Triplet Mining BUT in a Dynamic Manner: Batch-Incremental Triplet Sampling for Training Triplet Networks Using Bayesian Updating Theorem

Milad Sikaroudi*, Benyamin Ghojogh†, Fakhri Karray†, *Fellow, IEEE*,
Mark Crowley†, *Member, IEEE*, H.R. Tizhoosh*, *Senior Member, IEEE*
*Kimia Lab, University of Waterloo, Waterloo, ON, Canada
†Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada
Emails: {msikaroudi, bghojogh, karray, mcrowley, tizhoosh}@uwaterloo.ca

Abstract—**Triplet variant networks are robust entities for learning a discriminative embedding subspace. There exist different triplet mining approaches for selecting the most suitable training triplets. Some of these mining methods rely on the extreme distances of instances, and some others make use of sampling. However, sampling from stochastic distributions of data rather than sampling merely from the existing embedding instances can provide more discriminative information. In this work, we sample triplets from distributions of data rather than the existing instances. We consider a multivariate normal distribution for the embedding of each class. Using Bayesian updating and conjugate priors, we update the distributions of classes dynamically by receiving the new mini-batches of training data. The proposed triplet mining with Bayesian updating can be used with any triplet loss function, e.g., *triplet-loss* or Neighborhood Component Analysis (NCA) loss. Accordingly, Our triplet mining approaches are called Bayesian Updating Triplet (BUT) and Bayesian Updating NCA (BUNCA), considering which loss function is being used. Experimental results on MNIST and histopathology colorectal cancer (CRC) datasets substantiate the effectiveness of the proposed triplet mining method.**

I. Introduction

Siamese network variants contain several, typically two [1] or three [2], [3], sub-networks sharing their weights. The Siamese network variants are robust networks for learning a discriminative embedding space, i.e., explicit metric space, between the classes of data [4]. One of these variants is the triplet network in which anchor, positive and negative triplets are used for decreasing and increasing the distance of anchor-positive and anchor-negative pairs, respectively [2] resulting in increasing and decreasing the inter- and intra-class variances of data [5]. Two popular forms of loss function for training triplets are *triplet-loss* [2] and the softmax form [6]. Some examples for the latter are Neighborhood Component Analysis (NCA) [7] and Proxy-NCA [8].

Apart from the loss functions, there is another knob to turn, which is how the triplets are sampled. It is shown in [9] that sampling of the triplets also matters in deep embedding learning. Hence, proposing a decent sampling strategy has not less importance than a novel loss function. In other words, for the triplet networks, the more informative and stable triplets are drawn from the pool of patches, the more qualitatively salient embedding will be.

There are some triplet mining strategies in the literature. Instead of using all the triplets in a mini-batch of data, i.e., Batch All (BA) [10], one can mine the triplets like in Batch Semi-Hard (BSH) [2] and Batch Hard (BH) [11]. Some mining methods, such as Easy Positive (EP) [12], concentrate on the extreme distances of instances. However, some other triplet mining methods use the concept of sampling from the available triplets in the mini-batch of data [9].

In this work, we aim to draw the positive and negative samples for every anchor instance in a dynamic manner. The main idea is to sample the positive and negative instances of triplets for every anchor in a mini-batch of data from some distributions rather than from the embedded data points themselves. This gives the triplet network more opportunity to explore the embedding space for increasing and decreasing the inter- and intra-class variances because the triplet information is not restricted to only the embedded data but is stochastic. This is while the related work on triplet sampling sample the triplets from the existing embedded data instances [9] and thus do not use the stochastic information of embedding space.

We assume a multivariate normal distribution for the embedded data instances of every class. These distributions are updated dynamically by receiving new streaming embedded data for the different classes. For this dynamic updating, we leverage the theory of Bayesian distribution updating [13], [14] and conjugate priors [15], [16]. Sampling from dynamic distributions makes the task of sampling not only more robust to outliers but also more amenable to available data. The proposed approaches are called *Bayesian Updating for Triplet loss (BUT)* and *Bayesian Updating for NCA loss (BUNCA)*.

The rest of the paper is organized as follows. Section II introduces the necessary background on Bayesian updating and conjugate priors. The dynamic triplet sampling for training triplet networks is proposed in Section III. We report and discuss the experimental results in Section IV. Finally, Section V concludes the paper and indicates the possible future work.

The first two authors contributed equally to this work.

