2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Learning a Triplet Embedding Distance to Represent Gleason Patterns

Fabian León* and Fabio Martínez*

Abstract— **Gleason grade stratification is the main histological standard to determine the severity and progression of prostate cancer. Nonetheless, there is a high variability on disease diagnosis among expert pathologists (kappa lower than 0.44). End-to-end deep representations have recently deal with the automatic classification of Gleason grades, where each grade is limited to namely code high-visual-variability sharing patterns among classes. Such limitation on models may be attributed to the relatively few labels to train the representation, as well as, to the natural imbalanced sets, available in clinical scenarios. To overcome such limitation, this work introduces a new embedding representation that learns intra and inter-Gleason relationships from more challenging class samples (grades tree and fourth). The proposed strategy implements a triplet loss scheme building a hidden embedding space that correctly differentiates close Gleason levels. The proposed approach shows promising results achieving an average accuracy of 74% to differentiate between degrees three and four. For classification of all degrees, the proposed approach achieves an average accuracy of 62%.**

Clinical relevance— **The proposed approach properly model intra and inter variability of visual regions correlated with Gleason score. Such fact may contribute to support cancer aggressiveness stratification.**

I. Introduction

Prostate cancer is the second most common cancer, reporting around 1,276,000 new cases each year and more than 358,989 deaths [1]. The patients survival depends entirely on early and correct disease diagnosis to select the best treatment [2]. However, this procedure is tedious and depends entirely on experts, introducing a high variation in diagnosis. The Gleason grading is the main system for quantifying and characterizing cancer from histological images, analyzing, among others, glandular structures, color and general spatial arrangement [3], [4]. In general, the Gleason grading system is divided into two scales: from (1-5) to characterize local regions, and from (6-10) to report a general sample analysis. In the first scale (1-5), the primary levels present well-formed glands, while the last stages do not present glands, instead, there are solid nets and necrosis as show in Figure 1. The high scale (6-10) is reported as the sum of the two most predominant patterns present in the sample. For example, a diagnosis of grading 7 is the result of finding grades 3 and 4 in the sample (3 + 4 or 4 + 3) [5].

Despite the effectiveness of the Gleason system, high diagnosis variations are shown between pathologists, mainly due to challenging visual patterns. Different studies have reported this inter pathologist variation. For example, in [6], a study including eight pathologists reported a kappa value of 0.68, in the diagnosis of 81 prostate slides. In [7] 150 slides were sent to 3 pathologists in two stages, reporting a kappa value of 0.25 and 0.52, respectively, showing an improvement after a specific training course. In most of the cases, this discordance comes from biases of human nature, such as avoiding extreme ranges, preference of numbers, confirmation biases, among others [8].

Deep learning approaches have been widely implemented in recent years, showing promising results to support diagnosis and stratification diagnosis [9]. These approaches deeply decompose information and allow to represent local and global complex patterns associated to Gleason concepts. For instance, in [10] it was proposed an automatic method to classify benign, Gleason 3, 4, and 5 patches in tissue microarrays. Bulten et al. [11] proposed a semi-supervised method for classifying Gleason patterns on the whole slide images, using 3 different CNNs to segment and classify gland structures. In it [12] was compared different gland segmentation architectures to classify Gleason patterns in whole slide images. Despite advances in deep representations, the variability intra and inter-Gleason scores remain as an open problem, mainly due to rigid learning with stratified and balanced datasets, which results unreal in clinical domain.

Therefore, this work introduces a semi-hard triplet distance metric that tackles Gleason stratification from more variable and challenging positive and negative patches on a particular Gleason level. This learning scheme allows to deal with imbalanced set, learning embedding spaces that maximizes differences among disease stages. The main contributions of this work can be summarized as follows:

1) A feature embedding space, learning from a triplet loss scheme, that better discriminate among Gleason scores.
2) The exploration of semi-hard distances on histological classification problems and the respective integration with other transfer learning mechanics.
3) An experiment among challenging and close stages (Gleason 3 and 4) showing state-of-the-art performance with 74% of accuracy.

