Multimodal Deep Learning for Cervical Dysplasia Diagnosis

Tao Xu[1⋆], Han Zhang[2⋆], Xiaolei Huang[1], Shaoting Zhang[3], and Dimitris N. Metaxas[2]

[1] Computer Science and Engineering Department, Lehigh University, Bethlehem, PA, USA
[2] Department of Computer Science, Rutgers University, Piscataway, NJ, USA
[3] Department of Computer Science, UNC Charlotte, Charlotte, NC, USA

Abstract. To improve the diagnostic accuracy of cervical dysplasia, it is important to fuse multimodal information collected during a patient's screening visit. However, current multimodal frameworks suffer from low sensitivity at high specificity levels, due to their limitations in learning correlations among highly heterogeneous modalities. In this paper, we design a deep learning framework for cervical dysplasia diagnosis by leveraging multimodal information. We first employ the convolutional neural network (CNN) to convert the low-level image data into a feature vector fusible with other non-image modalities. We then jointly learn the non-linear correlations among all modalities in a deep neural network. Our multimodal framework is an end-to-end deep network which can learn better complementary features from the image and non-image modalities. It automatically gives the final diagnosis for cervical dysplasia with 87.83% sensitivity at 90% specificity on a large dataset, which significantly outperforms methods using any single source of information alone and previous multimodal frameworks.

1 Introduction

Cervical cancer ranks as the second most common type of cancer in women aged 15 to 44 years worldwide [13]. Screening can help prevent cervical cancer by detecting cervical intraepithelial neoplasia (CIN), which is the potentially precancerous change and abnormal growth of squamous cells on the surface of the cervix. According to the World Health Organization (WHO) [13], CIN has three grades: CIN1 (mild), CIN2 (moderate), and CIN3 (severe). Mild dysplasia in CIN1 only needs conservative observation while lesions in CIN2/3 or cancer (denoted as CIN2+ in this paper) require treatment. In clinical practice one important goal of screening is to differentiate normal/CIN1 from CIN2+ for early detection of cervical cancer.

Widely used cervical cancer screening methods today include Pap tests, HPV tests, and visual examination. Pap tests are effective, but they often suffer from low sensitivity in detecting CIN 2+ [10]. HPV tests are often used in conjunction with Pap tests, because nearly all cases of cervical cancer are caused by Human papillomavirus (HPV) infection. Digital cervicography is a non-invasive and low-cost visual examination method that takes a photograph of the cervix (called a Cervigram®) after the application of 5% acetic acid to the cervix epithelium. Recently, the automated Cervigram analysis techniques [14, 10] have shown great potential for CIN classification.

⋆ Indicates equal contribution

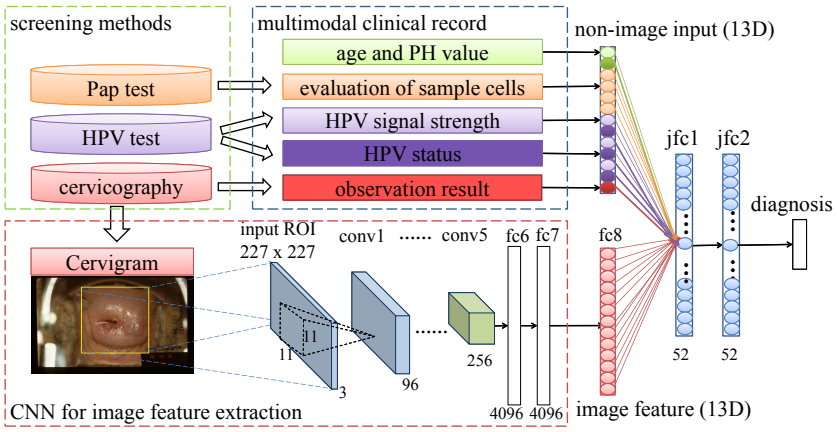

Fig. 1. Our multimodal deep network: (1) we apply a convolutional neural network (CNN) to learn image features from raw data in Cervigram ROIs; (2) we use joint fully connected (jfc) layers to model the non-linear correlations across all sources of information for CIN classification.

Previous works [3, 1, 10, 14] have shown that multimodal information from conventional screening tests can provide complementary information to improve the diagnostic accuracy of cervical dysplasia. DeSantis et al. [3] combined spectroscopic image information measured from the cervix with other patient data, such as Pap results. Chang et al. [1] investigated the diagnostic potential of different combinations of reflectance and fluorescence spectral features. In [10, 14], the authors hand-crafted pyramid histograms of color and oriented gradient features to represent Cervigram and directly utilized the clinical results to represent non-image modalities. Then either Support Vector Machine (SVM) [14] or k-nearest neighbor (K-NN) [10] was used to calculate the decision score for each group of modalities separately. The final decision was simply made by combining decision scores in all the modalities. Since those previous methods integrated multimodal information at the final stage, they did not fully exploit the inherent correlations across image and non-image modalities. Another limitation is that their hand-crafted features may require strong domain knowledge and it is difficult to manually design proper features that are fusible across different modalities.

