



Concept Detection in Medical Images using Xception Models - TUC_MC at ImageCLEFmed 2020

Nisnab Udas¹, Frederik Beuth² , and Danny Kowerko³ 

¹ Chemnitz University of Technology, Germany
nisnab.udas@gmail.com

² Chemnitz University of Technology, Germany
beuth@cs.tu-chemnitz.de

³ Chemnitz University of Technology, Germany
danny.kowerko@cs.tu-chemnitz.de

Abstract. This paper summarizes the approach and the results of the submission of the Media Computing group from the Chemnitz University of Technology (TUC_MC) at ImageCLEFmed Caption task, launched by ImageCLEF 2020. In the task, contents of medical images have to be detected, for the goal of a better diagnosis of medical diseases and conditions. In the context of a master thesis by Nisnab Udas, Xception model, which slightly outperformed InceptionV3 on the ImageNet dataset in 2017, was adapted to this caption task. Out-of-the-box, his approach achieved an F1 score of 35.1% compared to the best contribution with 39.4%, which places our team in the top-5. Part of his strategy was to optimize the confidence threshold, and to bring in a max pooling in the last layer which reduced the number of parameters making the model less prone to overfitting.

1 Introduction

Computer science challenges have been established in the last decades to advance diverse problems in text, audio and video processing [4]. In this tradition, challenges are organized within the established ImageCLEF or LifeCLEF lab since 2003 and 2014, respectively. Since 2003, medical (retrieval) tasks have been part of the challenge and been continuously developed into 3 subtasks, where one is called medical concept detection since 2017 [2,13,10,6]. It contains automatic image captioning and scene understanding to identify the presence and location of relevant concepts in a large corpus of medical images. The latter stem from the PubMed Open Access subset containing 1,828,575 archives. A total number of 6,031,814 image - caption pairs were extracted. A combination of automatic filtering with deep learning systems and manual revisions was applied to focus

Copyright © 2020 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0). CLEF 2020, 22-25 September 2020, Thessaloniki, Greece.

Table 1. Columns 2 to 4 show the F1 metric results in % for the best submission run of the top 3 teams at medical caption subtask since 2017. Note, that the number of registrations is usually considerably larger than the number of teams who submit results (last column).

Year	1st	2nd	3rd	No. of teams
2017	12.1	9.6	5.0	4
2018	25.0	18.0	17.3	5
2019	28.2	26.6	22.4	11
2020	39.4	39.2	38.1	7

merely on radiology images and non-compound figures. The origin of the biomedical images distributed in this challenge is a subset from the extended ROCO (Radiology Objects in COntext) dataset [11]. In ImageCLEF 2020, additional information regarding the modalities of all 80,747 images was distributed [6].

Evaluation is conducted in terms of set coverage metrics such as precision, recall, and combinations thereof. Leaderboards utilize the F1 metric summarized in Table 1. The results prove that the task remains challenging even though a continuous improvement from year to year is to be noted. The results of this year bring the top-3 group closer together for the first time. In the caption task of 2019 [10], Kougia *et al.* won the competition by combining their CNN (Convolutional Neural Network) image encoders with an image retrieval method or a feed-forward neural network and achieved an F1 score of 28.2% [7]. Xu *et al.* applied a multilabel classification model based on ResNet [5] and achieved 26.6% [14]. Guo *et al.* achieved 22.4% F1 score with a two-stage concept including the medical image pre-classification based on body parts with AlexNet ([8]) and the transfer learning-based multi-label classification model based on Inception V3 and Resnet152 [3].

2 Data Analysis

The amount of images has increased from 2019 to 2020. The concept detection task this year, contains training and validation images in 7 separate folders. In total, there are 64,753 training images, 14,970 validation images and 3,534 test images, respectively. Concept frequency was reduced from 5,528 last year to 3,047 in 2020 as low occurring concepts were removed by the organizers. Top-20 concepts in our training images are shown in Fig. 1. Concepts 'C0040405' and 'C0040398' both occur 25022 images in training images. The figure clearly shows how our concepts are imbalanced in the dataset.

