Deep Learning for Cervical Cancer Diagnosis: Multimodal approach

Hugo Barros
FEUP
up201404104@fe.up.pt

Jingjing Zheng
ISEP
zheng@isep.ipp.pt

Tomé Albuquerque
FEUP
up201708868@fe.up.pt

Abstract

Cervical cancer is one of the most common cancer types worldwide and has an elevated mortality rate. This is mainly due to the asymptomatic feature that is related with the first stages of the disease. Therefore it is extremely important to perform early diagnosis and effective screening in order to detect the problem in its initial stages. Several developed countries have already implemented screening programmes however the process is extensive composed by HPV test, cytology test (or Pap smear), colposcopy, and biopsy. Therefore tools to facilitate the process may highly benefit both the effectiveness and time saving aspect of any phase. In this paper it is suggested a deep learning approach to classify the risk of cancer by analyzing colposcopy images using two datasets with different labels but similar cervigram images. The Neural Model constructed features Multimodal and Multitask learning making use of the two different datasets in two different classification tasks with a shared section that learns information from both. The best model was able to achieve 92.97% AUC, 94.92% sensitivity and 83.12% specificity using ResNet architecture.

1. Introduction

1.1. Motivation

Cervical cancer is the fourth most frequent cancer in women with an estimated 570,000 new cases in 2018 representing 6.6% of all female cancers [1] worldwide. The survival rate for women with cervical cancer is reduced, in the USA the 5-year survival rate for all women with cervical cancer is 66% [2]. The main factor for high mortality rate is the asymptomatic characteristic of the condition in its initial stages[3] which promptly dignifies the need of a early diagnosis. Screening programs are implemented in most developed countries [4] and the process includes HPV test, cytology test (or Pap smear), colposcopy, and biopsy [1]. Human papillomavirus (HPV) screening is important because HPV is a group of viruses that has been proved to influence the risk of cancer [5].

A cervical cytology test is a test used to detect abnormal or potentially abnormal cells from the uterine cervix [6] and after that the conditions is further evaluated with colposcopy which consists in the use of a colposcope to capture images of the cervix in order to clarify inconclusive cytologic findings and give an assessment of the location, size, and extent of a lesion [7].

The colposcope is basically an image recorder device with a light source fixed to the body designed to capture images of the cervix. Posteriorly the images are analyzed by a professional to evaluate the severity of the lesion. There are even portable colposcopes such as Eva colpo [1].

It is also mentionable the fact that in low and middle-income countries HPV testing and cytologic testing are not widely available [5] [8]. Therefore the colposcopy could be an inexpensive method to diagnose cervical cancer in developed and specially in developing countries were the mortality rate is highly superior[8]. To transform this process into an economic reliable solution and to decrease the time spent on the image analyses, an automatic computer based diagnosis could be a solution.

1.2. Related Literature

Several methods have been implemented to analyze lesion in the cervix. Deep learning techniques have recently been the standard for state-of-art performances in cervix image analyses. In one of Xu et al. most recent works [9] a Convolution Neural Network (CNN) is used in the images to obtain a feature vector that is further united with non-image information to diagnose dysplasia. Using deep learning techniques Hu et al. [10] trained an algorithm based on Faster R-CNN with the goal of automatic detection of the cervix (segmentation) and prediction of cancer probability (classification) It is worth mention that data augmentation was applied in this work.

There are other works that rely mainly on feature ext ration followed by classification as well and constitute a large portion of research. Elayaraja P. et al. [11] used a Neural Network to classify the cervical image into normal and

[1] https://www.mobileodt.com/products/eva-colpo/?doing_wp_cron=1578935596.8403298854827880859375

abnormal by using Oriented Local Histogram to enhance edges, then the images are transformed to a multi resolution images to which features such as wavelet, Grey Level Co-occurrence Matrix (GLCM), moment invariant and Local Binary Pattern (LBP) are extracted and feed to the neural network. in 2018, Asiedu et al. [12] extracted colour and texture features, such as contrast, correlation, energy, homogeneity support and central tendencies channels of mean, median, mode, variance and Otsu threshold level in different color. Furthermore the features extracted were inputed in a vector machine model to classify cervigrams.