II. Materials and Methods

Learning semantic distances can be a key issue in histopatholo-gical Gleason classification, dealing with pattern variability without constrained stratification and artificial balanced of data. The resulting embedding space from different deep training strategies are observed in Figure 1. As observed, the typical representation is overlapped from a classical classification framework with no separation of classes. Specifically, the reported classification results are the contribution of boundary and outlier samples that result

*Biomedical Imaging, Vision and Learning Laboratory (BivL^{2}ab). Universidad Industrial de Santander (UIS), COLOMBIA. famarcar@sabaer.uis.edu.co

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

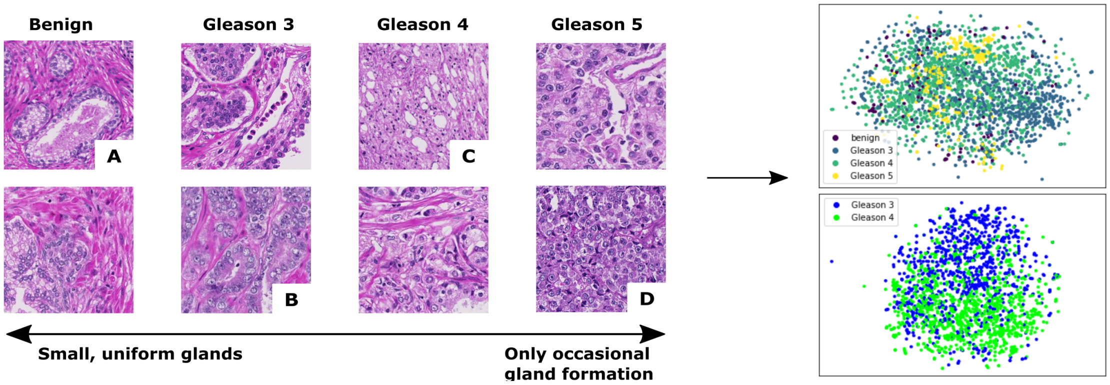

Fig. 1: Gleason score distribution. A) well-formed and uniform glands, B) discrete glans with stroma separation, C) cribriform pattern, D) solid nests, with their respective embedding representations. Top scatter embedding representation from benign tissue, Gleason score 3, 4, and 5. Bottom scatter embedding representation from Gleason score 3 and 4. High semantic overlapping and diffused class boundaries are observed from the embedding space, leading classification variability and weakens Gleason patterns.

different among other classes but especially of the same class.

A. *Learning distance metric*

This work explored and adapted triplet loss metric [13], which uses three samples batches to determine the best weight of semantic representation from raw visual data. Each image patch x_i is mapped trough a deep net f to obtain an embedding representation fx_i. Formally, the triplet is defined by an anchor representation sample fx_i, a positive sample representation fx_i^+, with same class, and a negative sample fx_i^-, *i.e.,* a different Gleason grade class w.r.t the anchor. Hence, a distance D is formulated to minimize the separation between similar pairs and maximize the space regarding the anchor and the negative class. The distance is defined as: $D(fx_i, fx_i^+) + \alpha < D(fx_i, fx_i^-)$, where α is a strength term to force class separation. Such relationship from deep convolutional representation f allows to learn semantic embedding spaces that trend to deal with a better Gleason grade separation. Then, the triplet loss is defined as:

$$Loss(x_i, x_i^+, x_i^-) = \\ max\{0, \alpha + D(fx_i, fx_i^+) - D(fx_i, fx_i^-) \quad (1)$$

The main advantage of this learning scheme, regarding typical cross-entropy, is the self-representation grade capability, which allows to directly model the strong intra-variability reported on Gleason score. The triplet loss can approximate KL divergence by learning cross-entropy w.r.t to negative samples, but also entropy regarding the same Gleason sample. This fact allows to create a robust representation that among others reduces sensitivity to noise samples, avoids adversarial examples, and enhances boundary margins among classes, a fundamental issue on cancer degree characterization.

