FASHION AND APPAREL CLASSIFICATION USING CONVOLUTIONAL NEURAL NETWORKS

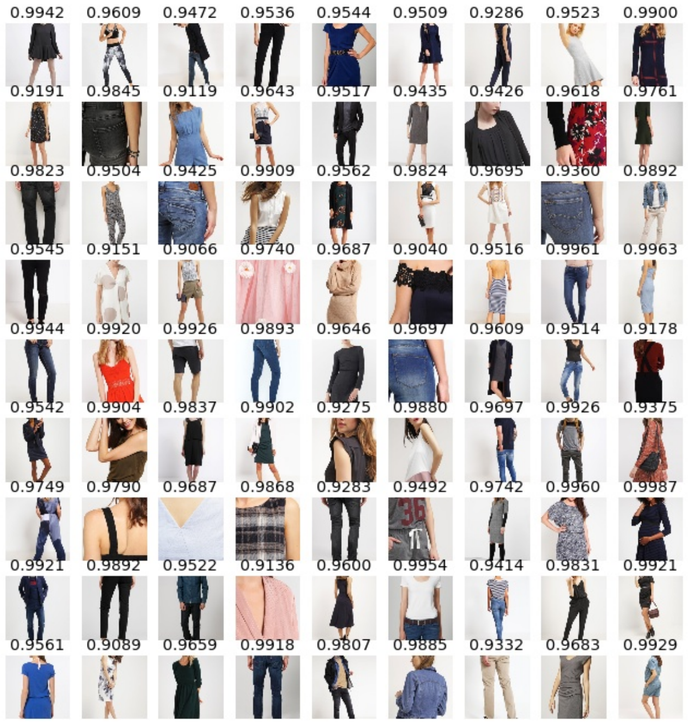
Alexander Schindler
alexander.schindler@ait.ac.at
Thomas Lidy
lidy@ifs.tuwien.ac.at
Stephan Karner
stephan.karner@monstyle.io
Matthias Hecker
matthias.hecker@monstyle.io

FASHION IMAGE CLASSIFICATION

- Online e-commerce access to product images
  - Asos-EU, Farfetch, Zalando
    - Images & metadat
- **Problem**
  - Metadata differs in
    - Quality, granularity, taxonomy
    - Taxonomy varies in depth of categorical hirarchy
- **Task**
  - use CNNs to
    - Consolidate Metadta
    - Enrich Metadata

BRIEF OVERVIEW

- Empirical study
- Applying deep Convolutional Neural Networks to fashion classification
- Evaluated five CNN architectures
- Custom and pre-trained models
- Evaluated on three tasks
  - Person detection
  - Product classification
  - Gender prediction