The work of Xu et al. [13, 14, 15, 16] also offers a variety of solutions. In [14] uses images and text information for classification and in [15] Xu et al. extracts features such as including Pyramid histogram in L*A*B* color space PLAB, Pyramid Histogram of Oriented Gradients PHOG, and Pyramid histogram of Local Binary Patterns PLBP and then uses adaboost algorithm to classify. In [13] the previous work is further extended by comparing seven classic machine-learning algorithms to differentiate images of high-risk patient visits from those of low-risk patient visits. In 2017 the paper [16] compares the hand-crafted pyramid features with CNN features for cervical disease classification. Xu et al. [9] also uses a Multimodal approach combining images and text information and feeding it to a CNN. Finally, Song et al. [17] integrates cervical images, Papa smear results, HPV test, and patient age in a Multimodel framework with improved results over the image based only.

1.3. Brief summary

The present work uses two different datasets discussed herein in section 3.1. A smaller one with labels describing the risk of cancer (high, low) and the larger one with different labels, not fitted for the problem. The smaller dataset also has incorporated non-image information regarding clinical data of the patient.

The goal is to use multitask and Multimodal learning to conjugate information from the large dataset and the non-image data but at the same time consider the smaller, main dataset with the pretended labels.

2. Methods

In this work all the used methods include two stages: train and test. In the training stage, the models are first pre-trained on the ImageNet dataset, and data pre-processing is applied on the cervical colposcopy datasets. Next, transfer learning is applied, whereby the pre-trained network parameters are used to initialize all the different classification models. The models are then fine-tuned. In the testing stage, the pre-processed testing images are fed into the fine-tuned models. The models score is obtained by comparison with the ground truth labels. Additional details are described below.

2.1. Data Pre-processing and Augmentation

First is worth mention that the dataset were divide in train, validation and test subsets. Proceeding further. the uniform size of the images is an essential requirement to use the images to feed the neural networks. This way, the resize of the images must be done to get 224 x 224 images. The last pre-processing that must be performed is the normalization which consists in a rescale of pixel values from the range of 0-255 to the range 0-1. The performance of deep learning algorithms is severely dependent of the quantity of data available to train the models. Thus, although deep learning algorithms can achieve much superior performance contrasted to traditional machine the demand of larger datasets is even higher. Data augmentation reduce the impact of the small databases used in deep learning. The main data set (NCI/NIH database) used in this work has a small number of labelled cervigrams (1,886), to overcome this problem and augment the dataset a series of random transformations are done in each training epoch to every single image. The transformations applied included width and height shift, image rotation with a range of 90, horizontal flip, zoom (in or out), and color saturation. Figure 1 shows three examples of random transformations. This augmentation is made online on the train subset.

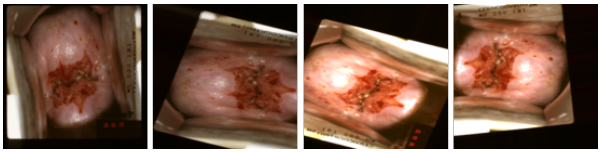


Figure 1. Example of Data Augmentation on the NCI/NHI database. The original image (left) and random transformations.

2.2. Neural Network Model

This project used a Neural Network (NN) [18] model divided into three different parts: First a convolutional neural network whose weights are shared by both datasets and a second and third part constituted of two different fully connected neural networks specific for each dataset used. The model is described in figure 2 and in the sections bellow.

2.2.1 Convolutional Neural Network

A convolutional neural network (ConvNet) is a deep learning algorithm [19] which is composed by numerous consecutive stages, namely convolutional (conv), activation functions, pooling (pool) layers, etc. ConvNet have learnable parameters (weights) that can be updated using matrix multiplication and its goal is to reduce the images into a form which is easier to process. The input of the ConvNet is the

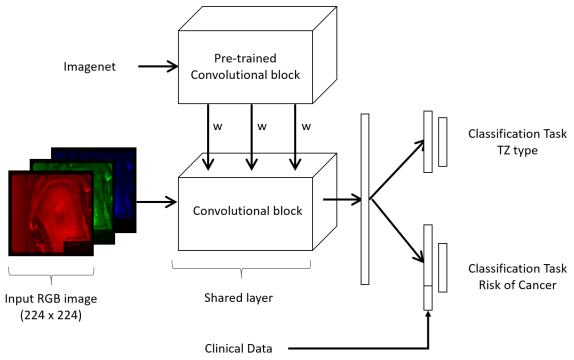


Figure 2. Multimodal and Multitask learning model using transfer learning with two tasks.

raw pixel intensity of the RGB images normalized to the scale of *0-1*.

•**Convolutional layer:** The first convolutional layer (conv) [19] takes local quadrilateral patches across the entire input image, in the consequent layers it gets feature maps as input, in this type of layers it is assumed local connectivity and space stationary, since instead of fully connected networks, there is a set of filters/kernels of optional shape that are convolved with the image.

