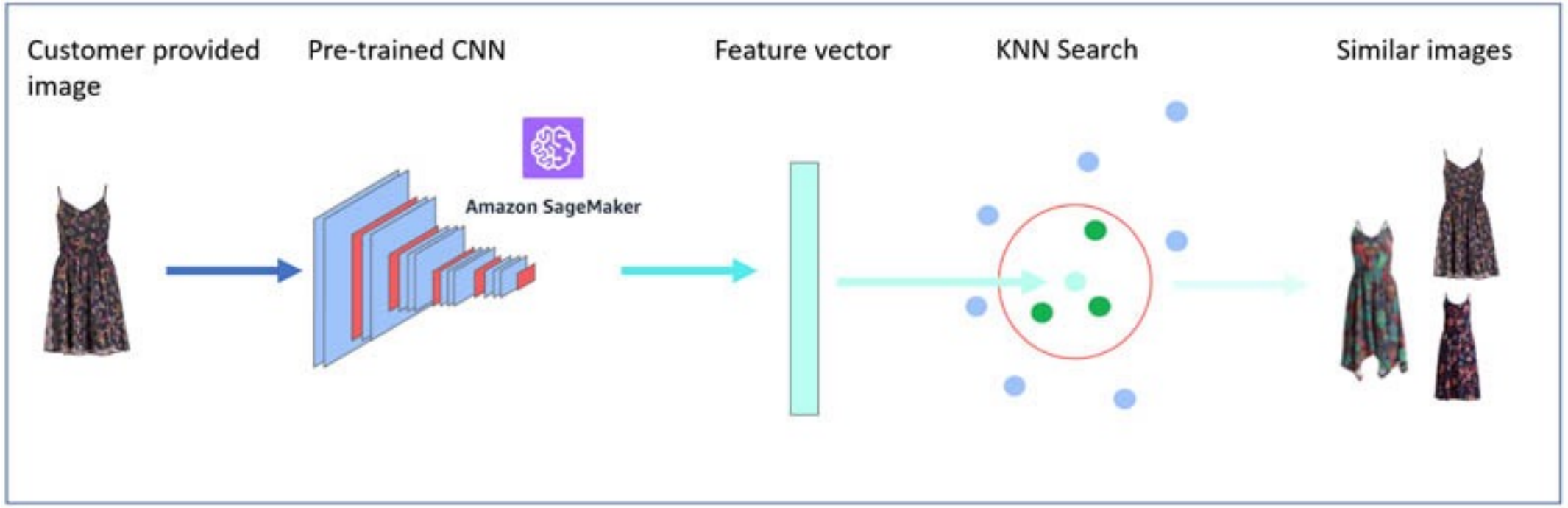


Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2019). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. CoRR, abs/1901.0. http://arxiv.org/abs/1901.06032

Classification: Confidential

Using a pre-trained CNN, we assume that objects of the same category should have small vector distances

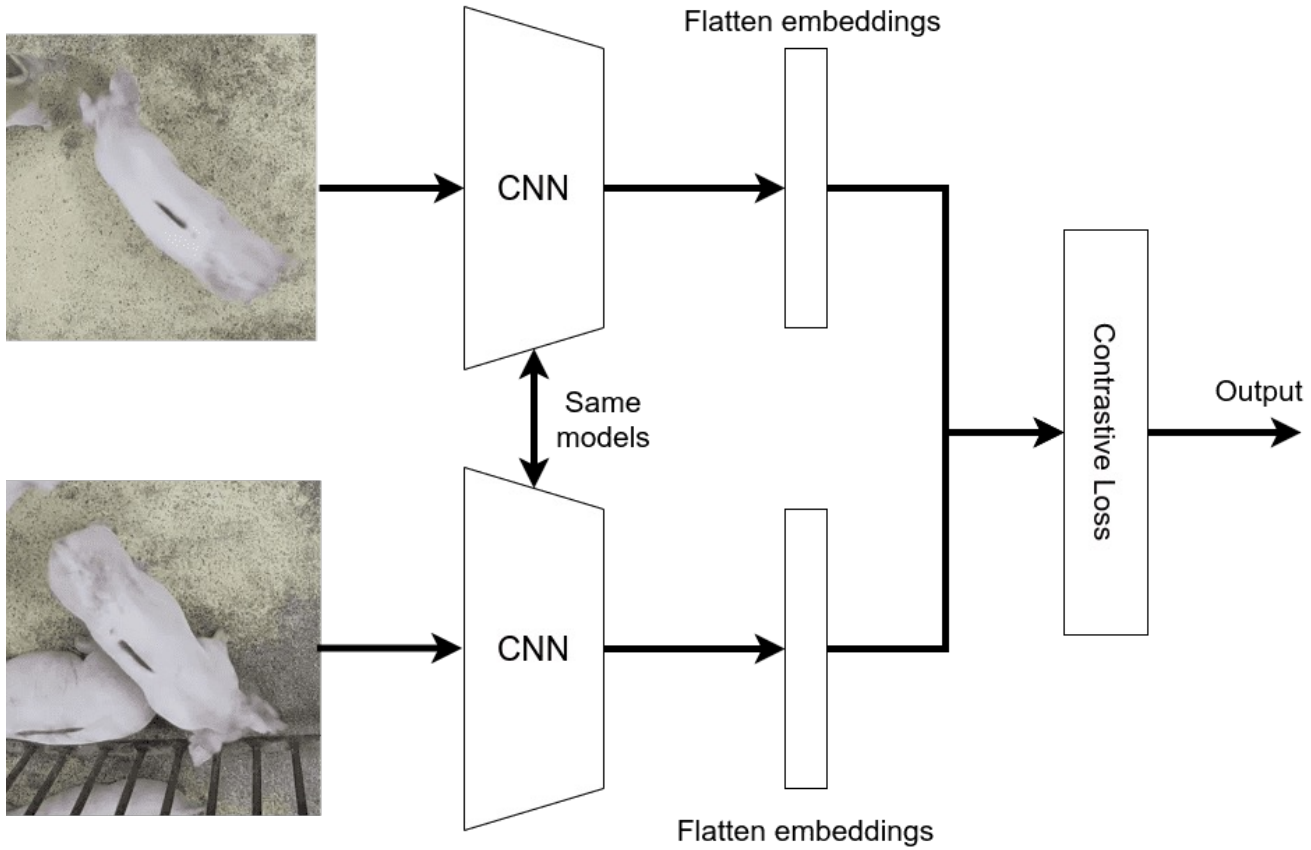
Similarity Search



Classification: Confidential

Siamese networks allow the input pairs and trains the model to produce similar values for positive pairs and vice-versa