Recently, deep learning has been exploited in medical image analysis to achieve state-of-the-art results [2, 12, 9, 8]. Besides learning data representations just from a single modality, deep learning is also able to discover the intricate structure in the multimodal datasets (e.g., video and audio) and improve the performance of the corresponding tasks [7]. However, this attractive feature is less well investigated in the medical domain. A pioneering work in this direction, Suk et al. [11] applied multimodal Deep Boltzmann Machine (DBM) to learn a unified representation from the paired patches of PET and MRI for AD/MCI diagnosis. However, considering that PET and MRI are both image modalities, this could be less complicated compared to the dilemma we face. Particularly, in a patient's medical record, the data is more heterogeneous. The raw medical image data is a high dimensional vector. It requires less human labor to obtain but it contains a large amount of undiscovered information. The clinical results

that are verified by clinicians have less feature dimensions, but they usually provide more instructional information. Therefore, it is challenging to combine the information from these modalities to perform improved diagnosis.

In this paper, we apply deep learning for the task of cervical dysplasia diagnosis using multimodal information collected during a patient's screening visit. The contribution is threefold. (1) We solve the challenging problem of highly heterogeneous input data by converting the low-level Cervigram data into a feature fusible with other non-image modalities using convolutional neural networks (CNN). (2) We propose a deep neural network to jointly learn the non-linear correlations among all modalities. (3) We unify the CNN image processing network and the joint learning network into an end-to-end framework which gives the final diagnosis for cervical dysplasia with 87.83% sensitivity at 90% specificity on the test dataset. The proposed multimodal network significantly outperforms methods using any single source of information alone and previous multimodal frameworks.

2 Our Approach

In our dataset, every screening visit of the patient has at least one Cervigram and the clinical results of Pap tests and HPV tests. As shown in Fig. 1, we use Cervigram as low-level image input. Motivated by the work of Song et al. [10], we construct a 13D high-level non-image input using four Pap test results (e.g., Cytyc ThinPrep), three pairs of HPV test results (e.g., high risk HPV 16 and HPV18), one Cervigram observation result, PH value and age of the patient. Not every visit has a complete set of clinical results for all Pap and HPV tests. Thus, for our non-image feature vector, we compute the average value of each dimension using available data of that dimension in the training dataset to estimate the missing value. This imputation method is widely used in training deep networks since it actually ignores the missing dimension after whitening the data. Next, we will describe each component of our proposed multimodal deep network.

2.1 Learning a deep representation for Cervigram

Inspired by the recent success of convolutional neural networks (CNN) in general recognition tasks [6], instead of hand-crafting features [14, 10], we propose to use a CNN to learn visual features directly from Cervigram images.

Fine-tune pre-trained model: We use AlexNet [6] as our network structure for the feature learning. This model contains five convolutional layers (conv1-conv5) and two fully connected layers (fc6 and fc7) and a final 1000-way softmax layer. Since our Cervigram dataset is relatively small compared to the general image classification datasets, we follow the transfer learning scheme to train our model. We first take the model which is pre-trained on ImageNet classification task and replace its output layer by a new 2-way softmax layer. Then, we fine-tune its parameters on our Cervigram dataset. We detect one cervix region of interest (ROI) for each Cervigram using the method proposed in [10]. Every ROI region is fed into the network which outputs the corresponding feature vector from its last fully connected layer (fc7). Since there are 4096 hidden units in the fc7 layer, we get a 4096D image feature embedding.

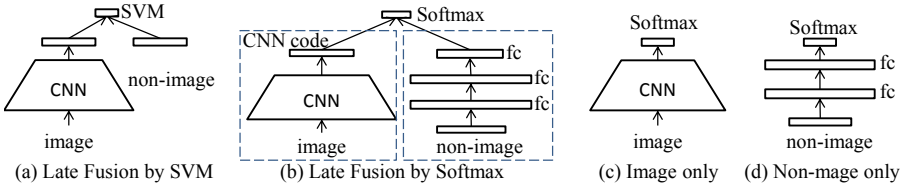

Fig. 2. Baseline methods.

