Integration of Services for Software Development in DH: A Case Study of Image Classification using Convolutional Neural Networks

Florian Niebling[1] Michael Haas[2] André Blessing[3]

Abstract: This paper presents a case study of integrating RESTful web services for software development in the Digital Humanities. Journalists today are able to utilize huge amounts of image data to illustrate their articles. To provide easy access, image databases need to be structured and photographic documents stored within need to be categorized. Here, methods based on Deep Learning have recently had a significant impact on image categorization. We highlight the potentials of incorporating machine learning with different models for named-entity recognition in image captions, to facilitate classification of images according to standard IPTC media topics. We finally discuss the results of our project employing service-oriented architectures integrating specialized components to simplify the creation of DH applications.

Keywords: Service-oriented Architecture; Digital Humanities; Classification of Images; Machine Learning

1 Introduction

IPTC categories have been developed as a classification method for images to standardize exchange of images between journalists, newspapers and news agencies. Today, photographic images in these areas are stored in media archives such as the dpa Bildfunk[4], together with categorizations as well as fitting captions. In our previous work, an adapted ResNet neural network [He15] showed promising results in classifying photographic images according to IPTC categories. For many categories such as different sports, that are relatively homogeneous themselves, while at the same time very distinct from other categories, this approach works reasonably well. Nevertheless, there are many categories in IPTC media topic classification scheme where humans without sufficient context struggle as well. For example, images tagged with *politics* can look very similar to images tagged with *crime, law and justice* or *economy, business and finance*.

[1] Julius-Maximilians-Universität Würzburg, Human-Computer-Interaction, Am Hubland, 97074 Würzburg, Germany. florian.niebling@uni-wuerzburg.de
[2] Julius-Maximilians-Universität Würzburg, Human-Computer-Interaction, Am Hubland, 97074 Würzburg, Germany. michael.haas@stud-mail.uni-wuerzburg.de
[3] Institut für Maschinelle Sprachverarbeitung, Pfaffenwaldring 5b, 70569 Stuttgart, Germany. andre.blessing@ims.uni-stuttgart.de
[4] https://www.dpa.com/de/produkte-services/dpa-bilder

doi:10.18420/inf2019_ws17

In this work, we are trying to augment and support the tagging of images with image classification techniques and algorithms from computational linguistics, i.e. named entity recognition, to provide knowledge about persons depicted in photographs to the utilized neural networks.

Most Natural Language Processing (NLP) tools (e.g. taggers, parsers) used in academia are research projects themselves, and as such require specific environments (libraries, operating systems, hardware resources) which often make a local installation non-trivial. The combination of different tools into a specific workflow chain is even harder, since many tools require different and largely incompatible formatting of data. Infrastructure projects such as CLARIN[5] and DARIAH[6] try to address these issues by providing service oriented solutions. Components created in these research infrastructures can be easily integrated into specialized applications for distinct use cases.

We combine these methods in a service-oriented architecture, enhancing reusability of the developed independent components. We enhance our existing image classification mechanism by recognizing faces in images to enable better matching of photographic images to categories, under the assumption that specific recurring persons are often depicted in similar contexts.

2 Related Work

Commonly used NLP toolkits such as OpenNLP, CoreNLP or NLTK are generally not following the same standards in representing data, which makes exchanging modules between those frameworks hard. Infrastructures such as CLARIN try to counter this by encapsulating methods of those frameworks in a common setup. Based on the CLARIN Infrastructure, several use cases for encapsulation already exist, e.g. WebLicht [HHZ10] or Textual Emigration Analysis [BK14]

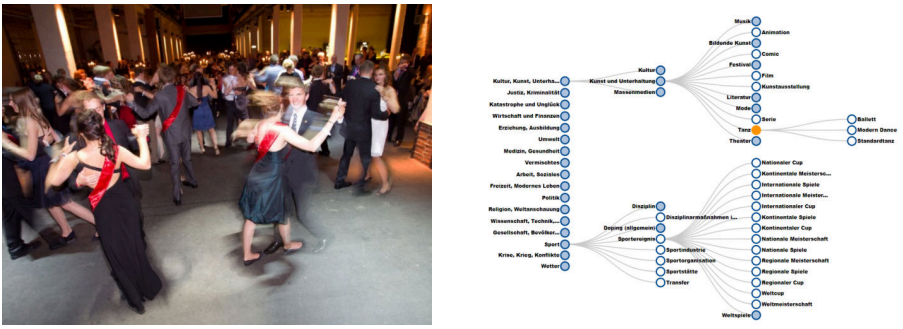
There have been huge improvements regarding image classification in the last years, especially since the introduction of ImageNet in 2009 [De09], a large-scale database classifying objects in images according to WordNet cognitive synonym categories (synsets). Crowdsourcing of the annotation process resulted in more than 14 million images tagged with 21.841 synsets. The Large Scale Visual Recognition Challenge (ILSVRC), a yearly image recognition competition, has been held from 2010 to benchmark algorithms based on a sample dataset containing 1000 ImageNet categories [Ru15]. Since images often cannot be satisfyingly categorized with a single class, e.g. due to multiple objects in a single image, evaluation according to top-1 error is problematic. In ILSVRC since 2012, top-5 error rate is predominantly used instead. Here, five predictions $c_{i1}, ...c_{i5}$ with the highest confidences are compared to the tagged class.