In Fig. 2, we show distribution of concept length in training dataset. Maximum number of images, 5,248, to be specific, have only 2 concepts. The second and third largest group of images have 3 and 4 concepts per image, respectively. The highest number of concepts occurring in an image is 140 which occurs one time.

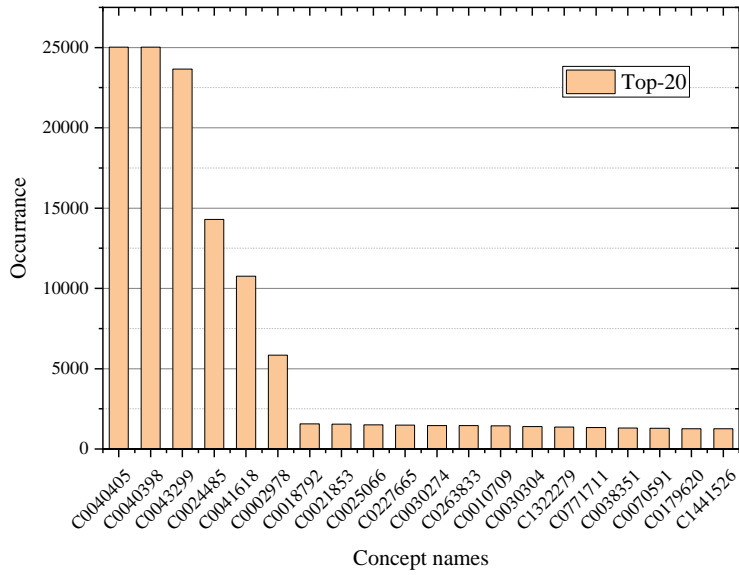


Fig. 1. Frequency distribution of Top-20 concepts.



Fig. 2. Frequency distribution of the concepts length.

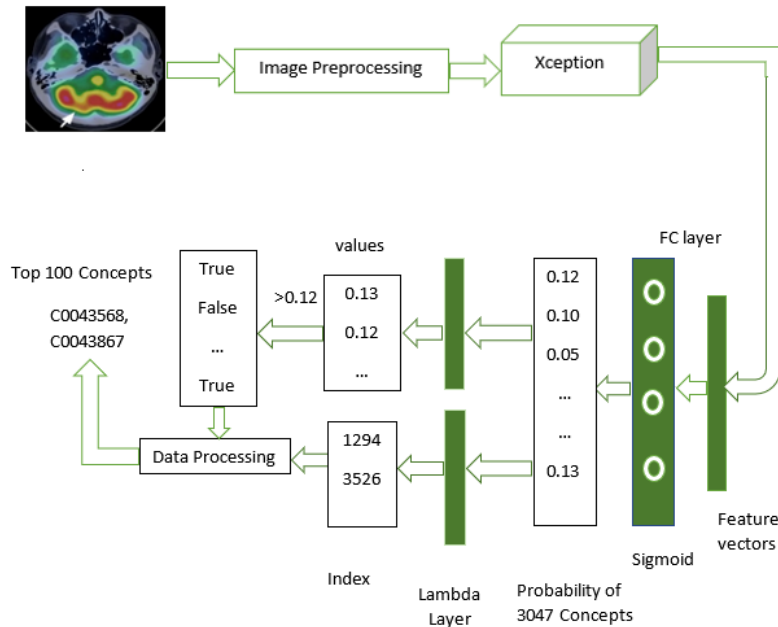


Fig. 3. Model architecture describing the mechanism of concept detection from pre-trained Xception model.

3 Proposed System

Our deep learning based architecture was based on Xception architecture [1] and is shown in Fig. 3. The Xception model slightly outperforms Inception V3 on the ImageNet dataset in 2017, and was chosen due to this performance and as our preliminary test found it well working on our medical detection task. For fine-tuning concerning model, we utilize transfer learning and use weights pre-trained on ImageNet Dataset [12]. We then eliminated the top classifier layer as is required in transfer learning. We froze the entire Xception model and made only the last six layers trainable.