CNN feature compression: The dimension of the CNN feature vector from fc7 layer is much higher than that of the non-image feature. Our experimental result shows that the high dimensional image feature can overwhelm the low dimensional non-image feature if we fuse them directly. Thus, we add another fully connected layer (fc8) with 13 units on top of fc7 to reduce the dimension of CNN feature to be comparable with non-image data. Thus, our image feature is non-linearly compressed to 13D.

2.2 Novel method for fusing multimodal information in a deep network

Increasing evidence shows that cues from different modalities can provide complementary information in cervical dysplasia diagnosis [14, 10, 3, 1]. However, it is challenging to integrate highly heterogeneous modalities. To motivate our multimodal deep network, we first discuss two simple fusion models and their drawbacks.

Baseline models: The previous multimodal methods [14, 10] assumed that the image and non-image data should be treated separately to make the prediction by themselves. The fusion between image and non-image modalities only happened when merging the decision scores from each modality. We call this type of fusion as Late Fusion. Based on this assumption, we will compare our proposed fusion model with two baseline late fusion frameworks. In the first one (Fig. 2a), the 13D CNN feature and the 13D non-image feature are directly concatenated and fed into a linear SVM for CIN classification. It is an intuitive approach without any feature learning or engineering in the non-image modalities. In the second baseline model (Fig. 2b), we simulate the feature learning strategy in the image modality to use a neural network to learn the features in non-image modalities and then combine them with CNN features for the final classification using softmax. In this setting, the hidden units in the deep neural networks are only modeling the correlations within each group of modalities.

Our model: Instead of using the above assumption, we assume that the data in the different modalities have a tighter correlation. For example, visual features (e.g., acetowhite epithelium) in the image can be treated as a complementary support of positive HPV or Pap. Therefore, those correlations can be used as a better representation to improve the classification accuracy. However, hand-engineering such complementary pairs is difficult and time-consuming. It is better to learn such correlations directly from the multimodal data. Therefore, we propose an early fusion framework to use deep neural networks to learn the highly non-linear correlations across all the modalities. As shown in Fig. 1, the 13D image feature and the 13D non-image feature (e.g., clinical results) are concatenated at an early stage and followed by joint learning layers.

To solve the problem that data in different modalities have different statistical properties, we applied the batch normalization (BN) transform [5] to fix the means and variances of the input in each modality. Given the input $x_1, x_2, ...x_m$ over a mini-batch, the output $\hat{x}_i$ is calculated as:

$$\hat{x}_i = \gamma \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta \tag{1}$$

where γ, β are the parameters to be learned by the network, $\mu = \frac{1}{m}\sum_{i=1}^{m} x_i$ and $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu)^2$. The batch normalization can regularize the model and allow us to have higher learning rates. Thus, we also apply it to joint fully connected layers.

In our network, the joint fully connected layer (jfc) is applied to learn the correlations across different modalities. Each node in the jfc layer is computed by Eq. 2,

$$\mathbf{z}^k = f(\mathbf{W}^k \hat{\mathbf{x}}^{k-1} + b^k) \tag{2}$$

where $\mathbf{z}^k$ indicate the activations in the k^{th} layer; $\mathbf{W}^k$ and b^k are weights and bias learned for the k^{th} layer; $\hat{\mathbf{x}}_{k-1}$ are the normalized output of the previous layer; $f(x) = \max(0, x)$ is the ReLU activation function. Compared to the previous framework [3, 1, 10, 14], the units in our jfc layers model the non-linear correlations across modalities. Also the output of jfc can be viewed as a better representation for the multimodal data.

Finally, a 2-way softmax (Eq. 3) layer is added upon the last joint fully connected layer to predict the diagnosis.

$$p(c = j|\hat{\mathbf{x}}; \mathbf{W}, b) = \frac{exp(\mathbf{W}_j \hat{\mathbf{x}} + b_j)}{exp(\sum_{l=0}^{1} \mathbf{W}_l \hat{\mathbf{x}} + b_l)} \tag{3}$$

where $p(c = j)$ indicates the probability of the input data belonging to the j^{th} category, here $j \in [0, 1]$; $\hat{\mathbf{x}}$ is the normalized output of the last joint fully connected layer; $\mathbf{W}$ and b are weights and bias learned for the softmax layer.

During the training process, we compute the cross-entropy loss and apply stochastic gradient descent (SGD) to train the whole network. The classification loss can also backpropagate to the image CNN layers to guide the CNN network to extract visual features that complement clinical features for the classification. The number of joint fully connected layers and the number of hidden units in each jfc layer are the hyper-parameters, and we choose them through cross validation.