•**Activation layer:** The activation layer exist because, simply explaining, the neurons do not have boundaries (-inf to +inf) therefore an activation layer is necessary to project the output value (Y). ReLu [20] was the activation function used.

•**Pooling layer:** The pooling operation (pool) is responsible for the down-sample in the feature maps by summarizing feature replies in each non-overlapping local patch. There are two types of pooling: Max Pooling and Average Pooling. Max Pooling has a great performance when compared with Average Pooling, once it can also work as a noise suppressant [21].

•**Batch Normalization:** BatchNorm normalizes the output of a previous activation layer. By subtracting the batch mean and dividing by the batch standard deviation it can prevent the amplification of small changes. The author of the paper batchnorm goes further to declare that it may help to prevent overfitting.

2.2.2 Fully Connected Networks

After the convolutional part it follows two different Fully Connected (FC) Networks that have as input the flatten feature map (into a feature vector) with origin in the convolutional output. Each Fully Connected Network is related to one task as seen in figure 2. The last layer of the two networks computes the classification probability for each class using softmax regression. The output layers are composed of several neurons each matching to the number of the classes that the network is classifying. For instance, existing two classes (high and low risk) there are two neurons/perceptrons in the last layer. It is also very important to notice that, for the tasks related with "risk of cancer", he clinical data present in the NCI/NIH dataset is concatenated into the feature vector before feeding it to the FC block. To reduce overfitting, "dropout" [22] is used to constrain the FC layers.

2.2.3 Network training and testing

To initialize training it also necessary to initialize the weights. They can be a random gaussian distribution or pre-defined values. In this case the Convolutional Network used is pre-trained on the ImageNet dataset [2] improve the performance of the classification algorithm, meaning the weights are initialized via Transfer learning (discussed in section 2.3). During training, these weights are iteratively optimized by minimizing the classification error (loss function) on the training set using the back-propagation algorithm [23]. The weights are updated using the gradients of the loss function, computed via stochastic gradient descent (SGD) over a mini-batch (size of 256) of training samples. The optimizer used was Adam [24] which automatically reduces the learning rate aperiodically along the epochs. The training process is terminated after a pre-determined number of epochs. The model with the lowest validation loss value is selected as the final network.

Five fold cross validation was also used by dividing into 5 different groups of training and validation subsets.

2.3. Transfer Learning

Transfer learning corresponds to the initialization of learning in a new task provided the knowledge of a previous, different task [25]. In this study, the first part of the model had its ConvNet pre-trained on the ImageNet classification dataset (ILSVRC), considering two different classes, which means that the model weights were not randomly initialized. After the mentioned step the model was fine-tuned in our dataset.

2.4. Multimodal Learning

Multimodal learning deals with different types of data (text, images, signals, etc) in order to help in the model development [26]. As mentioned herein in section 3.1 one of the datasets has also information about clinical data of the patient that is incorporated in the training, constituting a Multimodal learning model. The clinical data may contain information relevant for the decision and, as mentioned before, it is concatenated to the flatten feature map derived from the ConvNet block in the task related with the risk of cancer, as it is only present in the NCI/NIH dataset. In this

[2] http://www.image-net.org/to

work 4 different clinical features were included as an input in the classification task (age, time point, HPV test and worst histology diagnosis (Normal, abnormal; CIN2, CIN3, cancer)).

2.5. Multitask Learning

Multitask learning shares representations between similar tasks in order to retrieve information from both. It can be useful for generalization because the model learns from different tasks instead of being laser-focused on a single task [27]. For the problem at hands, because the dataset of interest was to small, it was used a second dataset with similar images, but with different labels. In figure 2 are visible two different classification tasks, each task is attributed to each dataset with its respective labels, however there is also a shared section that features the same weights for both tasks. The ConvNet section is a common branch whose weights are updated based on the classification error of both tasks which allows the model to learn features present in both datasets. The task associated with the NCI/NIH dataset corresponds to the classification of the risk of cancer (high or low) and the task related with the Intel MobileODT dataset is the classification of the transformation zone (type I, II or III - see figure A.1 in the appendix).

Although this is not a regular multitask problem, as it does not use the same dataset for the different tasks, it was taken in consideration the highly similar aspects of the two datasets and therefore can be treated as a multitask problem.

2.6. Losses

For the first mentioned task the loss used was the binary cross entropy displayed in eq. 1 and for the second task it was used categorical cross entropy (eq.l3).