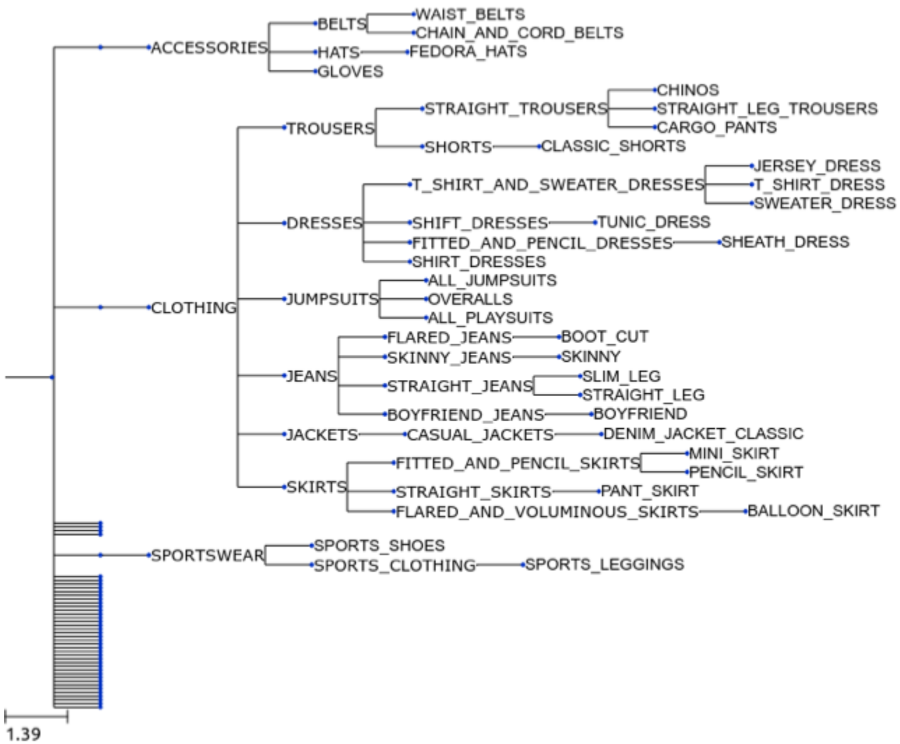


DATASETS

- **Person**
  - 7.833 images
    - 5.669 labeled as persons
    - 2.164 labeled as products
- **Products**
  - 234.884 images
  - 39.474 products
    - ~5,95 images per product
  - Ground-truth labels assignements
    - Product category
      - Label hirarchy
    - Gender
    - Age

Person

Product



DATA QUALITY / ISSUES

White background

Worn by persons

Text, Overlays

Close-up texture

Close-up fit

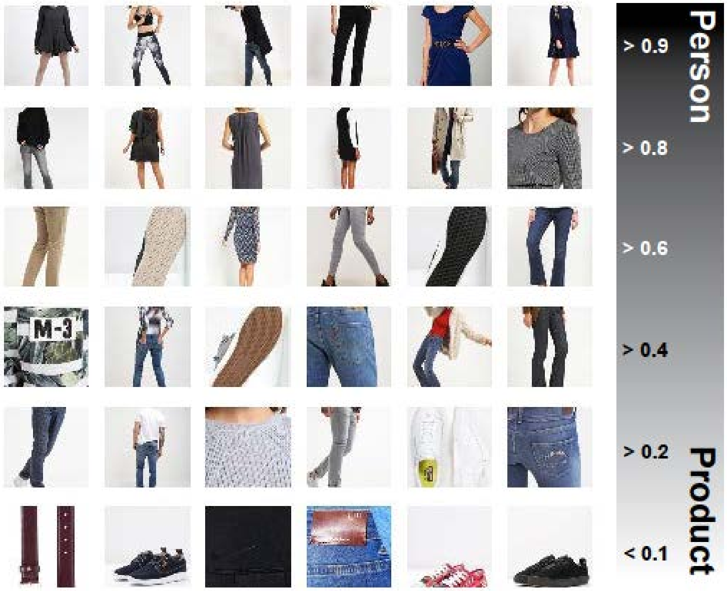
Multiple objects

Brand logo

Misc

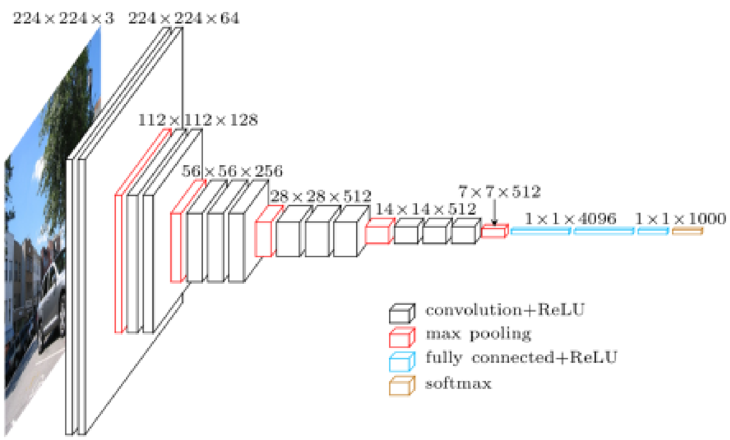
PERSON DETECTION

- **Products also presented by persons**
  - How they look when worn?
- **Problem**
  - Person wears multiple products
    - Single-label classification
    - Decission problem
- **Approach**
  - Train model to identify persons
  - Use model to filter images with persons
  - VGG-like custom model
- **Results**
  - **91.07%** accuracy on persons dataset