Generally, as in transfer learning, before adding classifier to a pre-trained model, the layer is flattened. Flattening transforms a 2D matrix of features to the vector, which can be provided to a fully-connected layer (FC layer). In our case, we used a max pooling layer of window size (2,2) followed by the dropout layer to reduce the number of free parameters and facilitate object-size dependent pooling as a special trick. Afterwards, the usual flattening layer is added.

Subsequently to adding the flattening layer, we used the ReLU activation function, followed by the dropout layer. The data contains 3,047 concepts in total, thus our final FC layer contains 3,047 units and sigmoid as our activation function because we are dealing with a multi-label problem. The white rings in the FC layer represent neurons (Fig. 3). The top lambda layer extracts top-100 highest probabilities. These probabilities are compared against a threshold

Table 2. Model summary with layers and feature maps changing shapes as it passes through these layers.

Layer	Output shape
Xception	$16 \times 5 \times 5 \times 2048$
Max pooling	$16 \times 2 \times 2 \times 2048$
Dropout	$16 \times 2 \times 2 \times 2048$
Flatten	16×8192
Activation	16×8192
Dropout	16×8192
Dense	16×3047

Table 3. Influence of the high-level max-pooling.

Method	Mean F1 score	Free trainable parameters
Base model	0.349	29,712,871
Without max-pooling	0.345	160,758,247

value, e.g. $t=0.12$, which generates boolean values for these 100 probabilities. The lower lambda layer gives the index of these individual neurons/concepts. In data processing, these indices are used to locate only the neurons with 'True' boolean value. In the data processing part, results are reformatted into the competition format. Table 2 shows how feature maps change shape after passing through each layer.

For training our network, a proper optimization is necessary and we conduct the following optimization methods. The major contribution to a satisfying F1 score had the optimization of the confidence threshold, along with the max-pooling, as shown in the next sections. Besides these two methods, we deploy other minor approaches to raise the performance. These include the tuning of the drop-out level and the data augmentation level. Drop-out is a well-suitable technique to avoid overfitting and the values were optimized for both neuronal layers of drop-out (Table 2) via conducting a cross-test of 25 different combinations. The best configuration has a drop-out value of 0.2 for the first layer and 0.5 for the second layer. Additionally, data augmentation was tuned, also increasing the F1 score value by 0.01–0.02 depending on the configuration. Each of the methods raise the F1 score by 0.01 – 0.02 only, but in total, these effects add up to an elevate of the F1 score by a level of 0.03 – 0.05.

4 Results

One of the ideas for improving the original Xception model [1] was the introduction of an additional max-pooling operation before the highest layer. It is shown in Table 2 in the second entry. This particular max-pooling operation reduces, on the application set, the spatial resolution, inducing a reduction of the free parameters in the next layer. In our dataset, the layer before the max-pooling operation had a 5×5 resolution, which is reduced by a 2×2 pooling to a

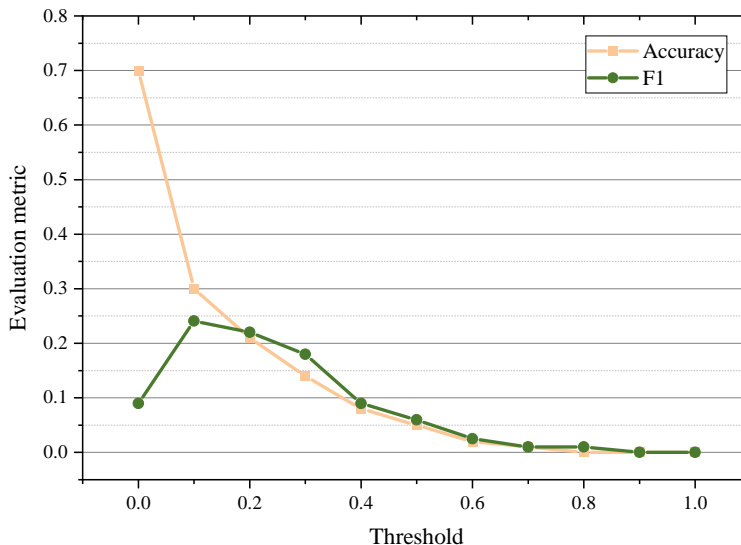


Fig. 4. Variation in F1 score and accuracy with respect to the threshold.