B. *semi-hard negative mining*

To find correct triplets are crucial during training to force learning a robust representation. Nonetheless, selecting all triplets can be computational expensiveness $O(n^3)$, and selecting only hard triplet can generate sample bias and appears fragile to outliers [14]. To overcome this limitation, this work adopts semi-hard negative mining to strongly penalize negative samples, selecting only triplets where the negative is bounded between $D(fx_i, fx_i^+)$ and the α factor. Then, a more specific definition of triplet loss mining is:

$$D(fx_i, fx_i^+) < D(fx_i, fx_i^-) < D(fx_i, fx_i^+) + \alpha \quad (2)$$

This negative-positive relationship is controlled with α parameter, allowing to contribute on positive loss on initial and middle epochs, reducing variance in gradients [13]. To avoid embedding space normalization, this semi-hard distance was implemented with a cosine distance among anchor and positive/negative samples, as:

$$D(fx_i, fx_i^{(+,-)}) = \frac{fx_i \cdot fx_i^{(+,-)}}{\sqrt{f^2x_i}\sqrt{f^2x_i^{(+,-)}}} \quad (3)$$

C. *Inception architecture as backbone*

As a backbone net, here it was adopted a CNN representation that fully characterizes cancer histological patterns, stratified according to Gleason degrees. These architectures allow to learn complex and multi-level patterns in a set of different layers and kernels. Specifically, in this work, for binary classification it was adopted an InceptionV3 architecture [15]. This architecture factorizes symmetric and asymmetric block convolutions, reducing the number of connections/parameters without decreasing the network efficiency. For multi Gleason classification it was adopted the Xception architecture [16], an extension of InceptionV3, replacing the standard modules by depth-wise separable convolutions. The main issue to train such large net architectures is to have a sufficient amount of data. To overcome this limitation and to take advantage of previously learned representation, in this work was initially used a Transfer Learning (TL) scheme. Two different TL schemes were here validated. Firstly, the selected backbone

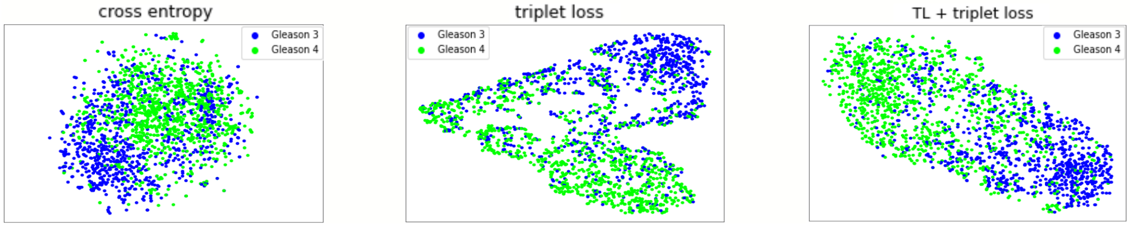

Fig. 2: Embeddings from last inceptionV3 layers for the pathologist 1. Triplet embedding with TL from classical schemes shows better Gleason differentiation, optimizing the space learning semantic concepts to contrast fine-grained Gleason patterns.

net was initialized from ImageNet and then adjusted with a triplet loss. A second TL option was validated in two steps: First, the model was training with a classical cross-entropy loss with a TL from ImageNet. Then, the previous weights were used to adjust embedding space from a coarse histopathological space using the triplet loss.

D. *Experimental setup*

Data. Evaluation was carried out over HARVARD Dataverse [17]. This dataset contains a total of 886 H&E images with benign tissue, Gleason 3, 4, and 5 scores at 40x resolution (0.23 microns per pixel) whit a size of 3100×3100 pixels. The dataset was divided into five tissue microarrays (TMAs) selecting three TMAs to train, one to validation and one to test. The assignation of Gleason scores was performed by one pathologist over all dataset and an additional pathologist to the test sub-set. To train, a patch extraction was performed over each image following the same scheme reported in [10]. In training, a total of 17785 patches were extracted from 641 histology images, coding: 2076 patches for benign tissue, 7226 for grade three, 5207 for grade four, and 4541 for grade five. In the test for pathologist one, was extracted 2239 patches from 245 images distributed as: 127 benign patches for benign, 1602 patches to grade three 2121 patches for grade four, and 387 for grade five.

Method setup. To evaluate the proposed embedding scheme two backbone nets were compared: the inceptionV3 and the Xception. The net was trained, follows two steps: 1) only top layers are trained with 5 epochs and 2) the model is fine-tuned using 20 epochs. An Adam optimizer was used with a learning rate of 0.001. Regarding triplet loss configuration, after a hyperparameter tuning, an empirical α factor of 10 was fixed, forcing an optimal and stable Gleason pattern differentiation without decreasing the network efficiency. Feature embedding vectors were fixed with a dimension of 512, following a semi-hard triplet negative mining.