II. Background on Bayesian Updating

A. Bayesian Updating

Let X and θ be two random variables where θ is a parameter of the distribution of X. According to Bayes' rule, we have:

$$\mathbb{P}(\theta|X) = \frac{\mathbb{P}(X|\theta)\,\mathbb{P}(\theta)}{\mathbb{P}(X)} \implies \mathbb{P}(\theta|X) \propto \mathbb{P}(X|\theta)\,\mathbb{P}(\theta), \quad (1)$$

which shows the relation of the posterior $\mathbb{P}(\theta|X)$, likelihood $\mathbb{P}(X|\theta)$, and prior $\mathbb{P}(\theta)$. Given some data X and the prior over the parameter of interest θ, we want to find the posterior using Eq. (1). This is the basic idea behind *Bayesian updating* in which the posterior over the parameter of interest is updated after receiving some new data, i.e., using the new data X, we have $\mathbb{P}(\theta) \mapsto \mathbb{P}(\theta|X)$ [13].

B. Conjugate Priors

If the posterior distribution $\mathbb{P}(\theta|X)$ and the prior distribution $\mathbb{P}(\theta)$ are in the same probability distribution family, they are called *conjugate distributions* and the prior is the *conjugate prior* for the likelihood $\mathbb{P}(X|\theta)$ [14].

Assume there already exist some data, denoted by X^0, and some new data, indicated by X', are received. The existing data X^0 has a distribution with some parameter(s) θ. The posterior of the parameter of interest, i.e., $\mathbb{P}(\theta|X)$, can be updated using the new data. Hence, this can be used to update the parameter(s) of the distribution of X using the newly received data [16].

Let the data X have a multivariate normal (or Gaussian) distribution, so its likelihood is $\mathbb{P}(X|\theta)$. Assume both the mean and covariance of likelihood are considered as random variables, so θ includes mean and covariance. Using the new data X_{new}, we want to update the parameters, mean and covariance, of the normal distribution. In this case, the likelihood $\mathbb{P}(X|\theta)$ has a multivariate normal distribution, and for updating the posterior, we should use the conjugate prior for the likelihood. The conjugate prior distribution for the multivariate normal distribution with both random mean and covariance is the normal-inverse-Wishart distribution [15]. In our analysis, we also require the skewed generalized Student-t distribution. In the following, we introduce these distributions.

C. Multivariate Normal, Inverse-Wishart, and Skewed Generalized Student-t Distributions

Multivariate Normal Distribution: The Probability Density Function (PDF) of the *multivariate normal distribution* is defined as [14]:

$$X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{1}{\sqrt{(2\pi)^d\,|\boldsymbol{\Sigma}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{\mathbf{-1}}(\boldsymbol{x}-\boldsymbol{\mu})\Big), \quad (2)$$

where d is the dimensionality of data, $|.|$ denotes the determinant of matrix, and $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{\mu} \in \mathbb{R}^d$, and $\boldsymbol{\Sigma} \in \mathbb{R}^{d\times d}$ are the data, mean, and covariance of data, respectively. The mean and covariance of the normal distribution can be estimated by the sample mean and sample covariance matrix, respectively.

Wishart and Inverse Wishart Distributions: The PDF of the *Wishart distribution* is defined as [14]:

$$X \sim \mathcal{W}_d(\boldsymbol{V}, \nu) := \frac{1}{2^{(\nu d)/2}\,|\boldsymbol{V}|^{\nu/2}\,\Gamma_d(\frac{\nu}{2})} |\boldsymbol{x}|^{(\nu-d-1)/2} \exp(-\frac{1}{2}\mathbf{tr}(\boldsymbol{V}^{-1}\boldsymbol{x})), \quad (3)$$

where ν is the degrees of freedom (which should be $\nu \geq d$), $\mathbb{R}^{d\times d} \ni \boldsymbol{V} \succ 0$ is the scale matrix, $\mathbf{tr}(.)$ denotes the trace of matrix, and $\Gamma_d(.)$ is the *multivariate gamma function* [17]:

$$\Gamma_d(a) := \int_{\boldsymbol{S}\succ 0} \exp\big(-\mathbf{tr}(\boldsymbol{S})\big)\,|\boldsymbol{S}|^{a-(d+1)/2}\,d\boldsymbol{S}. \quad (4)$$