3 Experiments

We evaluate our method on a dataset built from a large data archive collected by the National Cancer Institute (NCI) from 10,000 anonymized women in the Guanacaste project [4]. Each patient typically had multiple visits at different ages. During each visit, multiple cervical screening tests were performed. Since the Guanacaste project is a population-based study, only a small proportion of patient visits have the Worst Histology results: multiple expert histology interpretations were done on each biopsy taken during a visit; the most severe interpretation is labeled the Worst Histology and

serves as the "gold standard" ground truth for that visit in the database. From those labeled visits, we randomly sample 345 positive visits (CIN2+) and 345 negative visits (normal / CIN1) to build our visit-level dataset. And we use the same three-round three-fold cross validation to evaluate the proposed method and compare it with baseline models and previous works [14, 10, 3, 1].

Hyper-parameters of the proposed method: For our proposed models with different hyper-parameters, we compare their overall performance in Fig. 3. Their accuracy and sensitivity at high (90% and 92%) specificity are also listed in Table 1.

We first evaluate the importance of our feature compression layer (fc8) by comparing models "4096D-image+non-image" and "13D-image+non-image". In the former model, we directly concatenate the 4096 CNN feature from fine-tuned AlexNet with the 13D non-image feature and feed them into the softmax for the final classification. In the latter one, we first compress the 4096 CNN feature into 13 dimensions using an additional hidden layer, and then perform the concatenation and classification. The result shows that "13D-image+non-image" significantly outperforms "4096D-image+non-image", especially at high specificity. For example, the sensitivity is increased by about 10% at both 90% and 92% specificity using the compressed "13D-image+non-image". The reason is that the high dimensional image feature overwhelms the low dimensional non-image feature in "4096D-image+non-image".

We can further improve our model by adding joint fully connected (jfc) layers. After trying different depth (the number of jfc layers) and width (the number of units in each jfc layer), we get our best model "13D-image+non-image,2jfc". It has two jfc layers and each of them has 52 units. This deeper model achieves a better overall performance with another over 10% sensitivity increment at 90% and 92% specificity. It indicates that the information in image and non-image modalities needs to be jointly learned and non-linearly transformed in a deeper network. To test the effectiveness of our batch normalization, we remove the batch normalization transform in our best model. The new model "13D-image+non-image,2jfc(noBN)" has a decreased AUC 91.61% (in comparison to AUC 94% with BN). Thus it is important to use batch normalization to fix the means and variances to input in each modality and regularize the model. To conclude, our "13D-image+non-image,2jfc" with BN model gives the best performance with 88.91% accuracy and 87.83% sensitivity at 90% specificity. In the following experiments, we use this model for comparison.

Comparison with previous works: For fair comparison, we search the best hyper-parameters for alternative CIN classification methods shown in Fig. 2. We report the results of their best models as baselines in Fig. 4.

We first compare our model with the methods using image only or non-image only. From Fig. 4, it is clear that our model achieves a significant improvement over using any single group of modalities (our 94% AUC vs. image-only 88.77% and non-image only 86.06%). The sensitivity of our method is 38.26% higher than "non-image only" and 21.74% higher than "image only" at 90% specificity. It demonstrates the importance of fusing raw Cervigram information with other non-image modalities (e.g., Pap and HPV results) for cervical dysplasia diagnosis.

Fig. 4 also shows the comparison results of our early fusion method with "Late Fusion by SVM" and "Late Fusion by Softmax". The best model of "Late Fusion by

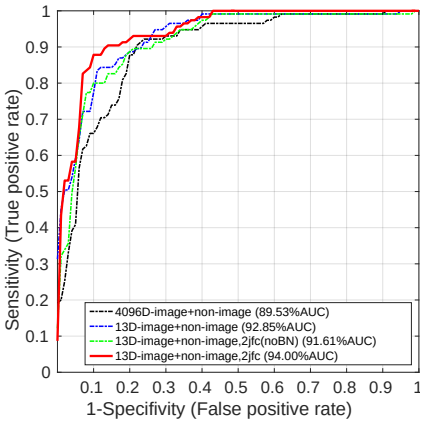

Fig. 3. Our models with different hyper-parameters (please view in color)

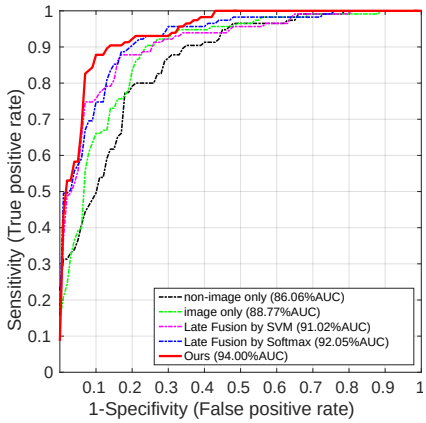

Fig. 4. Comparison with baseline methods (please view in color)

Table 1. Results of the proposed models with different hyper-parameters. ("noBN" indicates that batch normalization is not used in joint fully connected (jfc) layers.)