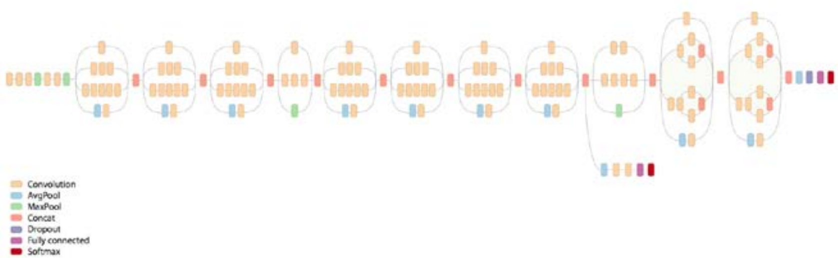


PRODUCT CLASSIFICATION

- **Deep Neural Network Architectures**
  - Vgg16 and Vgg19
  - InceptionV3
  - Custom CNN and Vgg-like
- **Experiments**
  - From-Scratch
  - Pre-Trained
- **Evaluation**
  - 4-Fold Crossvalidation
  - Grouped Stratification
- **Metrics**
  - Raw Accuracy
  - Max of Sum per product



David Fossard, https://www.cs.toronto.edu/~frossard/post/vgg16/



John Shlens, https://research.googleblog.com/2016/03/train-your-own-image-classifier-with.html

EXPERIMENTAL SET-UP

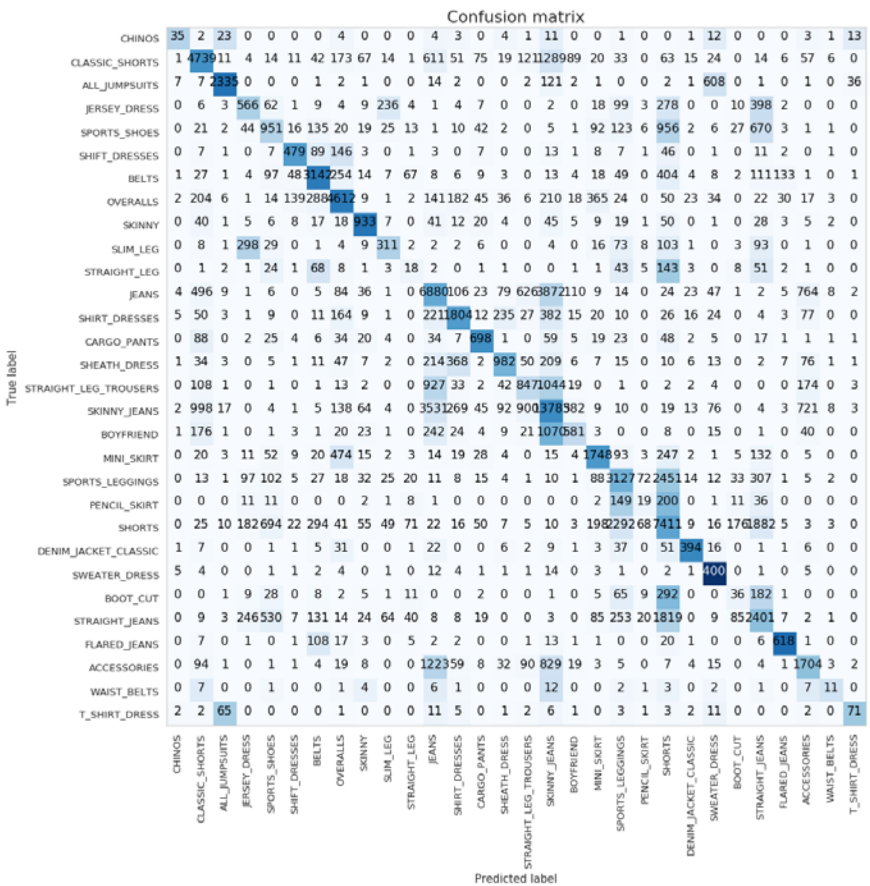
- **Small scale**
  - Subset of 23.305 images
- **Large scale**
  - 234.408 images
- **All Models**
  - Data Augmentation
    - 25% vertically and horizontally shifting
    - 25% zoom range
    - Horizontal flipping