Consider a variable with Wishart distribution, i.e. $Z \sim \mathcal{W}_d(\boldsymbol{V}, \nu)$. Then, the variable $X = Z^{-1}$ has the *inverse Wishart distribution* whose PDF is defined as [14]:

$$X \sim \mathcal{W}_d^{-1}(\boldsymbol{\Psi}, \nu) := \frac{|\boldsymbol{\Psi}|^{\nu/2}}{2^{(\nu d)/2}\,\Gamma_d(\frac{\nu}{2})} |\boldsymbol{x}|^{-(\nu+d+1)/2} \exp(-\frac{1}{2}\mathbf{tr}(\boldsymbol{\Psi}\boldsymbol{x}^{-1})), \quad (5)$$

where $\mathbb{R}^{d\times d} \ni \boldsymbol{\Psi} \succ 0$ is the scale matrix and we have $\boldsymbol{\Psi} = \boldsymbol{V}^{-1}$ [18].

From the moments of the inverse Wishart distribution, the mean of a random variable $X \sim \mathcal{W}_d^{-1}(\boldsymbol{\Psi}, \nu)$ is as follows [19]:

$$\mathbb{E}(X) = \frac{\boldsymbol{\Psi}}{\nu - d - 1}, \quad \forall\, \nu > d+1. \quad (6)$$

Skewed Generalized Student-t Distribution: The PDF of the *Student-t distribution* is defined as [14]:

$$X \sim t_\nu := \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})} (1 + \frac{x^2}{\nu})^{-(\nu+1)/2}, \quad (7)$$

where $\nu > 0$ is the degrees of freedom and $\Gamma(\nu) := (\nu - 1)!$ is the Gamma function. The Student-t distribution can be generalized which is called the *skewed generalized Student-t distribution* whose PDF is defined as [15], [20]:

$$X \sim t_\nu(\mu, \sigma^2) := \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\sigma\,\Gamma(\frac{\nu}{2})} \Big(1 + \frac{1}{\nu}\big(\frac{x-\mu}{\sigma}\big)^2\Big)^{-(\nu+1)/2}, \quad (8)$$

where μ and σ^2 are the mean and variance, respectively. The generalized Student-t distribution can be d-dimensional multivariate [21, Definition 2]:

$$X \sim t_\nu(\boldsymbol{\mu}, \boldsymbol{\Sigma}) := \frac{\Gamma(\frac{\nu+d}{2})}{(\nu\pi)^{\nu/2}\,\Gamma(\frac{\nu}{2})} \Big(1 + \frac{1}{\nu}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\Big)^{-(\nu+1)/2}, \quad (9)$$

where $\boldsymbol{\mu} \in \mathbb{R}^d$ and $\mathbb{R}^{d\times d}$ are the mean and covariance, respectively. The mean of the skewed generalized Student-t distribution is $\mathbb{E}(X) = \boldsymbol{\mu}$ [15].

D. The Normal-Inverse-Wishart Distribution

As was mentioned before, the prior distribution for the multivariate normal distribution with both mean and covariance as random variables is the inverse Wishart distribution. Recall that we have some existing data denoted by X^0. We show the set of existing data vectors by $\{\boldsymbol{x}_i^0\}_{i=1}^{n_0}$ where n_o is the sample size of the existing data. Assume that data have a multivariate normal distribution $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let $\mathbb{R}^d \ni \boldsymbol{\mu}^0 := (1/n_0)\sum_{i=1}^{n_0}\boldsymbol{x}_i^0$ and $\mathbb{R}^d \ni \boldsymbol{\mu}' := (1/n')\sum_{i=1}^{n'}\boldsymbol{x}_i'$ denote the sample mean of the existing and new data, respectively. Likewise, $\mathbb{R}^{d\times d} \ni \boldsymbol{\Sigma}^0 := (1/n_0)\sum_{i=1}^{n_0}(\boldsymbol{x}_i^0 - \boldsymbol{\mu}^0)(\boldsymbol{x}_i^0 - \boldsymbol{\mu}^0)^\top$ and $\mathbb{R}^{d\times d} \ni \boldsymbol{\Sigma}' := (1/n')\sum_{i=1}^{n'}(\boldsymbol{x}_i' - \boldsymbol{\mu}')(\boldsymbol{x}_i' - \boldsymbol{\mu}')^\top$ are the sample covariance matrix over the existing and new data, respectively.