2×2 layer resolution. This operation reduces the free trainable parameters from 160,758,247 parameters to 29,712,871 parameters in total, which fabricates a more robust and stable model. As a second argument, the operation allows the recognition of concepts in the image more independently from the position. In the original ImageNet data set, objects are larger on average than in our medical image data set. To compensate for this difference in size, we increase the pooling as the objects in the original dataset cover large portions of the image, while our concepts are typically appearing in a smaller region. The pooling operation allows both, a recognition independent of the concept's place, and smaller object sensitive filters facilitating the recognition of smaller objects. The difference in F1 score and performance, i.e. free trainable parameters, is shown in Table 3.

4.1 Confidence threshold optimization

Fig. 4 shows the threshold variation against accuracy and F1. A classical accuracy metric is not optimal for training a model in this challenge as we have a large class imbalance. Therefore, the F1 score metric is used.

Confidence threshold selection plays a crucial role in the multi-label problem. The threshold determines over which predicted probability a concept is mapped to our image or not. When a class is predicted, the network outputs a probability and only probabilities exceeding a certain threshold are counted as that this concept is in that image. Given that our model is well trained, an unoptimized threshold may still have a substantial effect on our result. And determining the optimum threshold can often be tricky.

Therefore, we varied the threshold systematically and tuned the value of the threshold on the validation set, shown in Fig. 4. The maximum performance with respect to the confidence threshold is identified to range around 0.1 to 0.25. Hence, we submitted several runs with different threshold values between $\theta = 0.12$ and $\theta = 0.25$ (see Table 4). As expected from the Fig. 4, improvements of F1 are within a large amount.

4.2 Optimization techniques

There are plenty of ways of managing limited data volume and imbalanced datasets such as eliminating outliers, expanding the data set, augmentation, etc. In the medical image domain, few type of disease or conditions occurs less frequently to humans resulting in less sample numbers. Thus, to tackle these problems, we decided to use image data augmentation. Available methods are for example in Keras the following, which we employ as parameterized below:

- Rotation is performed by randomly rotating an image around its center of up to 5° .
- Vertical and horizontal flip. Flipping images is one of the most widely implemented techniques popularized by [8].
- Height and width shift range: The images are randomly shifted horizontally or vertically up to 5% of the total height and width respectively.
- Zoom: Objects in images are randomly zoomed in a range of $\pm 5\%$.
- Brightness shift: The image is randomly darkening or brightening in range of 80 – 120% of the initial brightness.
- Samplewise center: To eliminate the problem of vanishing gradients or saturating values, data are normalized in such a way that the mean value of each data sample becomes 0.
- Samplewise standard normalization: This pre-processing method approaches the same concept as sample-wise centering, but rather it fixes the standard deviation to 1.

The enabling of data augmentation increases the F1 score and contributes to a more robust working of the system.

The competition requires only 100 concepts per image. Therefore, to ensure that, probabilities were sorted in descending order and Top-100 probabilities were selected.

4.3 Run description

We submitted ten runs (Table 4). The runs often utilize the same base-structure, an Xception model, and all use transfer learning from ImageNet. The runs vary in meta-parameters as we tested different ones. We vary primarily: (i) the threshold in the last layer, (ii) slightly different base-models, and (iii) with and without max-pooling in the highest layers.

Table 4. Test results of our 10 submitted runs. For details see text.