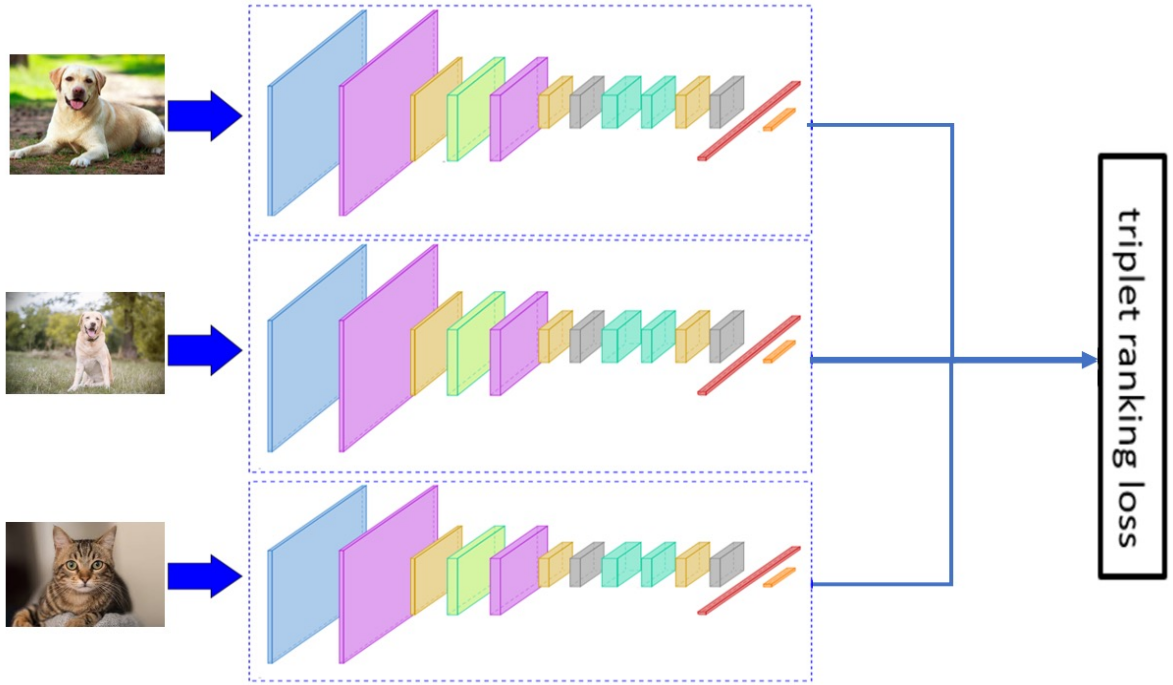
Similarity Search



Classification: Confidential

Triplet based similarity networks achieve better results due to their capability to develop relative similarity

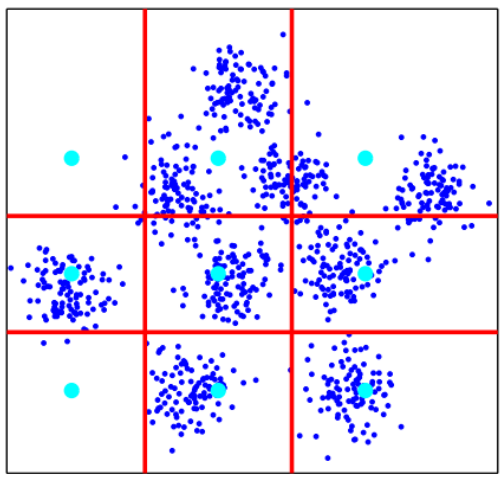
Similarity Search



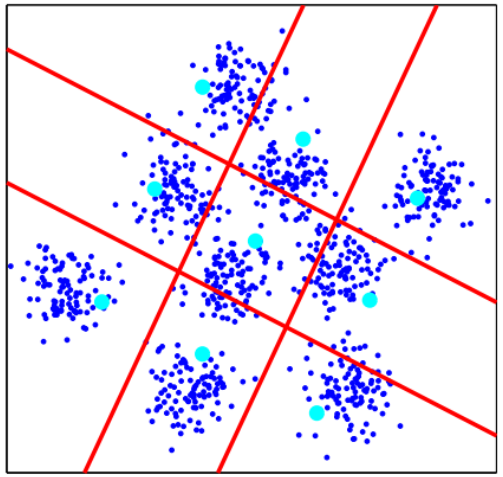
Classification: Confidential

Traditional and isolated learning hash functions perform worse than integrating them in the training of a CNN

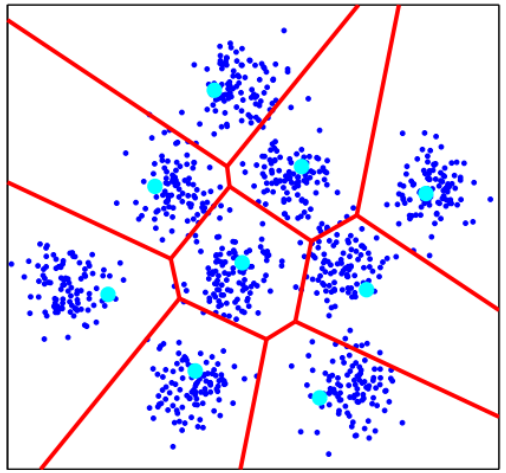
Dimensionality Reduction – Deep Quantization



Product Quantization



Cartesian k-means



Composite Quantization

Wang, Jingdong, Zhang, T., Song, J., Sebe, N., & Shen, H. T. (2016). A Survey on Learning to Hash. CoRR, abs/1606.0. http://arxiv.org/abs/1606.00185