N represents the number of classes, *y* is a binary indicator (0 or 1) that takes the value 1 if class label *c* is the correct classification for observation *i*, and *p* is the predicted probability of observation *i* to belong to class *c*.

$$L_{rc}(y,p) = -(y\log(p) + (1-y)\log(1-p)) \quad (1)$$

$$L_{tz}(y,p) = -\sum_{c=1}^{N} y_{i,c}\,(p_{i,c}) \quad (2)$$

$$L_{reg} = \sum_{i=1}^{N} |w_i| \quad (3)$$

$$L_{total} = (\alpha)(\omega)L_{rc} + (\mathbf{1}-\alpha)(\mathbf{1}-\omega)L_{tz} + L_{reg} \quad (4)$$

The lasso regression was also used for regularization in order to maintain the values of the weights as close to zero as possible (eq. 3). The total loss (eq.4) results from the weighted sums of the described losses, with $\alpha \in \{0,1\}$ being 0 or 1 according to the activated task being TZ or risk of cancer (rc) classification, respectively. Each loss (for the respective task) is also weighted with a larger weight attributed to the task performed by the NCI/NIH dataset which has the pretended labels, ωwas set to 0.7 and $\omega \in [0,1]$.

In resume for section 2.2, the first dataset passes through the pre-trained ConvNet block and follows for the FC block responsible for the classification task on the risk of cancer, where the clinical data is concatenated, and then, by back propagation, the weights of both blocks are updated. Afterwards the second dataset passes through the same ConvNet into the second FC block responsible for the transformation zone classification after which the weights are updated again.

3. Experimental details

3.1. Dataset

Two datasets were used:

•**The NCI/NIH dataset**: The American National Cancer Institute (NCI) from National Institutes of Health (NIH) collected a dataset composed by digital colposcopy images (cervigrams) from 10,000 women [27]. A subset of this dataset containing 2,120 cervigrams is available for technical works. Besides the images, the dataset includes information about the patient's age, HPV test, and histology results. Some of this data is missing, including the results from histologic examination, which should be the ground truth for cancer risk classification problems, resulting in a total of 913 labeled cervigrams. The annotations for histology results correspond to the neoplasia progression level, being categorized as normal, abnormal, CIN2, CIN3, and cancer. In colposcopy examination, each case is classified as high risk or low risk, to convert the categories of the dataset to this two classes, normal and abnormal cases are considered as low risk and CIN2, CIN3, and cancer cases are considered as high risk. This dataset can be used to train and validate decision support systems for cervical cancer and cervical intra-epithelial neoplasia diagnosis, being the dataset used in this work.

•**Intel & MobileODT dataset**: Intel & MobileODT submitted a dataset with about 8000 images for a Kaggle competition. The dataset covers the main colposcopy stages and has annotations about the cervix type regarding transformation zone. There are no annotation concerning neoplasia or risk of cancer, but it is considered that every images corresponds to normal cases.

3.2. NN Architecture

Several architectures were used in the convolutional block to achieve the best results.

•**VGG16** is a convolutional neural network first proposed by K. Simonyan et al. [28], and it is considered to be one of the excellent vision model architecture to date. The most unique thing about VGG16 is that instead of having a large amount of hyper-parameter they focused on having convolution layers of 3x3 filter with a stride 1 and always used same padding and maxpool layer of 2x2 and stride 2. It follows this arrangement of convolution and max pool layers consistently throughout the whole architecture. The full architecture can be seen in figure A.2.

•**MobileNet** uses depthwise separable convolutions and pointwise convolutions (1x1 convolutions) as seen in figure A.3, and has lightweight architectural features [29]. The overall architecture of the MobileNet is described on figure A.4.

•**ResNet50** or Residual Network, tackles the problem of vanishing gradients [30] by reusing activations from a previous layer. The authors explicitly reformulate the layers to learn the residual function with reference to the layer inputs, instead of learning unreferenced functions. It uses simple building block with skip connections (figure A.5) for that same reason. The overall architecture can be seen in figure A.6.

For every architecture the BatchNorm layer was added after each individual block (both ConvNet block and FC block) and droupout after each FC layer for regularization.

4. Results and Discussion

To evaluate the performance of the models several metrics were taken in consideration. In table 1 those metrics are elucidated and the values correspond to the mean ans standard deviation of the 5 best models extracted in the 5-fold cross validation. The model was tested for the three different architectures with best results for ResNet50 in all the metrics except precision and specificity. The ResNet50 is recent and very deep architecture and seems to be able to extract the most compelling features resulting in the best classification.

Going a step further and comparing the results with other works featuring Multimodal learning it is noticeable that the evaluation metrics are very close to the state-of-the-art results with special emphasis to the sensitivity which has the best value overall (94.92% for ResNet50) on the contrary to the specificity which is lower than expected with 83.15% being the best result using MobileNet versus 92.82% for Song et al.[17].