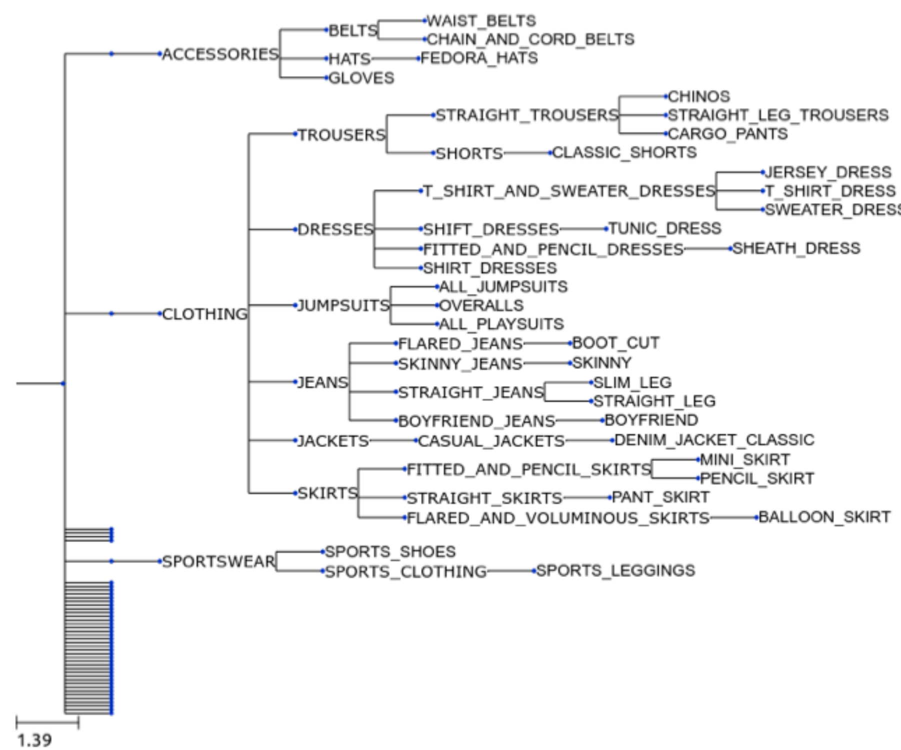
RESULTS – SMALL SCALE (24K)

- **Best results:** Pre-trained + fine-tune entire model
  - Freezing network + training only top layers not as good
- **Person filter** did not improve performance
- Small custom models have advantage of speed, but not as accurate

Description	best fold	best fold cum max	Mean cum max
InceptionV3, pretrained, fine-tuned	0.706	0.794	0.791
InceptionV3, pretrained, fine-tuned	0.658	0.729	0.716
VGG16, pretrained, fine-tuned	0.646	0.711	0.691
InceptionV3, pretrained, fine-tuned, person filter model as layer	0.569	0.685	0.658
VGG19, pretrained, fine-tuned	0.579	0.673	0.634
InceptionV3, pretrained, fine-tuned, no augmentation	0.564	0.673	0.647
VGG19, pretrained, train only top-layers	0.578	0.669	0.343
VGG16, pretrained, train only top-layers	0.603	0.652	0.368
InceptionV3, pretrained, train only top-layers	0.585	0.650	0.643
InceptionV3, pretrained, fine-tuned - person filtered metadata	0.640	0.636	0.614
InceptionV3, clean	0.492	0.594	0.580
Custom CNN, augmentation	0.506	0.568	0.538
Custom CNN	0.463	0.556	0.523
Custom VGG-like	0.438	0.549	0.519
VGG16, clean	0.439	0.455	0.443
VGG19, clean	0.437	0.447	0.430
VGG19, pretrained, train only top-layers	0.819	0.887	0.880
InceptionV3, pretrained, fine-tuned	0.798	0.863	0.836
VGG19, pretrained, fine-tuned	0.762	0.846	0.830