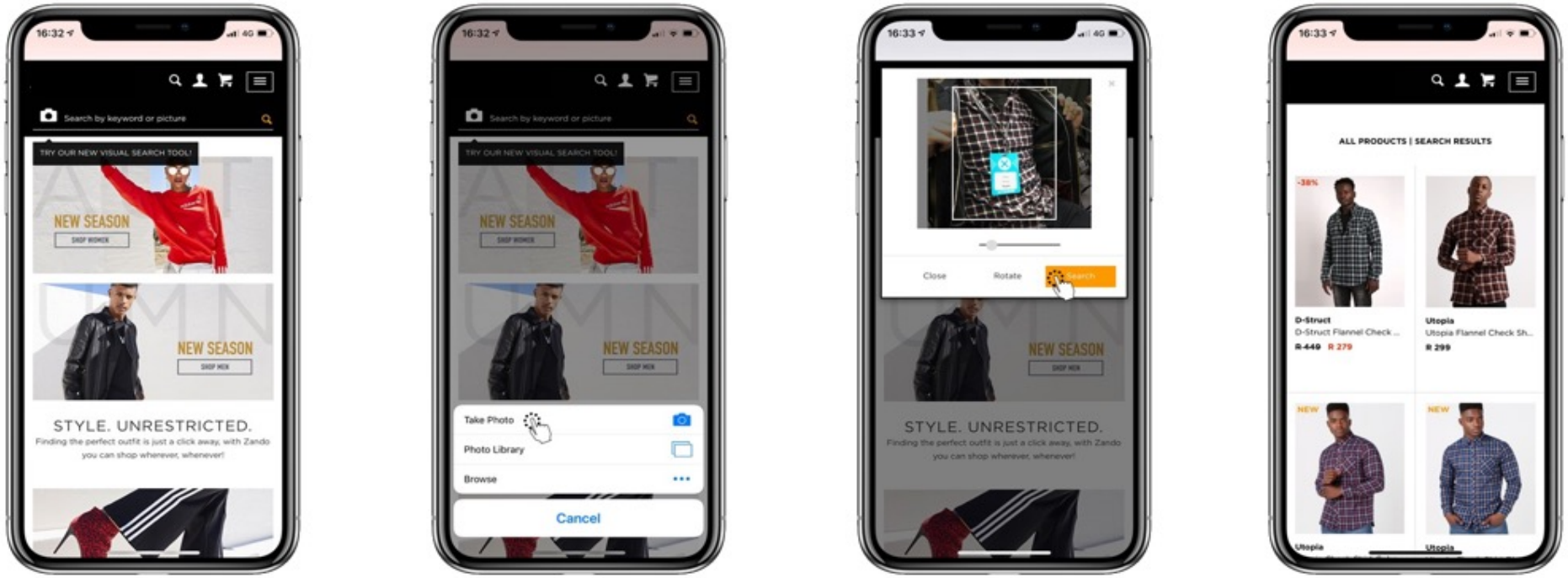
Classification: Confidential

Agenda

Classification: Confidential

Online recommendations are provided through an interface that processes a user submitted image in real-time