The prior of covariance is $\boldsymbol{\Sigma} \sim \mathcal{W}_d^{-1}(\boldsymbol{\Sigma}'^{-1}, n')$ and the distribution of mean given covariance is $\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}(\boldsymbol{\mu}', (1/n')\boldsymbol{\Sigma})$ [14], [15]. The joint distribution of the mean and covariance is the *Normal-Inverse-Wishart (NIW) distribution* [14], [15]:

$$\begin{aligned}\mathbb{P}(\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \text{NIW}(\boldsymbol{\mu}',\nu_1',\boldsymbol{\Sigma}',\nu_2') \\ &:= \frac{|\boldsymbol{\Sigma}'|^{\nu_2'/2}|\boldsymbol{\Sigma}|^{-((\nu_2'+d)/2+1)}}{2^{(\nu_2'd)/2}\Gamma_d(\frac{\nu_2'}{2})(\frac{2\pi}{\nu_1'})^{d/2}}\times \\ &\exp\Big(-\frac{1}{2}\textbf{tr}(\boldsymbol{\Sigma}'\boldsymbol{\Sigma}^{-1}) - \frac{\nu_1'}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}')^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}')\Big),\end{aligned} \tag{10}$$

where ν_1' and ν_2' are the sample sizes of new data used for calculating the new mean and covariance matrix. In this work, we have $\nu_1' = \nu_2' = n'$.

The posterior of mean and covariance of data is again a NIW distribution [14], [15]:

$$\begin{aligned}\mathbb{P}(\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{x}^0,\boldsymbol{\mu}',\nu_1',\boldsymbol{\Sigma}',\nu_2') \\ = \text{NIW}\big(\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\boldsymbol{\eta},\nu_1'+n_0,\boldsymbol{\Upsilon},\nu_2'+n_0\big),\end{aligned} \tag{11}$$

$$\mathbb{R}^d \ni \boldsymbol{\eta} := \frac{\nu_1'\boldsymbol{\mu}' + n_0\boldsymbol{\mu}^0}{\nu_1' + n_0}, \tag{12}$$

$$\mathbb{R}^{d\times d} \ni \boldsymbol{\Upsilon} := \nu_2'\boldsymbol{\Sigma}' + n_0\boldsymbol{\Sigma}^0 + \frac{\nu_1' n_0}{\nu_1' + n_0}(\boldsymbol{\mu}^0 - \boldsymbol{\mu}')(\boldsymbol{\mu}^0 - \boldsymbol{\mu}')^\top. \tag{13}$$

The marginal distributions of mean and covariance of data are [14], [15]:

$$\mathbb{P}(\boldsymbol{\mu}\,|\,\boldsymbol{x}^0) = t_{\nu_2'+n_0-d+1}\Big(\boldsymbol{\eta}, \frac{\boldsymbol{\Upsilon}}{(\nu_1'+n_0)(\nu_2'+n_0-d+1)}\Big), \tag{14}$$

$$\mathbb{P}(\boldsymbol{\Sigma}\,|\,\boldsymbol{x}^0) = \mathcal{W}_d^{-1}(\boldsymbol{\Upsilon}^{-1}, \nu_2' + n_0), \tag{15}$$

respectively. The Eqs. (14) and (15) can be used to update the parameters of a multivariate normal distribution upon receiving the new data.

III. Dynamic Triplet Sampling for Training Triplet Networks

A. Preliminaries and Notations

Consider a q-dimensional training dataset $\{\boldsymbol{z}_i\}_{i=1}^n$ where $\boldsymbol{z}_i \in \mathbb{R}^q$. The class labels of instances are $\{y_i\}_{i=1}^n$. Suppose we have c number of classes in the dataset. We use the mini-batch (of size b) stochastic gradient descent for training the network. Let n^j denote the training sample size per class in a mini-batch. We show the i-th training instance of the j-th class in a mini-batch by $\boldsymbol{z}_i'^j$. Let $\boldsymbol{x}_i'^j \in \mathbb{R}^d$ denote the embedding of $\boldsymbol{z}_i'^j$ by the triplet network where the dimensionality of embedding space is d.

The data of each class are accumulated by receiving new mini-batches of data. Let n_0^j denote the sample size of accumulated data for the j-th class so far. The sample size per j-th class in a mini-batch is denoted by n'^j. In this work, we have $n'^1 = \dots = n'^c = n' = \lceil b/c \rceil$ and $n_0^1 = \dots = n_0^c = n_0$ because we take the same sample size per class in the mini-batch. This n' is the sample size of new incoming data per class in every mini-batch. The accumulated data for the j-th class so far are denoted by $\boldsymbol{x}^{0,j}$. Also, $\boldsymbol{\mu}^j$ and $\boldsymbol{\Sigma}^j$ are the mean and covariance of the distribution of the j-th class, respectively.