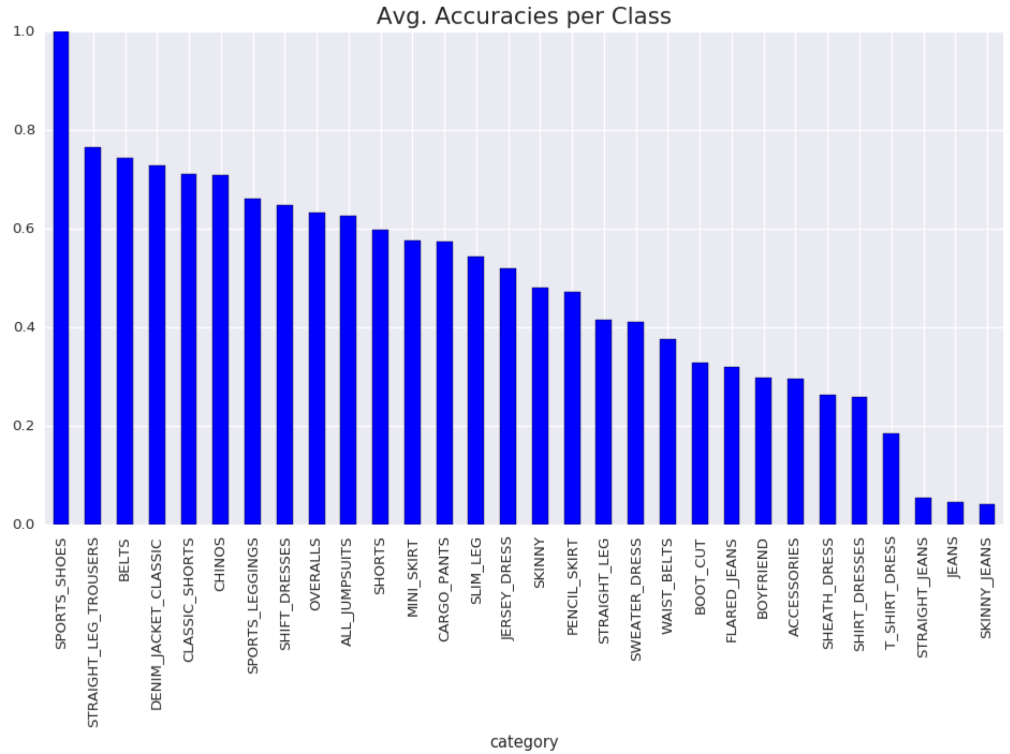
CONFUSIONS – SMALL SCALE (24K)





PER CLASS ACCURACIES – LARGE SCALE (234K)

- Problem of different granularity of provided ground truth
- Parent/Child nodes used interchangeably
  - Misclassification of child as parent is not wrong
  - Model does not consider hirarchy



CONCLUSIONS

- Despite large dataset and reduced number of classes
  - **Pretrained models outperform** from-scratch training
    - Product classification – 79.1%
    - Gender prediction – 88.0%

- **Custom small model** enough to learn **binary task**
  - person/product classification – 91.07%

- **Preprocessing** of **ground-truth** required
  - Flatten hierarchy, remove ambiguities and overlaps
  - Use **hierarchical CNNs**
  - Use **attention** (person images)

THANK YOU!

Alexander Schindler, 08.06.2018