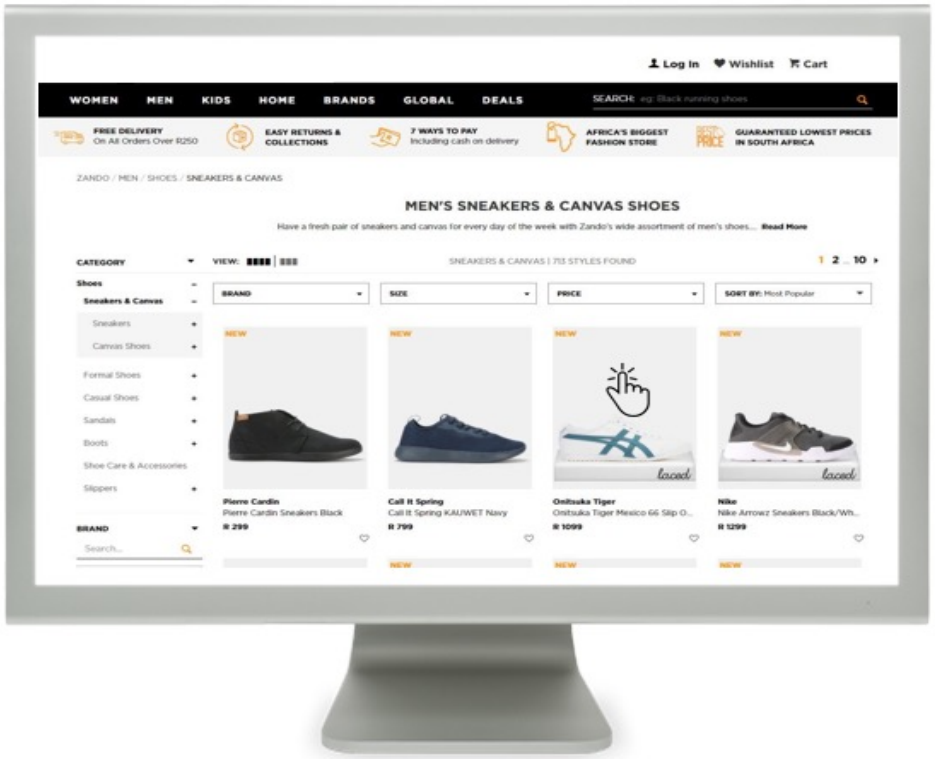
Online Recommendations

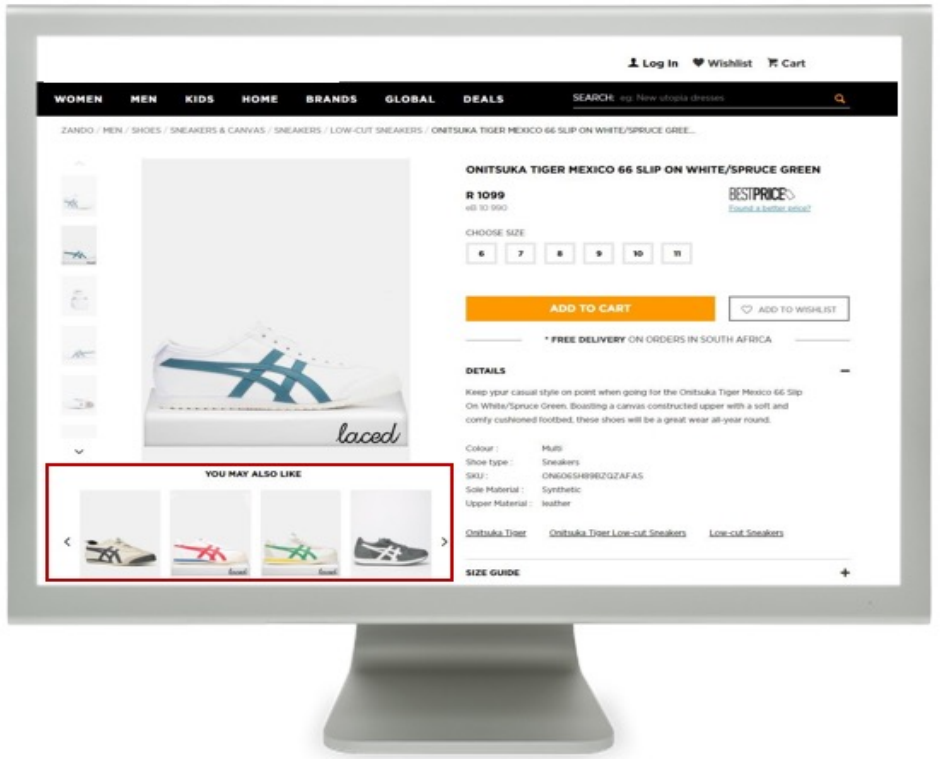


Classification: Confidential

Offline recommendations can be provided by displaying a banner of similar items on a product page

Offline Recommendations





Classification: Confidential

Agenda

Classification: Confidential

To create the training dataset, several in shop and street views of different items were collected and hierarchized

Triplet Generation

Classification: Confidential

The core CNN was created from pretrained models available on Keras and selected based on external benchmarks[1]

CNN Selection

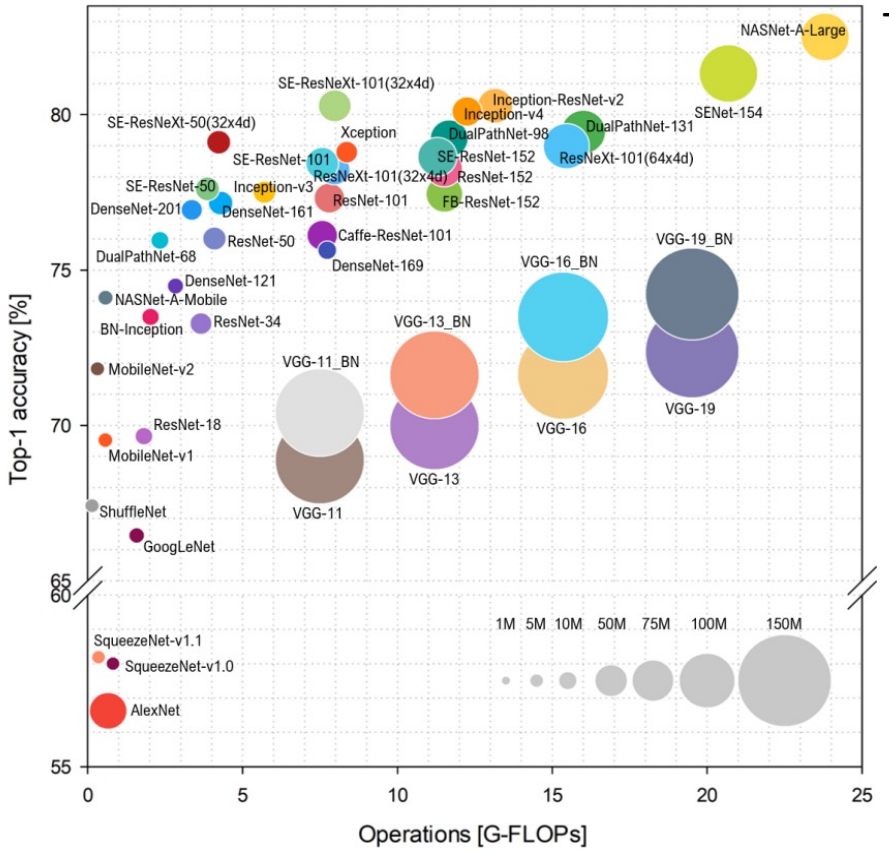


Table 1 - Benchmark times for different batch sizes on Nvidia Jetson TX1 (ms)

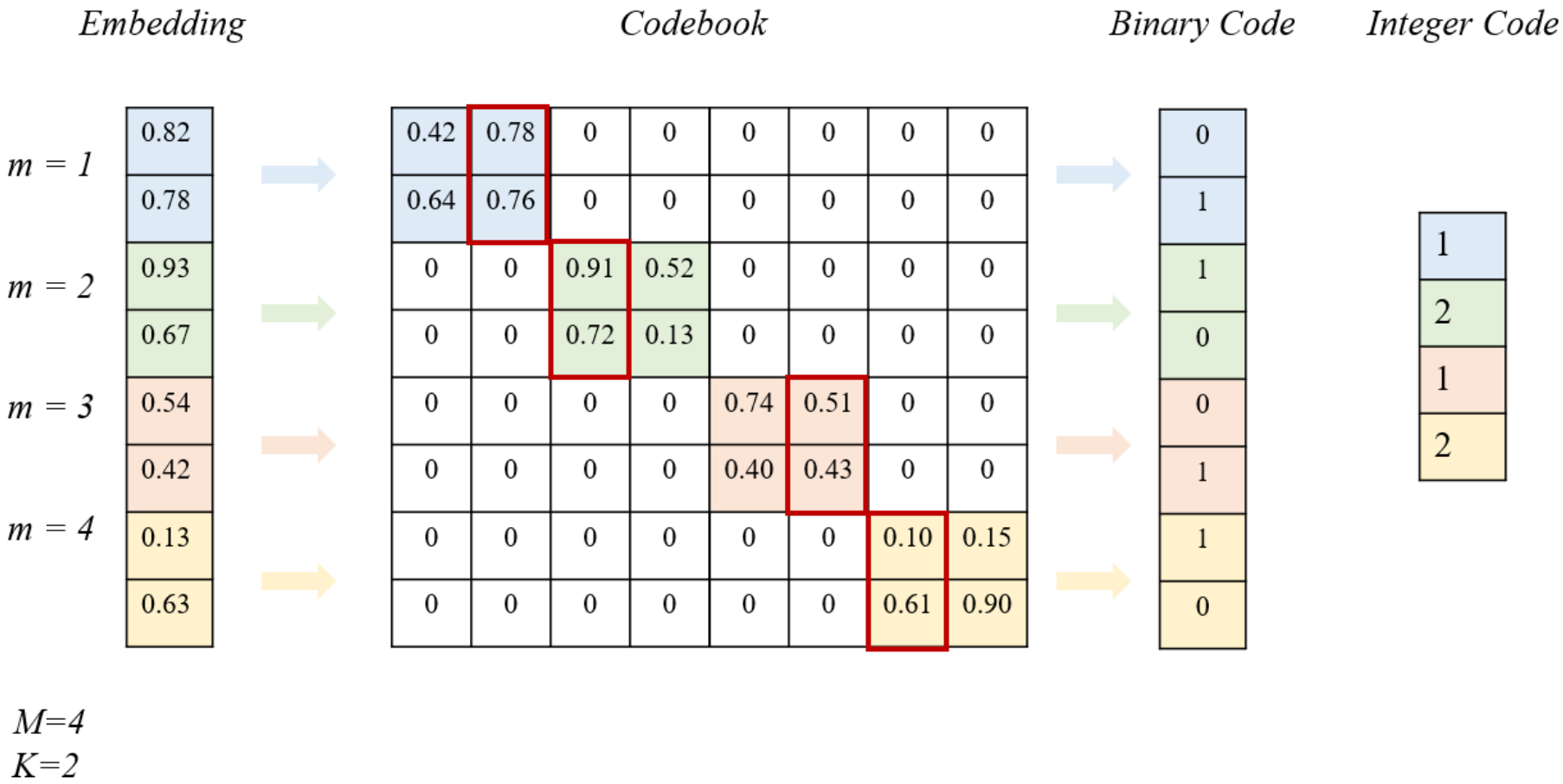
DNN	1	2	4	8	16	32	64
DenseNet-169	137.96	130.27	110.82	100.56	92.97	88.94	
DenseNet-201	84.57	61.71	62.62	53.73	49.28	46.26	
Inception-ResNet-v2	198.95	141.29	127.97	130.25	117.99	116.47	
Inception-v3	79.39	59.04	56.46	51.79	47.6	46.85	
MobileNet-v1	15.06	11.94	11.34	11.03	10.82	10.58	10.55
MobileNet-v2	20.51	14.58	13.67	13.56	13.18	13.1	12.72
NASNet-A-Large	437.2	399.99	385.75	383.55	389.67		
NASNet-A-Mobile	133.87	62.91	33.72	30.62	29.72	28.92	28.55
ResNet-101	84.52	77.9	71.23	67.14	58.11		
ResNet-152	124.67	113.65	101.41	96.76	82.35		
ResNet-50	53.09	44.84	41.2	38.79	35.72		
Xception	98.96	93.4	90.49	87.65	86.89		