Run id	Method	Name	Mean F1 score	Rank
68077	Early stopping	model_thr0_18.csv	0.351	20
68078	CNN2, $\theta = 0.25$	streamlined1_thr0_25.csv	0.349	21
68034	CNN2, $\theta = 0.20$	streamlined1_thr0_20.csv	0.349	22
68074	CNN2, $\theta = 0.15$	streamlined1.csv	0.349	23
68029	CNN1, $\theta = 0.20$	basemodel_thr0_20.csv	0.347	24
68045	Slow learning	model_low_lr_thr0_20.csv	0.345	25
68067	No max-pooling	streamlined1_nomax.csv	0.345	27
68024	CNN1, $\theta = 0.15$	basemodel.csv	0.343	28
68073	CNN2, $\theta = 0.12$	streamlined1_thr0_12.csv	0.342	29
68076	Exp. Normalizing	model_weighting.csv	0.332	32

Run_ID 1/68024 We deploy an Xception model and utilize transfer learning from ImageNet. We ran our model for N=100 epochs and set the learning rate to 1e-3. The model uses the confidence threshold θ in the last layer to map concepts probabilities to true/false for the concepts. We tuned the threshold to 0.15 from the validation set, selected the top 100 concepts and submitted our results.

Run_ID 2/68029 This run again uses the Xception model and generally the configuration of Run_ID 1. It optimizes the threshold further, setting it to 0.20.

Run_ID 3/68034 We again deploy an Xception model and utilize transfer learning from ImageNet. This submission has a more streamed-lined source code structure and explores different meta-parameters: We ran our model for N=30 epochs and set the learning rate to 1e-2. We tuned the threshold to 0.15 from the validation set.

Run_ID 4/68045 We again deploy an Xception model and utilize the configuration of run 1. This submission explores different meta-parameters: We ran our model for N=50 epochs and set the learning rate to 1e-4. We tuned the threshold to 0.20 from the validation set.

Run_ID 5/68067 This run again uses the Xception model and generally the configuration of Run_ID 3, while exploring which effect has the max-pooling layer before the highest layer. The max-pooling was removed here to show the effect.

Run_ID 6/68073 This run uses the more streamed-lined source code structure and explores again different meta-parameters: We ran our model for N=30 epochs and set the learning rate to 1e-2. We tuned the threshold to 0.12 from the validation set.

Run_ID 7/68074 This run again uses the Xception model and general the configuration of Run_ID 6, while tuning the threshold to 0.25.

Run_ID 8/68076 We again deploy the standard configuration of Run_ID 1. This submission focuses on an experimental normalizing of the dataset, but was not very successfully.

Run_ID 9/68077 We again deploy an Xception model and utilize transfer learning from ImageNet. This submission explores an early stop strategy: the best, i.e. lowest loss, was used over a run period of N=30 epochs. The learning rate was 1e-3 and the threshold was tuned to 0.18.

Run_ID 10/68078 This run deploys the more streamed-lined source code structure and explores a different threshold: 0.20.

In Table 5, we listed the top teams with their best F1 score in percent. Our team, TUC_MC occupied 5th position in terms of team ranking with F1 score of 0.3512.

Table 5. Top-7 team performance in concept detection problem 2020. F1 metrics are given in percent. [9]

Group Name	F1 score
AUEB_NLP_Group	39.4
PwC_MedCaption_2020	39.2
essexgp2020	38.1
iml	37.5
TUC_MC	35.1
Morgan_CS	16.7
saradadevi	13.5

5 Conclusion and Outlook

Our approach of adapting an Xception model for the medical caption task 2020 achieves an F1 score of 35.1% which is better than the 2019 results and close to the best contributions of 2020 which achieved 39.4%. Our strategies to rely on a modern Xception neural network proved to be successful. It also shows that transfer learning, with weights pre-learned on ImageNet, is very usable on an indeed different image material such as medical images. The introduction of a max pooling in the last layer, and to optimize the confidence threshold, have boosted the performance of our Xception model. Further investigation could lead in the direction of optimization learning through entropy-based analysis concepts of neural networks. Moreover, a more in-depth analysis of certain concept classes might be carried out in order to better understand the errors in the present classification task.