5. Conclusion and Future Work

The main goal of this work was to develop an accurate image-based algorithm to support a real-time medical decision for cervical cancer screening. The classification model must support medical decision during a colposcopy examination, however, the real diagnosis is only performed by biopsy. Thus the main archive for a screening model, its to minimize the number of false negatives. A false positive has a cost of an unnecessary biopsy, while a false negative can cost a human life. Even the accuracy and specificity metrics of our models are lower than the literature models as shown in table 1 the sensitivity and negative precision are higher than the best model found in the literature for cervical cancer screening using Multimodal information. The Multimodal ResNet50 shows the best results for the represented metrics however it was not able to outperform literature results. The limitation of data available and the imbalance classes were the greatest limitations of this work, this way comparing this work with the literature it is not a fair comparison because the state-of-art works had access to the entire NCI data set and no only a subset. There is a lot to improve in this classification models, for future work a larger data set will be a good starting point for better results. However, according to gynaecologists, taking a decision based on a single image is a tricky task that might lead to misdiagnosed cases. For future work several parameters may be fine-tuned and different variations, for instance, not using Multimodal learning may be applied.

References

[1] World Health Organization. Cervical cancer. https://www.who.int/cancer/prevention/diagnosis-screening/cervical-cancer/en/, 2019. (Accessed: 2020-01-11).

[2] cancer.net. Cervical cancer: Statistics. https://www.cancer.net/cancer-types/cervical-cancer/statistics, 2019. (Accessed: 2020-01-11).

[3] NHS. Overview-cervical cancer. https://www.nhs.uk/conditions/cervical-cancer/, 2019. (Accessed: 2020-01-11).

[4] World Health Organization. Cervical cancer screening in developing countries. Technical report, World Health Organization, 2002.

[5] World Health Organization. Human papillomavirus (hpv) and cervical cancer. https://www.who.int/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer, 2019. (Accessed: 2020-01-11).

[6] Lab Test Online uk. Cervical cytology. https://labtestsonline.org.uk/tests/cervical-cytology, 2019. (Accessed: 2020-01-11).

Table 1. Evaluation performance on several metrics for the 3 used architectures. It was also compared the NN model using Multimodal and Multitask learning described in this paper with state-of-the-art works related with Multimodal learning of Song et al.[17] and Xu et al. [17].

			Accuracy %	Balanced acc. %	AUC %	Sensitivity %	Specificity %	Precision %	Negative Precision %
Multimodal + Multitask	*MobileNet*	**Mean**	84.76	87	92.97	91.61	83.15	57.03	97.56
		STD	4.59	2.97	3.11	3.50	5.83	10.52	1.30
	ResNet50	**Mean**	85.10	89.02	92.97	**94.92**	83.12	56.92	98.18
		STD	2.87	2.79	3.11	6.03	3.32	11.00	2.40
	VGG16	**Mean**	83.69	86.30	92.97	90.86	81.74	56.12	97.18
		STD	4.59	5.11	3.11	7.92	5.67	3.88	2.49
Literature	*Xu et al.*		**88.91**	-	**94.00**	87.83	90.00	-	-
	Song et al.		87.79	-	-	82.79	**92.82**	-	-

[7] Maureen L Harmon and Kumarasen Cooper. *Gynecologic Pathologylo1995artificial*. Elsevier, 2005.

[8] World Health Organization. Un joint global programme on cervical cancer prevention and control. `https://www.who.int/ncds/un-task-force/un-joint-action-cervical-cancer-leaflet.pdf?ua=1`, 2016. (Accessed: 2020-01-11).

[9] Tao Xu, Han Zhang, Xiaolei Huang, Shaoting Zhang, and Dimitris N Metaxas. Multimodal deep learning for cervical dysplasia diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 115–123. Springer, 2016.

[10] Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S Jaiswal, Brian Befano, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *Obstetrical & Gynecological Survey*, 74(6):343–344, 2019.

[11] P Elayaraja and M Suganthi. Automatic approach for cervical cancer detection and segmentation using neural network classifier. *Asian Pacific Journal of Cancer Prevention: APJCP*, 19(12):3571, 2018.

[12] Mercy Nyamewaa Asiedu, Anish Simhal, Usamah Chaudhary, Jenna L Mueller, Christopher T Lam, John W Schmitt, Gino Venegas, Guillermo Sapiro, and Nirmala Ramanujam. Development of algorithms for automated detection of cervical pre-cancers with a low-cost, point-of-care, pocket colposcope. *IEEE Transactions on Biomedical Engineering*, 2018.