1 Bianco, S., Cadène, R., Celona, L., & Napoletano, P. (2018). Benchmark Analysis of Representative Deep Neural Network Architectures. CoRR, abs/1810.0. http://arxiv.org/abs/1810.00736

Classification: Confidential

A codebook matrix is trained to transform the embeddings from the previous layer into compact binary codes

Quantization Process



Classification: Confidential

The objective is to improve the similarity analysis (L) while improving the quantization error and efficiency (Q[1])

Objective Function

Triplet loss

$$L = \sum_{i=1}^{N_t} L_i = \sum_{i=1}^{N_t} \max\left(0, \delta - \|z_i^a - z_i^n\|^2 + \|z_i^a - z_i^p\|^2\right)$$

Quantization Error

$$Q_1 = \sum_{i=1}^{N_t} \sum_{*\in\{a,p,n\}} \left\| z_i^* - \sum_{m=1}^{M} C_m b_{mi}^* \right\|^2$$

Weak-Orthogonality

$$Q_2 = \gamma \sum_{m=1}^{M} \sum_{m'=1}^{N} \|C_m^T C_{m'} - I\|^2$$

$$\min_{\theta, C, B^*} \quad L + \lambda Q$$

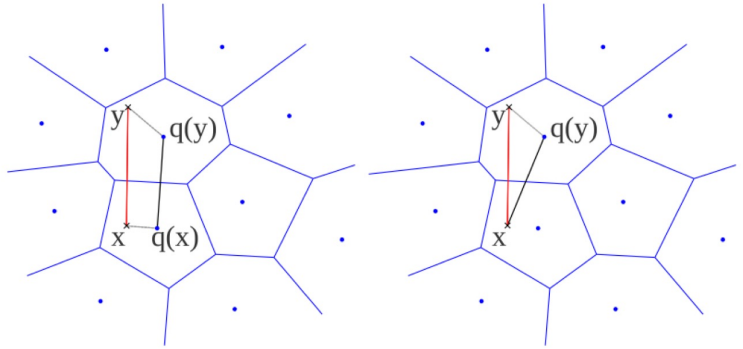
1 The parcel Q is the sum of Q_1 and Q_2

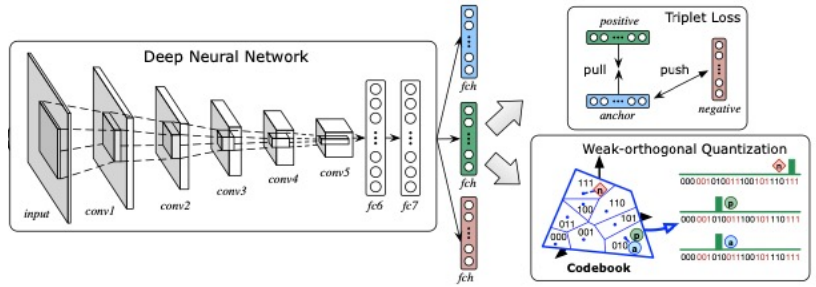
Classification: Confidential

Agenda

Classification: Confidential

The first benchmark aimed at the comparison of retrieval methods, binary encoding lengths and street vs studio images