Model	AUC(%)	At 90% specificity		At 92% specificity	
		accu(%)	sensi(%)	accu(%)	sensi(%)
4096D-image+non-image	89.53	78.04	66.09	77.30	62.61
13D-image+non-image	92.85	83.70	77.39	82.09	72.17
13D-image+non-image,2jfc(noBN)	91.61	85.00	80.00	84.70	77.39
13D-image+non-image,2jfc	**94.00**	**88.91**	**87.83**	**87.74**	**83.48**

Softmax" has two hidden layers with 104 units in each layer. Our method outperforms the best results of both late fusion methods, especially at high specificity region. For instance, our model achieves more than 10% higher sensitivity at 90% specificity than both of them. This comparison result proves our assumption that the information in image and non-image modalities has a tighter correlation and the proposed "Early Fusion" assumption is better than the "Late Fusion" assumption used in [14, 10].

Our model achieves 88.91% accuracy and 87.83% sensitivity at 90% specificity. Compared with other previous multimodal methods [14, 10, 3, 1], ours is state-of-the-art in terms of visit level classification. For example, the method by DeSantis et al. [3] only achieved an accuracy of 71.3% and the approach in [1] gave 82.39% accuracy. Two previous works [14, 10] utilized the same multimodal information as ours. The work in [14] performed visit level classification. However, our performance is much better than theirs (our 88.91% accuracy vs. their 79.68%). Song et al. [10] utilized patient-level (multiple visits) information and achieved similar performance as ours (their 89.00% accuracy vs. our 88.91%). However, their patient-level method could not tell which visit of the patient was diagnosed as high risk (i.e., CIN2+).

4 Conclusions

In this paper, we propose a multimodal deep network for the task of cervical dysplasia diagnosis. We integrate highly heterogeneous data collected during a patient's screening visit by expanding conventional CNN structure with joint fully connected layers. The proposed model can learn better complementary features for the image and non-image modalities through backpropagation. It automatically gives the final diagnosis for cervical dysplasia with 87.83% sensitivity at 90% specificity on a large dataset, which is the state-of-the-art performance in visit level classification.

References

1. Chang, S.K., Mirabal, Y.N., et al.: Combined reflectance and fluorescence spectroscopy for in vivo detection of cervical pre-cancer. J. Biomedical Optics 10(2), 024–031 (2005)
2. Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 411–418. Springer, Heidelberg (2013)
3. DeSantis, T., Chakhtoura, N., Twiggs, L., Ferris, D., Lashgari, M., et al.: Spectroscopic imaging as a triage test for cervical disease: a prospective multicenter clinical trial. J. Lower Genital Tract Disease 11(1), 18–24 (2007)
4. Herrero, R., Schiffman, M., Bratti, C., et al.: Design and methods of a population-based natural history study of cervical neoplasia in a rural province of costa rica: the guanacaste project. Rev Panam Salud Publica 1, 362–375 (1997)
5. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1106–1114 (2012)
7. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. pp. 689–696 (2011)
8. Roth, H.R., Lu, L., Seff, A., Cherry, K.M., Hoffman, J., et al.: A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Golland, P., Hata, N., Barillot, C., Hornegger, J., Howe, R. (eds.) MICCAI 2014, Part I. LNCS, vol. 8673, pp. 520–527. Springer, Heidelberg (2014)
9. Shin, H., Orton, M., Collins, D.J., Doran, S.J., Leach, M.O.: Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. TPAMI 35(8), 1930–1943 (2013)
10. Song, D., Kim, E., Huang, X., Patruno, J., Munoz-Avila, H., Heflin, J., Long, L., Antani, S.: Multi-modal entity coreference for cervical dysplasia diagnosis. TMI 34(1), 229–245 (2015)
11. Suk, H., Lee, S., Shen, D.: Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. NeuroImage 101, 569–582 (2014)
12. Suk, H., Shen, D.: Deep learning-based feature representation for AD/MCI classification. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 583–590. Springer, Heidelberg (2013)
13. WHO: Human papillomavirus and related cancers in the world. In: Summary report. ICO Information Centre on HPV and Cancer (Aug 2014)
14. Xu, T., Huang, X., Kim, E., , Long, L., Antani, S.: Multi-test cervical cancer diagnosis with missing data estimation. In: SPIE Medical Imaging. pp. 94140X–94140X–8 (2015)