[13] Tao Xu, Cheng Xin, L Rodney Long, Sameer Antani, Zhiyun Xue, Edward Kim, and Xiaolei Huang. A new image data set and benchmark for cervical dysplasia classification evaluation. In *International Workshop on Machine Learning in Medical Imaging*, pages 26–35. Springer, 2015.

[14] Tao Xu, Xiaolei Huang, Edward Kim, L Rodney Long, and Sameer Antani. Multi-test cervical cancer diagnosis with missing data estimation. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140X. International Society for Optics and Photonics, 2015.

[15] Tao Xu, Edward Kim, and Xiaolei Huang. Adjustable adaboost classifier and pyramid features for image-based cervical cancer diagnosis. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 281–285. IEEE, 2015.

[16] Tao Xu, Han Zhang, Cheng Xin, Edward Kim, L Rodney Long, Zhiyun Xue, Sameer Antani, and Xiaolei Huang. Multi-feature based benchmark for cervical dysplasia classification evaluation. *Pattern recognition*, 63:468–475, 2017.

[17] Dezhao Song, Edward Kim, Xiaolei Huang, Joseph Patruno, Héctor Muñoz-Avila, Jeff Heflin, L Rodney Long, and Sameer Antani. Multimodal entity coreference for cervical dysplasia diagnosis. *IEEE transactions on medical imaging*, 34(1):229–245, 2014.

[18] Tony Yiu. Understanding neural networks. `https://towardsdatascience.com/understanding-neural-networks-19020b758230`, 2019. (Accessed: 2020-01-11).

[19] Shih-Chung B Lo, Heang-Ping Chan, Jyh-Shyan Lin, Huai Li, Matthew T Freedman, and Seong K Mun. Artificial convolution neural network for medical image pattern recognition. *Neural networks*, 8(7-8):1201–1214, 1995.

[20] Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

[21] Víctor Suárez-Paniagua and Isabel Segura-Bedmar. Evaluation of pooling operations in convolutional architectures for drug-drug interaction extraction. *BMC bioinformatics*, 19(8):209, 2018.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[23] Yann A. LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller. *Efficient BackProp*, pages 9–48. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[26] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.

[27] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2014.

[29] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[31] Joe Jordan, Marc Arbyn, Pierre Martin-Hirsch, Ulrich Schenck, J-J Baldauf, Daniel Silva, A Anttila, Pekka Nieminen, and W Prendiville. European guidelines for quality assurance in cervical cancer screening: Recommendations for clinical management of abnormal cervical cytology, part 1. *Cytopathology : official journal of the British Society for Clinical Cytology*, 19:342–54, 01 2009.

[32] Bibo Shi, Rui Hou, Maciej Mazurowski, Lars Grimm, Yinhao Ren, Jeffrey Marks, Lorraine King, Carlo Maley, E. Hwang, and Joseph Lo. Learning better deep features for the prediction of occult invasive disease in ductal carcinoma in situ through transfer learning. page 98, 02 2018.

Appendix

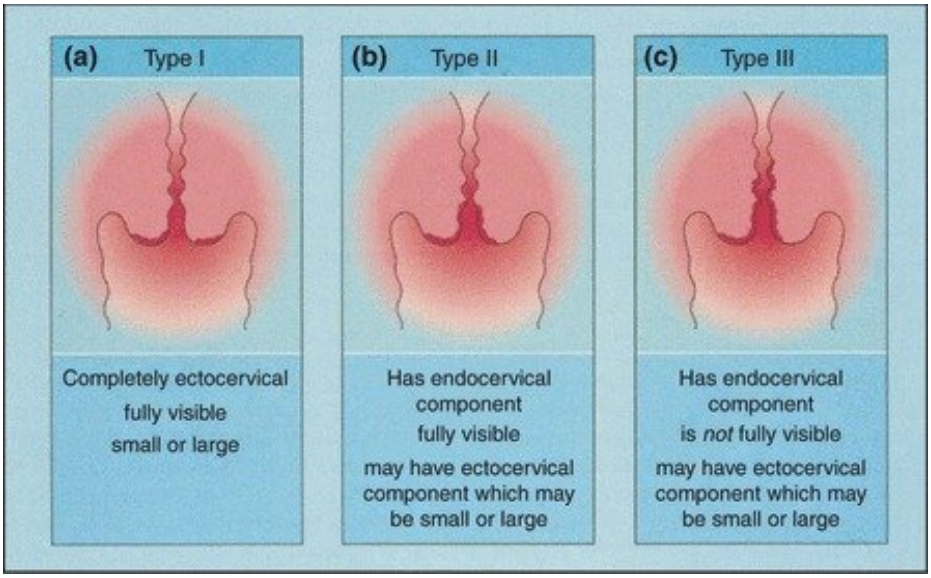


Figure A.1. Types of transformation zones in the cervix [31].