- Exact match
- Products w/ same pattern
- Product w/ same category

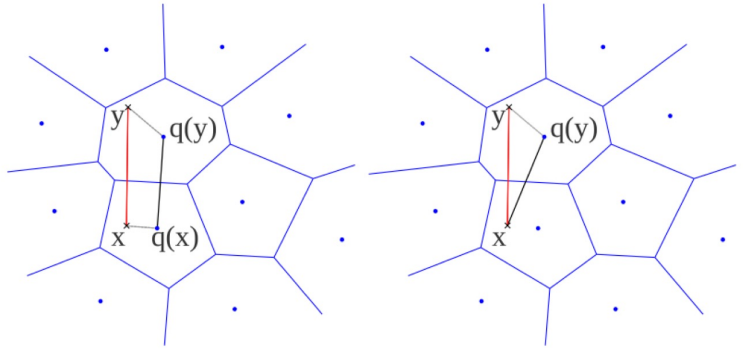
Street vs Studio Images

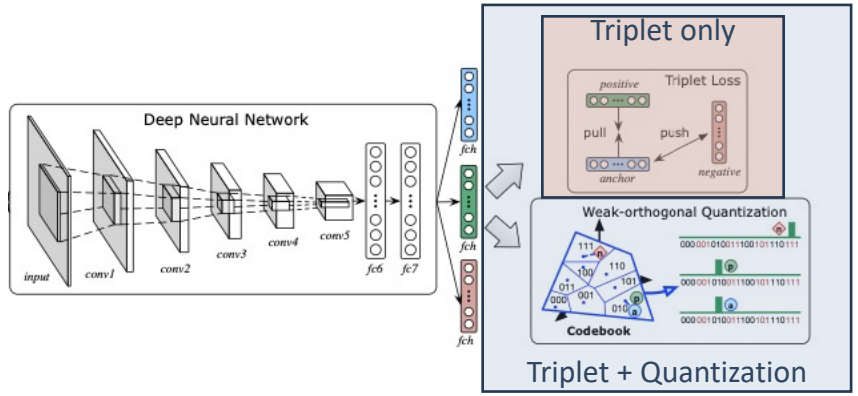
- Symmetric Distance Computation (SDC)
- Asymmetric Distance Computation (ADC)

- Triplet without Quantization (TNQ)
- Triplet with Quantization (TQ)
  - Different binary bit encodings
    - 32 bits
    - 48 bits
    - 64 bits

Classification: Confidential

The first benchmark aimed at the comparison of retrieval methods, binary encoding lengths and street vs studio images





- Exact match
- Products w/ same pattern
- Product w/ same category

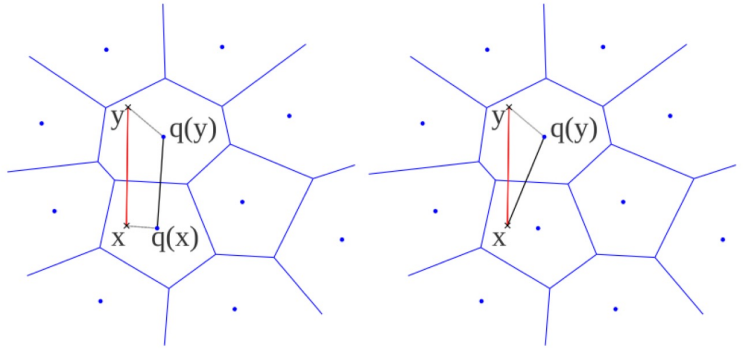
Street vs Studio Images

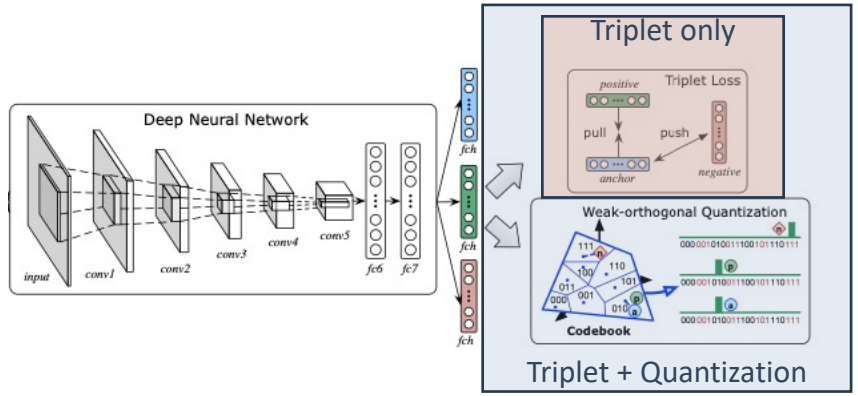
- Symmetric Distance Computation (SDC)
- Asymmetric Distance Computation (ADC)

- Triplet without Quantization (TNQ)
- Triplet with Quantization (TQ)
  - Different binary bit encodings
    - 32 bits
    - 48 bits
    - 64 bits

Classification: Confidential

The first benchmark aimed at the comparison of retrieval methods, binary encoding lengths and street vs studio images





- Exact match
- Products w/ same pattern
- Product w/ same category

Street vs Studio Images

- Symmetric Distance Computation (SDC)
- Asymmetric Distance Computation (ADC)

- Triplet without Quantization (TNQ)
- Triplet with Quantization (TQ)
  - Different binary bit encodings
    - 32 bits
    - 48 bits
    - 64 bits

Classification: Confidential

TQ - 64 bits for “Very Good” achieves 93.8%, 2.9 p.p. behind TNQ and 4.6 p.p. behind Human Team

Practical Use Case

Table 2 - Human evaluation of TQ vs TNQ vs Human Team (%)