[5] www.clarin.eu
[6] www.dariah.eu

3 Use Case: Image Classification

Categorization of images often is performed by photo journalists under the impression of a current event or situation. This often leads to suboptimal tagging of images, impeding or even preventing re-use, since images might be hard to recover by a too narrow or misleading classification alone. Our previous work on image classification using an adapted ResNet network shows the differences between IPTC and ImageNet categories. Where classes used in ILSVRC contain relatively homogeneous motives, the complexity and diversity within news categories makes it difficult to extract reliable classifiers. Several of these classes were shown to be hard to distinguish even for humans without sufficient context knowledge in a user evaluation.

The approach presented here tries to add context knowledge to the neural network, in the form of persons depicted in images. The idea is that detecting people that can be categorized makes it easier to differentiate between categories consisting of images containing ambiguous settings such as e.g. *politics* and *economy, business and finance*. For this, we build and validate a training dataset containing faces and corresponding names of the depicted persons extracted using named-entity recognition on accompanying image captions.



(a) Example Image (b) IPTC Classification

Fig. 1: Tagging of images according to IPTC categories

The VGG face-descriptor based on the ILSVRC VGG architecture was trained on a dataset created using Google and Bing image search. The results of the second fully-connected layer, a 4096-dimensional feature vector, can be used to distinguish faces of different persons.

4 Method

The images contained in the dpa Bildfunk stream are categorized with one or several IPTC tags. As persons depicted in the images are not explicitly declared, they have to be extracted from the accompanying image captions. In 2018, we captured more than 12,000 images and captions over several months, trying to avoid depictions of a single event to dominate the dataset, and also trying to ensure a certain heterogeneity in the appearances of persons.

Several pre-processing steps had to be performed to make captured images usable for CNNs. First, regions containing faces were extracted from images using a face detector. Then, an automatic alignment of heads according to facial landmarks was performed. The faces were then scaled to a given resolution in a square aspect ratio, allowing mapping of pixels to the square input vector of our CNN.

To create adequate training data, named-entity recognition algorithms (NER) can be employed to extract names from image captions. We used a mixture of 2 NER-approaches: IMS German NER[7] and SFS German NER[8], which are both CLARIN webservices. Before we apply the NER methods, the image captions have to be pre-processed by applying the following services: *TEXT TO TCF*[9], *IMS Tokenizer*[10] and *IMS TreeTagger*[11].

5 Integration

The detection of faces in images inside the dataset was performed using the software library dlib [Ki09]. The actual recognition of faces was achieved through the TensorFlow-based implementation of the VGG-Face-Descriptor [PVZ15]. We used Keras as an interface to the TensorFlow framework.

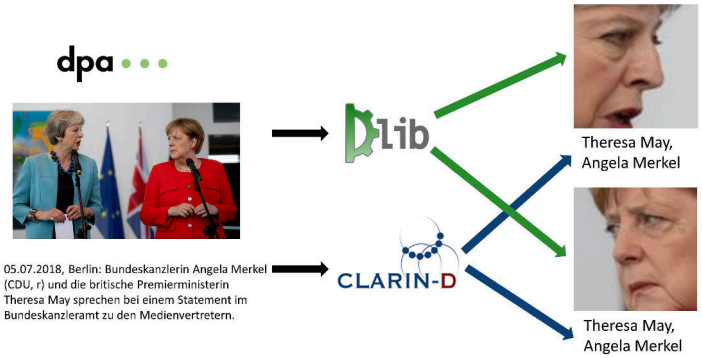


Fig. 2: Pre-processing of data from dpa: Facial extraction using dlib. Filtering of names from image caption using CLARIN-D-NER API. A mapping of faces to names according to descriptions (e.g. „r.“') is not possible due to inconsistent labeling.

The methods previously described in section 4 were combined into a single workflow to generate a training dataset. We combine the data generated by the face recognition algorithms with the output of the NER services, by initially linking every distinct face determined by the dlib library to each of the names returned from NER (see fig. 2). As there

[7] http://hdl.handle.net/11022/1007-0000-0000-8E2F-D
[8] http://hdl.handle.net/11022/0000-0000-1D84-B
[9] http://hdl.handle.net/11858/00-1778-0000-0004-BA56-7
[10] http://hdl.handle.net/11022/1007-0000-0000-8E1F-F
[11] http://hdl.handle.net/11022/1007-0000-0000-8E22-A

is often more than one person depicted in a single image, and often more than one entity is contained in the image caption, this leads to a huge amount of incorrect entries at first.