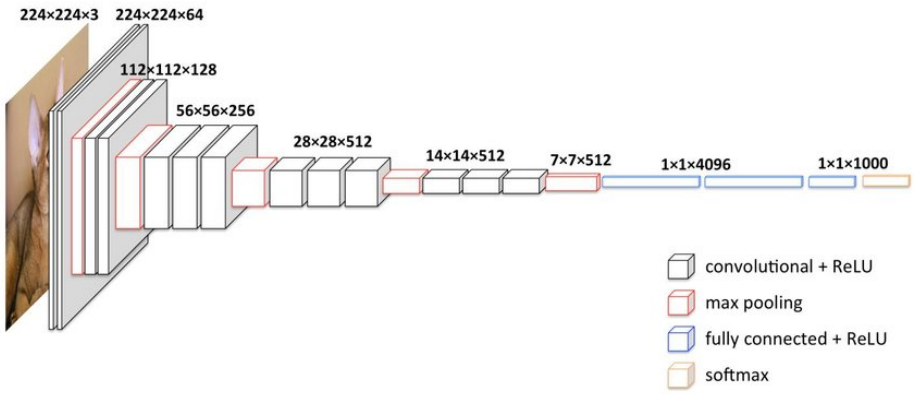
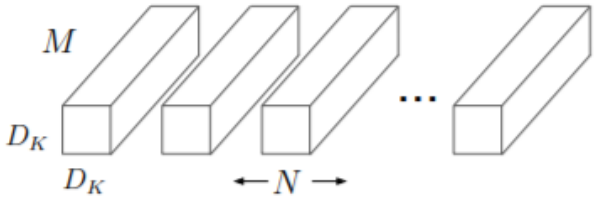
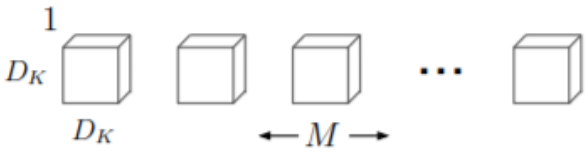


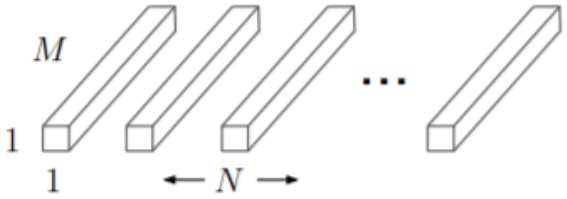
Figure A.2. Architecture of VGG16 [32].



(a) Standard Convolution Filters



(b) Depthwise Convolutional Filters



(c) 1×1 Convolutional Filters called Pointwise Convolution in the context of Depthwise Separable Convolution

Figure A.3. Depthwise and pointwise layers for MobileNet [29].

Type / Stride	Filter Shape	Input Size
Conv / s2	$3 \times 3 \times 3 \times 32$	$224 \times 224 \times 3$
Conv dw / s1	$3 \times 3 \times 32$ dw	$112 \times 112 \times 32$
Conv / s1	$1 \times 1 \times 32 \times 64$	$112 \times 112 \times 32$
Conv dw / s2	$3 \times 3 \times 64$ dw	$112 \times 112 \times 64$
Conv / s1	$1 \times 1 \times 64 \times 128$	$56 \times 56 \times 64$
Conv dw / s1	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 128$	$56 \times 56 \times 128$
Conv dw / s2	$3 \times 3 \times 128$ dw	$56 \times 56 \times 128$
Conv / s1	$1 \times 1 \times 128 \times 256$	$28 \times 28 \times 128$
Conv dw / s1	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 256$	$28 \times 28 \times 256$
Conv dw / s2	$3 \times 3 \times 256$ dw	$28 \times 28 \times 256$
Conv / s1	$1 \times 1 \times 256 \times 512$	$14 \times 14 \times 256$
$5\times$ Conv dw / s1	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
$5\times$ Conv / s1	$1 \times 1 \times 512 \times 512$	$14 \times 14 \times 512$
Conv dw / s2	$3 \times 3 \times 512$ dw	$14 \times 14 \times 512$
Conv / s1	$1 \times 1 \times 512 \times 1024$	$7 \times 7 \times 512$
Conv dw / s2	$3 \times 3 \times 1024$ dw	$7 \times 7 \times 1024$
Conv / s1	$1 \times 1 \times 1024 \times 1024$	$7 \times 7 \times 1024$
Avg Pool / s1	Pool 7×7	$7 \times 7 \times 1024$
FC / s1	1024×1000	$1 \times 1 \times 1024$
Softmax / s1	Classifier	$1 \times 1 \times 1000$

Figure A.4. Arquitechture of MobileNet [29].