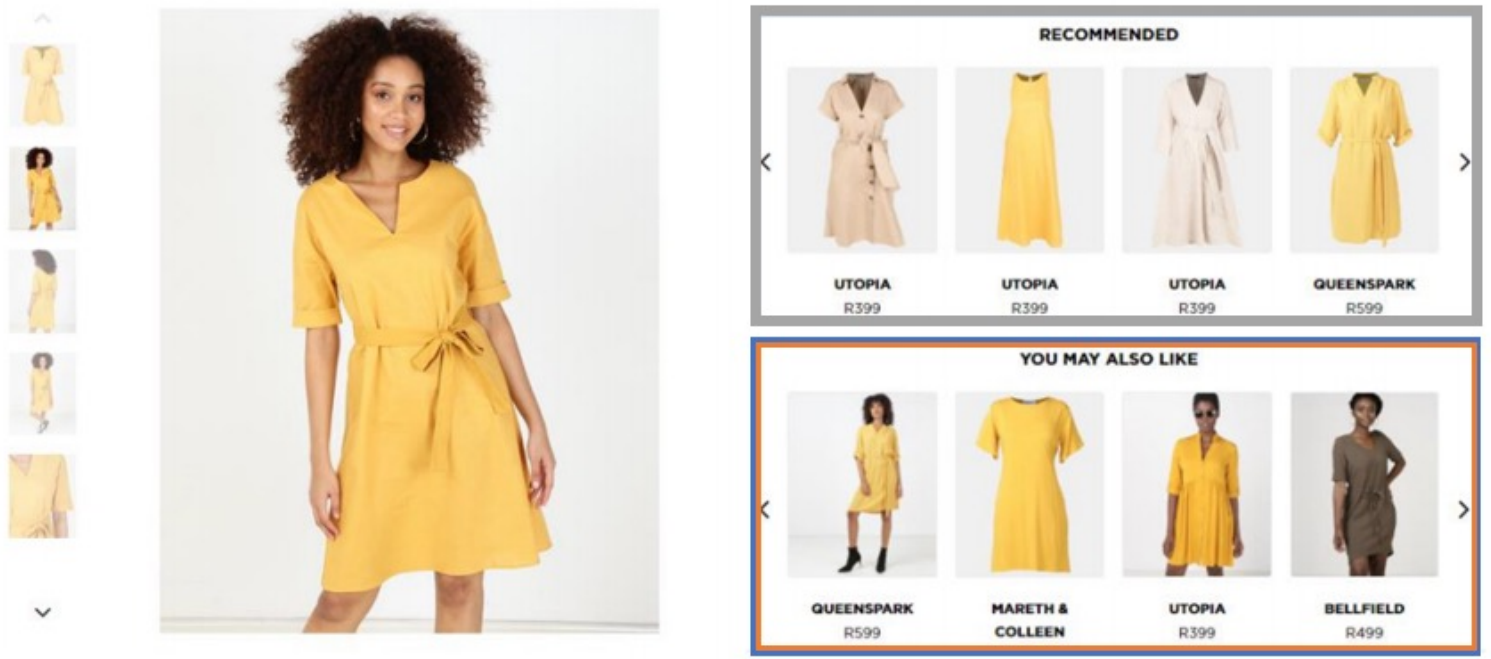
Classification	TQ - 32 bits	TQ - 48 bits	TQ - 64 bits	TNQ	Human Team
Very Poor	0.2	0.2	0.2	0.0	0.0
Poor	0.0	0.0	0.0	0.0	0.0
Medium	2.4	1.9	1.6	0	0.1
Good	3.7	4.2	4.4	3.3	1.5
Very Good	93.7	93.8	93.8	96.7	98.4

(a) (b)

Classification: Confidential

A/B Testing enables the real-world application of TQ and the analysis of its impact on business KPIs

A/B Testing – Offline Recommendations



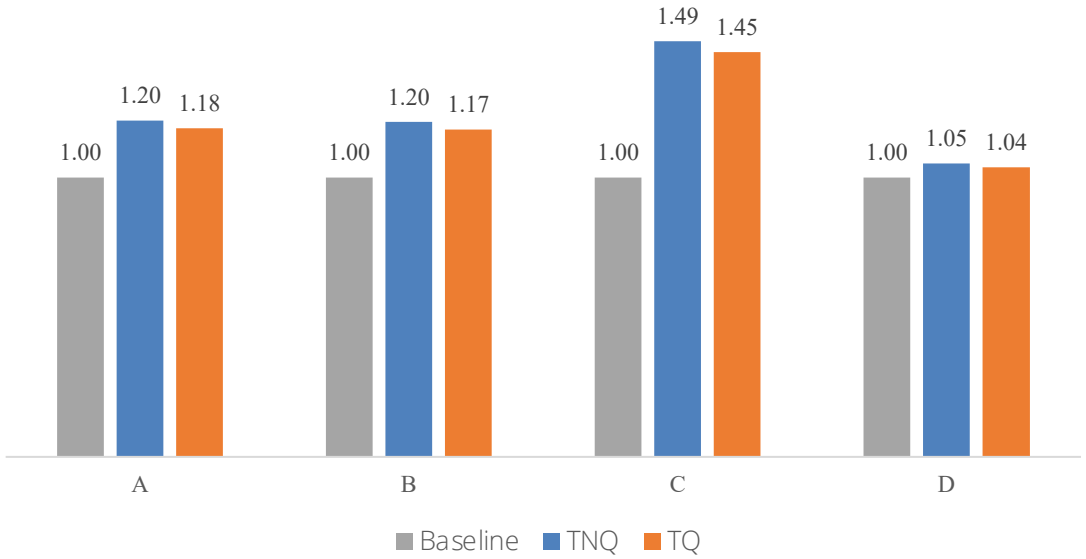
Baseline Recommender

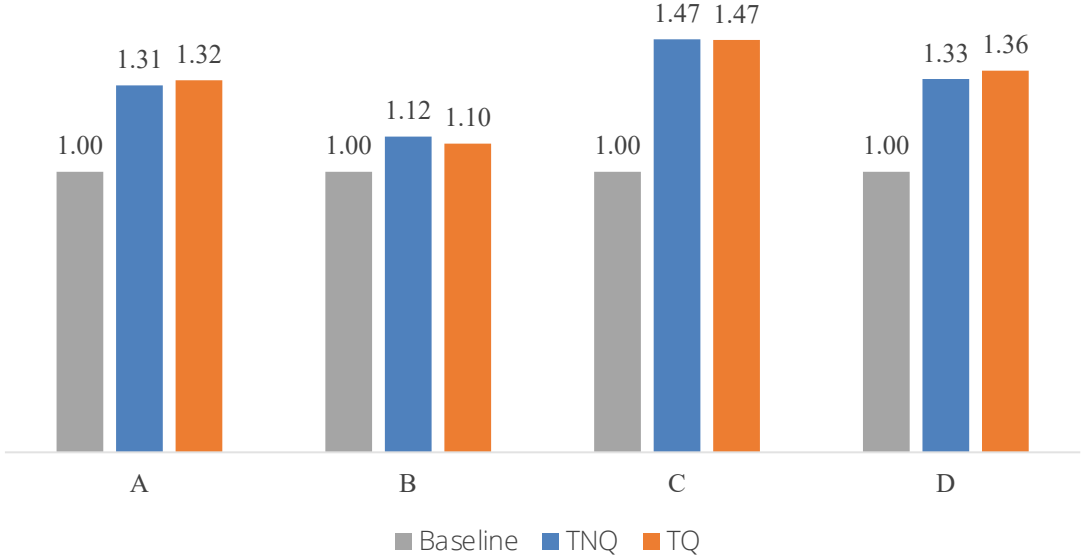
TQ or TNQ

Classification: Confidential

TQ and TNQ perform better than Baseline for all metrics, with no significant difference between them

A/B Testing





Classification: Confidential

Agenda

Classification: Confidential

Accurate and compact binary codes can be produced to efficiently compute similarity between images

Conclusions & Future Works

1. Study the application of Visual Transformers – recently developed architectures that achieved state of the art results on the ImageNet dataset by employing the transformer model to image processing and classification.