B. Sampling Algorithm

We assume a multivariate normal distribution for the embedded data of every class. This assumption makes sense according to the central limit theorem [22] and the fact that the normal distribution is the most common continuous distribution. In the first batch, where there is not already any embedding of training data, we use Maximum Likelihood Estimation (MLE) for the estimates of distribution parameters. The mean and covariance of the embedded data of every class are estimated by the sample mean and covariance matrix, respectively.

In later batches after the first batch, we do have some existing data per class, denoted by $n_0^j, \forall j$. According to Bayesian updating, the mean and covariance of distribution of every class are updated by Eqs. (14) and (15), respectively. We update the mean and covariance matrix of the distribution of every class by the expectation of Eqs. (14) and (15) which are the generalized Student-t and the inverse Wishart distributions, respectively. According to the expectations of these two distributions which were introduced in Section II, the updates of mean and covariance of the j-th class are as follows:

$$\boldsymbol{\mu}^{0,j} \leftarrow \mathbb{E}(\boldsymbol{\mu}^j\,|\,\boldsymbol{x}^{0,j}) = \boldsymbol{\eta}^j \overset{(12)}{=} \frac{n'\boldsymbol{\mu}'^j + n_0\boldsymbol{\mu}^{0,j}}{n' + n_0}, \tag{16}$$

$$\boldsymbol{\Sigma}^{0,j} \leftarrow \mathbb{E}(\boldsymbol{\Sigma}^j\,|\,\boldsymbol{x}^{0,j}) \overset{(6)}{=} \frac{\boldsymbol{\Upsilon}^{-1}}{n' + n_0 - d - 1}, \forall\, n' + n_0 > d + 1, \tag{17}$$

where, in Eq. (13), we use $\nu_1' = \nu_2' = n'$ and calculate $\boldsymbol{\mu}'^j$, $\boldsymbol{\mu}^{0,j}$, $\boldsymbol{\Sigma}'^j$, and $\boldsymbol{\Sigma}^{0,j}$ by sample mean and sample covariance matrix using the new batch of data. Note that for $n' + n_0 \leq d + 1$ which is in very first mini-batches of first epoch, we update the covariance matrix by MLE.

The proposed dynamic triplet sampling is summarized in Algorithm 1. As this algorithm reports, the mean and covariance of every class are estimated by MLE at the initial batch. In the next batches, Bayesian updating is exploited

```
Procedure: TrainTripletNetwork($\{\boldsymbol{z}_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$)
Input: training data: $\{\boldsymbol{z}_i\}_{i=1}^n$, training labels: $\{y_i\}_{i=1}^n$
for all required epochs do
    for all batches in epoch do
        $\{\boldsymbol{x}_i\}_{i=1}^b \leftarrow$ Feed $\{\boldsymbol{z}_i\}_{i=1}^b$ to the triplet-loss network
        for class $j$ from $1$ to $c$ do
            if it is first mini-batch then
                $\boldsymbol{\mu}^{0,j} := (1/n') \sum_{i=1}^{n'} \boldsymbol{x}_i'^j$
                $\boldsymbol{\Sigma}^{0,j} := (1/n') \sum_{i=1}^{n'} (\boldsymbol{x}_i'^j - \boldsymbol{\mu}^{0,j})(\boldsymbol{x}_i'^j - \boldsymbol{\mu}^{0,j})^\top$
            else
                $\boldsymbol{\mu}'^j := (1/n') \sum_{i=1}^{n'} \boldsymbol{x}_i'^j$
                $\boldsymbol{\mu}^{0,j} := (n'\boldsymbol{\mu}'^j + n_0\boldsymbol{\mu}^{0,j})/(n' + n_0)$
                if $n' + n_0 > d + 1$ then
                    $\boldsymbol{\Upsilon} := n'\boldsymbol{\Sigma}'^j + n_0\boldsymbol{\Sigma}^{0,j} + \frac{n' n_0}{n' + n_0}(\boldsymbol{\mu}^{0,j} - \boldsymbol{\mu}'^j)(\boldsymbol{\mu}^{0,j} - \boldsymbol{\mu}'^j)^\top$
                    $\boldsymbol{\Sigma}^{0,j} := \boldsymbol{\Upsilon}^{-1}/(n' + n_0 - d - 1)$
                else
                    $\boldsymbol{\Sigma}^{0,j} := (1/n') \sum_{i=1}^{n'} (\boldsymbol{x}_i'^j - \boldsymbol{\mu}'^j)(\boldsymbol{x}_i'^j - \boldsymbol{\mu}'^j)^\top$
        for instance $i$ from $1$ to $b$ do
            anchor $\leftarrow \boldsymbol{x}_i$
            for class $j$ from $1$ to $c$ do
                if $j = y_i$ then
                    Sample $(c-1)$ positive instances $\sim \mathcal{N}(\boldsymbol{\mu}^{0,j}, \boldsymbol{\Sigma}^{0,j})$
                else
                    Sample a negative instance $\sim \mathcal{N}(\boldsymbol{\mu}^{0,j}, \boldsymbol{\Sigma}^{0,j})$
        Minimize the triplet/NCA loss with the $(b \times (c-1))$ triplets.
```