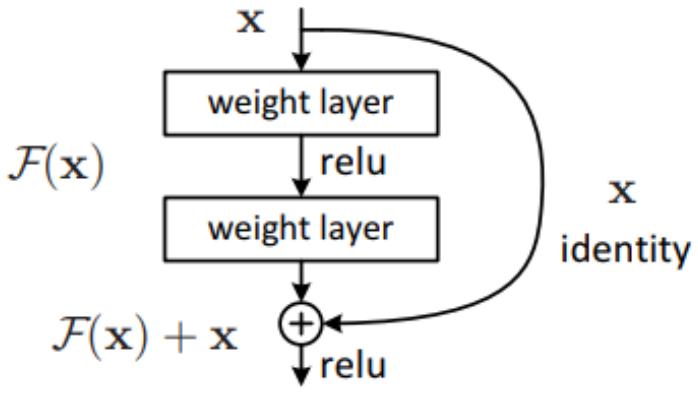


Figure A.5. Building block of ResNet featuring a skip connection [30].

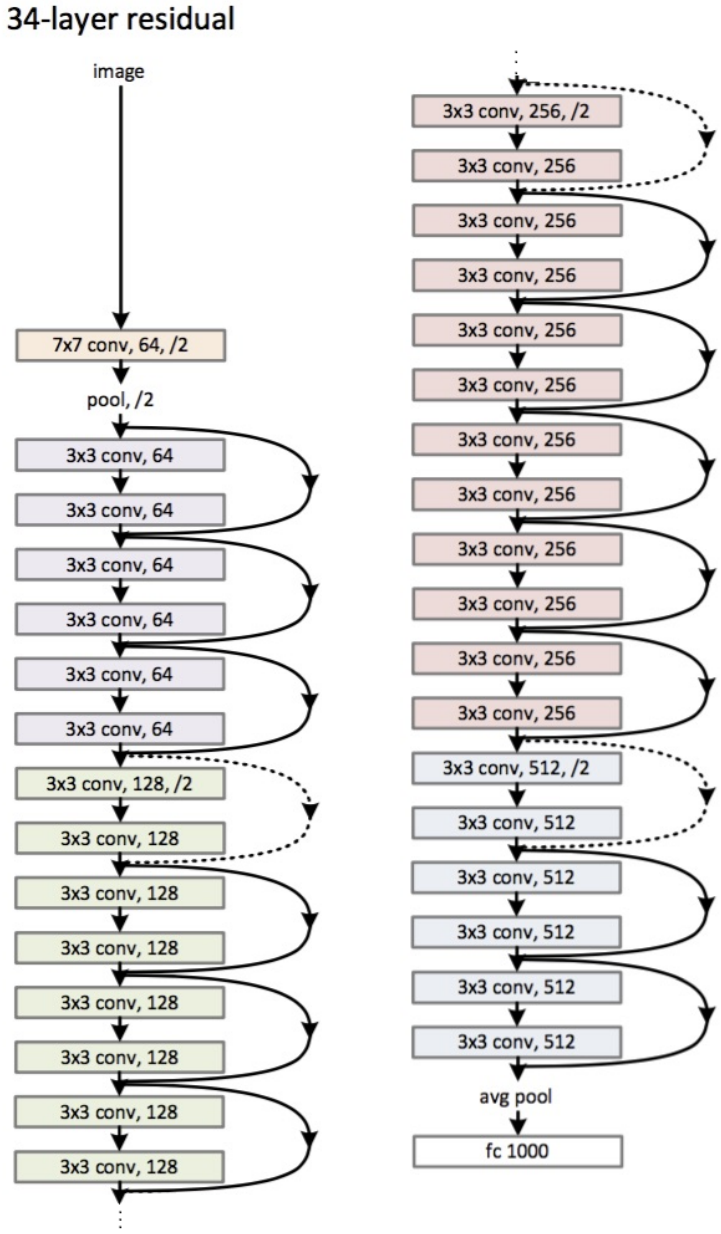


Figure A.6. Overall architecture of ResNet50 [30].