2. Study other deep hashing methods – within the quantization processes, DRQ[1] has drawn attention due to its capability to learn different binary size representations without performing several training stages – a drawback observed on DTQ[2].

1 Song, J., Zhu, X., Gao, L., Xu, X.-S., Liu, W., & Shen, H. T. (2019). Deep Recurrent Quantization for Generating Sequential Binary Codes. CoRR, abs/1906.0. http://arxiv.org/abs/1906.06699

2 Liu, B., Cao, Y., Long, M., Wang, J., & Wang, J. (2019). Deep Triplet Quantization. CoRR, abs/1902.0. http://arxiv.org/abs/1902.00153

Classification: Confidential

DSPT
DATA SCIENCE PORTUGAL

SHOPAI.CO

U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

VISUAL SIMILARITY WITH DEEP TRIPLET QUANTIZATION: APPLICATION TO THE FASHION INDUSTRY

PEDRO ESMERIZ| JOÃO GAMA

SCHOOL OF ECONOMICS AND MANAGEMENT OF THE UNIVERSITY OF PORTO

16TH MARCH 2023

ation: Confidential

Useful links

- How does product quantization work?
- Example of triple loss usage on PyTorch
- Cool animated explanation of neural networks
- Thesis of the project
- Deep Triplet Quantization paper

Classification: Confidential

TQ and TNQ achieve better results with ADC and TQ is better than TNQ – 64 bits by an average of 3.6 p.p.

Accuracy Results

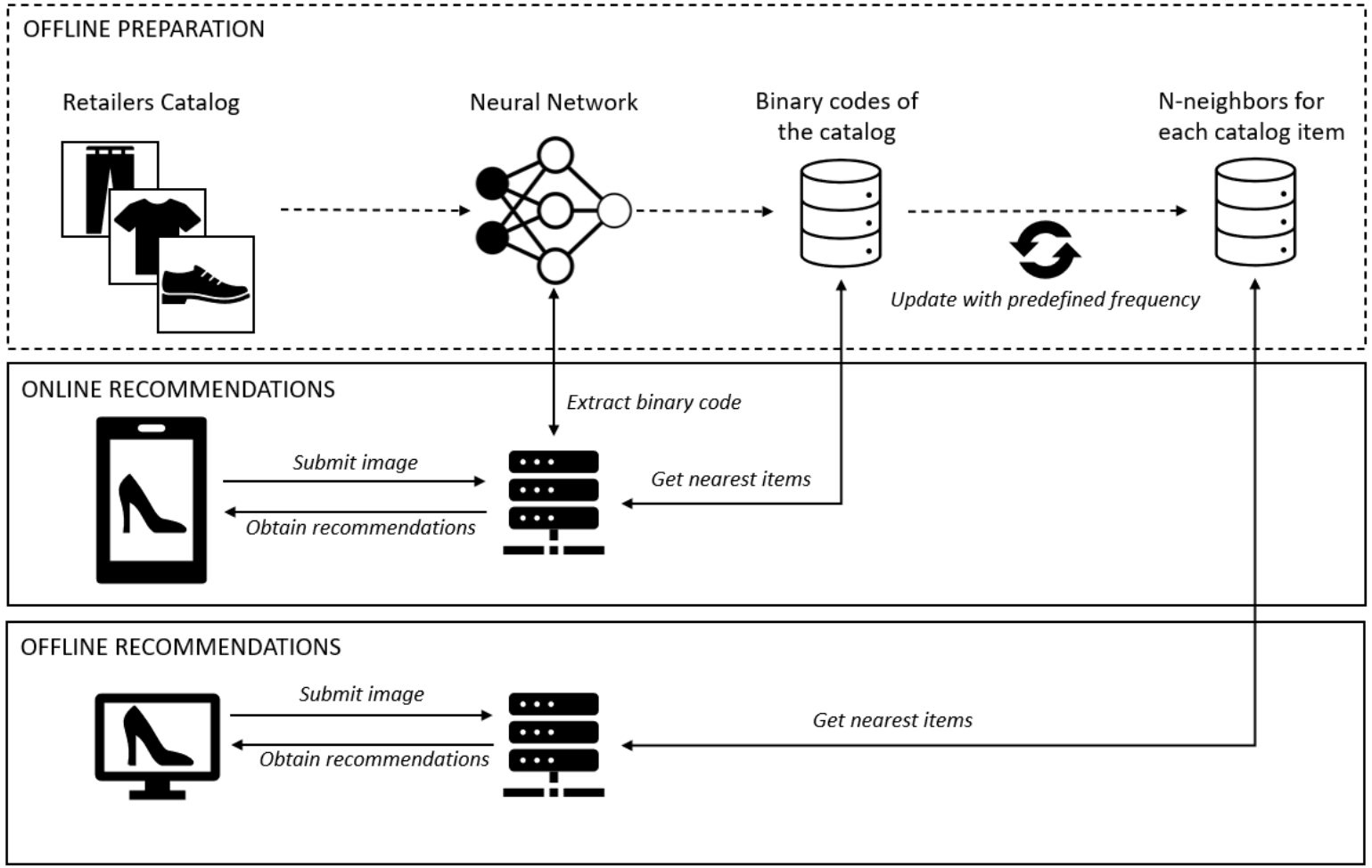
Table 3 - SDC mAP of TQ vs TNQ (%)

Conditions	Metric	TQ - 32 bits	TQ - 48 bits	TQ - 64 bits	TNQ
In Shop - Category	SDC	90.5	90.9	91.6	96.5
In Shop - Pattern	SDC	80.7	82.8	84.1	87.5
Street - Category	SDC	69.4	73.1	75.5	82.5
Street - Pattern	SDC	58.9	61.4	63.6	72.1
Street - Exact product	SDC	45.7	47.3	49.5	62.3
In Shop - Category	ADC	93.0	93.1	93.2	96.5
In Shop - Pattern	ADC	85.5	85.6	85.8	87.5
Street - Category	ADC	78.5	78.9	79.3	82.5
Street - Pattern	ADC	68.1	68.3	68.6	72.1
Street - Exact product	ADC	55.2	55.5	56.0	62.3

Classification: Confidential

The setup of both these recommendations involves the reprocessing of the retailers' catalog for every new item

Solution Architecture



Classification: Confidential