III. Evaluation and Results

A first experiment was conducted to built an embedding space using different learning representation schemes. Taking into account the main Gleason challenge, the experiment was carried out to differentiate between grades two and three. The set of resultant embedding vectors are projected into a two-dimensional space following a t-SNE projection. As show in Figure 2, the embedding space that results from triplet loss approach achieves a significant separation of grades, which results promising to understand cancer severity patterns. In contrast, the embedding space from classical cross-entropy has a significant overlapping of classes.

Scheme	**Accuracy**	
	InceptionV3	Xception
Cross-entropy loss	71% ± 1.89%	67% ± 1.76%
Triple loss	0.71% ± 2.14%	0.69% ± 1.47%
TL + Triple loss	**0.73%** ± **1.65%**	**0.70%** ± **2.35%**

TABLE I: Evaluation results between inceptionV3 and Xception architectures for the different setup experiments.

Table I summarizes the Gleason grade classification obtained for different learning schemes. In this bi-modal classification task, the two pathologists obtained an average accuracy of 71%, which evidenced the challenge of visual pattern quantification. Each scheme used the InceptionV3 and Xpception nets. In general, the schemes that implement triplet loss achieve better scores, being the recovered samples closer to the centroid of the embedding class representation. The best configuration was achieved by the InceptionV3 (73%), using the classical TL with a further triplet loss adjustment and for Xception (70%) adjusting directly from triplet loss. The achieved results show the robustness of the proposed learning scheme, being more confident the classified samples, *i.e.,* the points in embedding space are closer to their respective centroid.

Figure 3 shows a more detailed analysis, obtained from the confusion matrices between two pathologist references. It should be noted that there exists a remarkable difference between pathologists to classify Gleason grade three, obtaining an average accuracy (agreement) of 47%. The proposed approach, in contrast, achieves balanced results between grades and among pathologists. This fact stands out the robustness of representation and clearly evidence the significant support that could provide on clinical scenarios.

Finally, the multi-grading severity stratification with Xception reported the best results. For training was used an initial TL adjustment, followed by the triplet loss learning. In Figure 4 is shown the confusion matrix obtained from the proposed representation. A main reported drawback is reported for ”Benign” class, which could be attributed to patches that consider abundant background because the automatic delineation on the original dataset. This approach

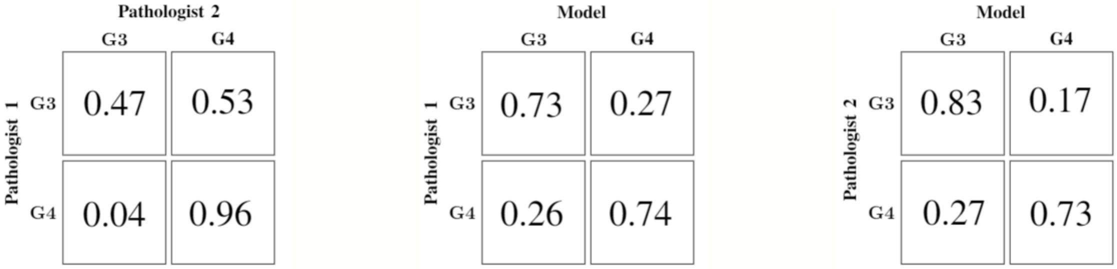

Fig. 3: Confusion matrix of triplet loss with TL for Gleason score 3 (G3) and 4 (G4). This approach show remarkable Gleason differentiation, especially for G3, with more robustness and stable classification.

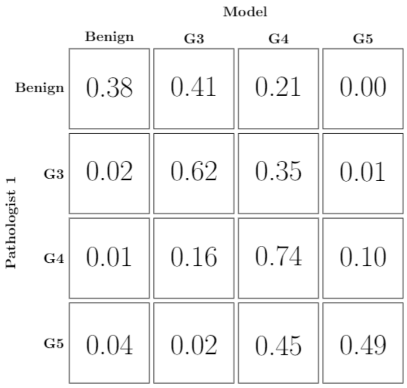

Fig. 4: Model evaluation on pathologist 1 for Gleason scores and benign tissue. Benign classification show a drawback mainly for delineation artifacts in the original dataset.

reaches an average accuracy of 62%, with no static difference to the accuracy of Arvaniti et al. [10] of 63%.