References

1. Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1800–1807. IEEE, Honolulu, HI (Jul 2017). <https://doi.org/10.1109/CVPR.2017.195>, <http://ieeexplore.ieee.org/document/8099678/>
2. Eickhoff, C., Schwall, I., Müller, H.: Overview of ImageCLEFcaption 2017 – Image Caption Prediction and Concept Detection for Biomedical Images. Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum **1866**, 1–10 (Sep 2017), <http://ceur-ws.org/Vol-1866/>
3. Guo, Z., Wang, X., Zhang, Y., Li, J.: ImageSem at ImageCLEFmed Caption 2019 Task: a Two-stage Medical Concept Detection Strategy. Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum **2380**, 1–8 (Sep 2019), http://ceur-ws.org/Vol-2380/paper_80.pdf
4. Hanbury, A., Müller, H., Balog, K., Brodt, T., Cormack, G.V., Eggel, I., Gollub, T., Hopfgartner, F., Kalpathy-Cramer, J., Kando, N., Krithara, A., Lin, J., Mercer, S., Potthast, M.: Evaluation-as-a-Service: Overview and Outlook. arXiv:1512.07454 [cs] pp. 1–28 (Dec 2015), <http://arxiv.org/abs/1512.07454>
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
6. Ionescu, B., Müller, H., Péteri, R., Abacha, A.B., Datla, V., Hasan, S.A., Demner-Fushman, D., Kozlovski, S., Liauchuk, V., Cid, Y.D., Kovalev, V., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Ninh, V.T., Le, T.K., Zhou, L., Piras, L., Riegler, M., Halvorsen, P., Tran, M.T., Lux, M., Gurrin, C., Dang-Nguyen, D.T., Chamberlain, J., Clark, A., Campello, A., Fichou, D., Berari, R., Brie, P., Dogariu, M., Ștefan, L.D., Constantin, M.G.: Imageclef 2020: Multimedia retrieval in medical, lifelogging, nature, and internet applications. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020), vol. 12260. LNCS Lecture Notes in Computer Science, Springer, Thessaloniki, Greece (September 22-25 2020)
7. Kougia, V., Pavlopoulos, J., Androutsopoulos, I.: AUEB NLP Group at ImageCLEFmed Caption 2019. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. vol. 2380, pp. 1–8. Lugano, Switzerland (Sep 2019), http://ceur-ws.org/Vol-2380/paper_136.pdf
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012), <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
9. Pelka, O., Friedrich, C.M., García Seco de Herrera, A., Müller, H.: Overview of the ImageCLEFmed 2020 concept prediction task: Medical image understanding. In: CLEF2020 Working Notes. CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece (September 22-25 2020)
10. Pelka, O., Friedrich, C.M., Seco De Herrera, A.G., Müller, H.: Overview of the ImageCLEFmed 2019 Concept Detection Task. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum. vol. 2380, pp. 1–13. Lugano, Switzerland (Sep 2019), http://ceur-ws.org/Vol-2380/paper_245.pdf

11. Pelka, O., Koitka, S., Rückert, J., Nensa, F., Friedrich, C.M.: Radiology Objects in COntext (ROCO): A Multimodal Image Dataset. In: Stoyanov, D., Taylor, Z., Balocco, S., Sznitman, R., Martel, A., Maier-Hein, L., Duong, L., Zahnd, G., Demirci, S., Albarqouni, S., Lee, S.L., Moriconi, S., Cheplygina, V., Mateus, D., Trucco, E., Granger, E., Jannin, P. (eds.) *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*, vol. 11043, pp. 180–189. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-01364-6_20, http://link.springer.com/10.1007/978-3-030-01364-6_20, series Title: Lecture Notes in Computer Science
12. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* **115**(3), 211–252 (Dec 2015). <https://doi.org/10.1007/s11263-015-0816-y>, <http://link.springer.com/10.1007/s11263-015-0816-y>
13. Seco De Herrera, A.G., Eickhoff, C., Andrearczyk, V., Müller, H.: Overview of the ImageCLEF 2018 Caption Prediction Tasks. In: *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*. vol. 2125, pp. 1–12. Avignon, France (Sep 2018), http://ceur-ws.org/Vol-2125/invited_paper_4.pdf
14. Xu, J., Liu, W., Liu, C., Wang, Y., Chi, Y., Xie, X., Hua, X.: Concept detection based on multi-label classification and image captioning approach - DAMO at ImageCLEF 2019. In: *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*. vol. 2380, pp. 1–10. Lugano, Switzerland (Sep 2019), http://ceur-ws.org/Vol-2380/paper_141.pdf