After having processed a large number of images and assigned names to the contained faces, we determine which of the multiple entities is actually the correct one for each face. We therefor perform two automated revision steps.

5.1 First Revision: Local Neighborhood Analysis in Vector Space

First, we estimate the appropriate name by analyzing the closest neighbors of each face in the 4096-dimensional vector space generated by the VGG Face Descriptor (see fig. 3). If a name was assigned to the ten nearest neighbors of a face vector which are closer than a given ϵ distance in vector space, we assume that name to be an appropriate match. This is possible under the following two assumptions:

- A1: For the majority of extracted faces, face descriptors are actually closer to descriptors of depictions of the same person, than to the others.
- A2: Each pair of two persons are only depicted together in a small fraction of all images in the dataset.

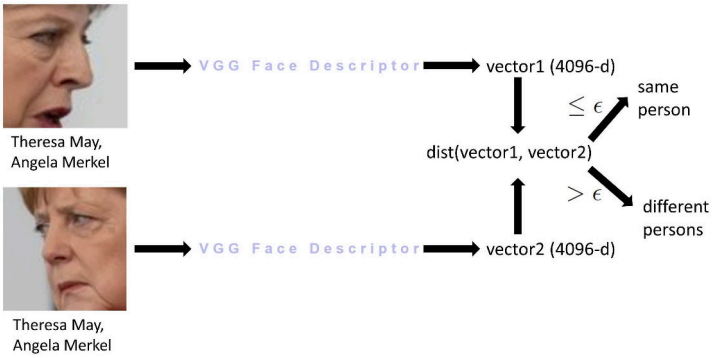


Fig. 3: Face recognition results in a 4096-dimensional vector. A distance metric is used to evaluate if two extracted faces are depictions of the same person.

5.2 Second Revision: Clustering of Face Vectors

In the second revision step, we remove outliers in the face descriptor vector space. After calculating median vectors of all face vectors assigned to a name, we only keep face descriptors with a cosinus distance from this median smaller than a selected threshold.

The validation dataset was built at the beginning of 2018. We selected pictures of the 30 persons whose names were returned most often by performing NER on the respective image captions. Wrong entries in this dataset were removed manually, to ensure a clean validation set. The resulting dataset contained from 117 depictions of a person (Donald Trump), over 90 (Carles Puigdemont), down to only 18 depictions each of Gustl Mollath und Philipp Kohlschreiber.

6 Conclusion

In this paper, we presented a method to extract named faces from customary news archives containing captioned images. We combined CLARIN-D webservices for named entity recognition with face extraction methods using previously trained networks to link all faces in an image to potential names extracted from the captions. The correct names are then identified by performing two revisions of the data. First, the most probable name for a face is identified by selecting the most frequent name assigned to the local neighborhood of its VGG face vector. Outliers are then reduced by removing faces whose face vector deviate strongly from the median face vector of all faces assigned to a name. The chosen service oriented approach facilitated the comparison of different NER models and allowed us to select an implementation that best suited our requirements.

Bibliography

[BK14] Blessing, Andre; Kuhn, Jonas: Textual Emigration Analysis (TEA). In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland, may 2014.

[De09] Deng, Jia; Dong, Wei; Socher, Richard; Li, Li-Jia; Li, Kai; Fei-Fei, Li: ImageNet: A Large-Scale Hierarchical Image Database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255, 2009.

[He15] He, Kaiming; Zhang, Xiangyu; Ren, Shaoqing; Sun, Jian: Deep Residual Learning for Image Recognition. CoRR, abs/1512.03385, 2015.

[HHZ10] Hinrichs, Erhard; Hinrichs, Marie; Zastrow, Thomas: WebLicht: Web-based LRT services for German. In: Proceedings of the ACL 2010 System Demonstrations. Association for Computational Linguistics, pp. 25–29, 2010.

[Ki09] King, Davis E: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research, 10(Jul):1755–1758, 2009.

[PVZ15] Parkhi, Omkar M; Vedaldi, Andrea; Zisserman, Andrew: Deep Face Recognition. In: British Machine Vision Conference (BMVC). volume 1, p. 6, 2015.

[Ru15] Russakovsky, Olga; Deng, Jia; Su, Hao; Krause, Jonathan; Satheesh, Sanjeev; Ma, Sean; Huang, Zhiheng; Karpathy, Andrej; Khosla, Aditya; Bernstein, Michael et al.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3):211–252, 2015.