IV. Conclusion

This work introduced the triplet loss training scheme that naturally models intra and inter-class variation on Gleason grade stratification. The proposed approach built an embedding space that properly groups samples of tree and four Gleason degrees, which are the most variable and challenging grades for expert pathologists. The achieved results showed a significant performance on patch identification, allowing to support clinical agreement tasks. Future works include detailed analysis over new representation space, using larger datasets, and trying to capture the space topology to properly classify among Gleason degree patterns.

V. Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in open access by HARVARD Dataverse [17]. Ethical approval was not required confirmed by the license attached with the open-access data.

Acknowledgements

The authors acknowledge the Vicerrectoriá de Investigación y Extensión of the Universidad Industrial de Santander for supporting this research work by the project: ”Identificación de características asociadas con pleomorfismo nuclear en imágenes histológicas de cáncer de mama utilizando algoritmos de aprendizaje profundo”, with SIVIE code 2692.

References

[1] International Agency for Research on Cancer, “The global cancer observatory ”globocan”,” https://gco.iarc.fr/today/home, 2018.

[2] Takashi *et al.* Fukagai, “Discrepancies between gleason scores of needle biopsy and radical prostatectomy specimens,” *Pathology international*, vol. 51, no. 5, pp. 364–370, 2001.

[3] Rebecca *et al.* Arora, “Heterogeneity of gleason grade in multifocal adenocarcinoma of the prostate,” *Cancer: Interdisciplinary International Journal of the American Cancer Society*, vol. 100, no. 11, pp. 2362–2366, 2004.

[4] Athanase *et al.* Billis, “The impact of the 2005 international society of urological pathology consensus conference on standard gleason grading of prostatic carcinoma in needle biopsies,” *The Journal of urology*, vol. 180, no. 2, pp. 548–553, 2008.

[5] Jonathan Epstein et al., “The 2014 international society of urological pathology (isup) consensus conference on gleason grading of prostatic carcinoma,” *The American journal of surgical pathology*, vol. 40, no. 2, pp. 244–252, 2016.

[6] J *et al.* Melia, “A uk-based investigation of inter-and intra-observer reproducibility of gleason grading of prostatic biopsies,” *Histopathology*, vol. 48, no. 6, pp. 644–654, 2006.

[7] Alireza Abdollahi et al., “Inter-observer reproducibility before and after web-based education in the gleason grading of the prostate adenocarcinoma among the iranian pathologists.,” *Acta Medica Iranica*, pp. 370–374, 2014.

[8] Richard *et al.* Zarbo, “Error detection in anatomic pathology,” *Archives of Pathology and Laboratory Medicine*, vol. 129, no. 10, pp. 1237–1245, 2005.

[9] Metin *et al.* Gurcan, “Histopathological image analysis: A review,” *IEEE reviews in biomedical engineering*, vol. 2, pp. 147–171, 2009.

[10] Eirini a*et al.* Arvaniti, “Automated gleason grading of prostate cancer tissue microarrays via deep learning,” *Scientific reports*, vol. 8, no. 1, pp. 1–11, 2018.

[11] Wouter Bulten et al., “Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study,” *The Lancet Oncology*, vol. 21, no. 2, pp. 233–241, 2020.

[12] Nathan *et al.* Ing, “Semantic segmentation for prostate cancer grading by convolutional neural networks,” in *Medical Imaging 2018: Digital Pathology*. International Society for Optics and Photonics, 2018, vol. 10581, p. 105811B.

[13] Florian *et al.* Schroff, “Facenet: A unified embedding for face recognition and clustering,” in *CVPR*, 2015, pp. 815–823.

[14] Ye *et al.* Yuan, “In defense of the triplet loss again: Learning robust person re-identification with fast approximated triplet loss and label distillation,” in *CVPR*, 2020, pp. 354–355.

[15] C *et al.* Szegedy, “Rethinking the inception architecture for computer vision,” in *CVPR*, 2016, pp. 2818–2826.

[16] François Chollet, “Xception: Deep learning with depthwise separable convolutions,” in *CVPR*, 2017, pp. 1251–1258.

[17] Eirini Arvaniti et al., “Replication Data for: Automated Gleason grading of prostate cancer tissue microarrays via deep learning.,” 2018.