Algorithm 1: Dynamic Triplet Sampling with Bayesian Updating

for updating the mean and covariance of classes. After the means and covariances are updated, we sample the triplets. For every instance of a batch, considered as an anchor, a negative instance is sampled from each different class resulting in $(c-1)$ negatives per anchor. Accordingly, $(c-1)$ positive instances are also sampled from the same class of anchor. Overall, $(b \times (c-1))$ triplets are sampled in every mini-batch while the distributions of classes are being updated dynamically.

C. Optimization of the Loss Functions

In a mini-batch, let the anchor, positive, and negative instances be indexed by i, k, ℓ, respectively. Using the $(b \times (c-1))$ sampled triplets, the triplet loss function can be employed to train the *triplet-loss* network [2]:

$$\text{minimize} \sum_{i=1}^{b} \sum_{k=1}^{c-1} \sum_{\ell=1}^{c-1} \Big[m + \|\boldsymbol{x}_i - \boldsymbol{x}_k\|_2^2 - \|\boldsymbol{x}_i - \boldsymbol{x}_\ell\|_2^2\Big]_+, \tag{18}$$

where $[.]_+ := \max(., 0)$ denotes the standard Hinge loss and m is a small margin (e.g., 0.25). We name the approach of dynamic triplet sampling used with the triplet loss by *Bayesian Updating for Triplet loss (BUT)*.

As was mentioned before, the triplet loss tries to increase and decrease the inter- and intra-class variances to have a discriminating embedding space for classes of data. This intuition can also be implemented in a softmax form [6] which is referred to as NCA [7]. We can use this form to train the network:

$$\text{minimize} \; -\sum_{i=1}^{b} \sum_{k=1}^{c-1} \ln\Big(\frac{\exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|_2^2)}{\sum_{\ell=1}^{c-1} \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_\ell\|_2^2)}\Big). \tag{19}$$

The name of using dynamic triplet sampling with the NCA loss function we put *Bayesian Updating for NCA loss (BUNCA)*.

IV. Experiments

A. Datasets

For experiments, we used two different datasets. The first dataset is the MNIST digits data [23] with 60,000 training instances and 10,000 test instances of size 28×28 pixels. The second dataset we used is the large colorectal cancer (CRC) histopathology dataset [24] with 100,000 stain-normalized (224×224)-pixel patches. The large CRC dataset includes nine classes of tissues, namely adipose, background, debris, lymphocytes (lympho), mucus, smooth muscle, normal colon mucosa (normal), cancer-associated stroma, and colorectal adenocarcinoma epithelium (tumor). Note that literature has shown the effectiveness of triplet variants networks for histopathology data, either with triplet loss [25] or with NCA loss [26]; this shows the importance of validating our approaches on this domain.

B. Experimental Setup

For the MNIST dataset, we split the training data into 70% and 30% portions for training and validation sets. The test set with 10,000 images was used for the test. The CRC data were split into training, validation, and test sets with 70%, 15%, and 15% portions, respectively. We used ResNet-18 network [27] as the backbone of *triplet* network. Using the validation set, early stopping [28] was employed, and the maximum number of epochs was set to 50. The batch size was 50 and 45 for the MNIST and CRC data, respectively, where every batch contains five instances per class (i.e., $n' = 5$). The learning rate was set to 10^{-5}, and the dimensionality of the embedding space was 128.

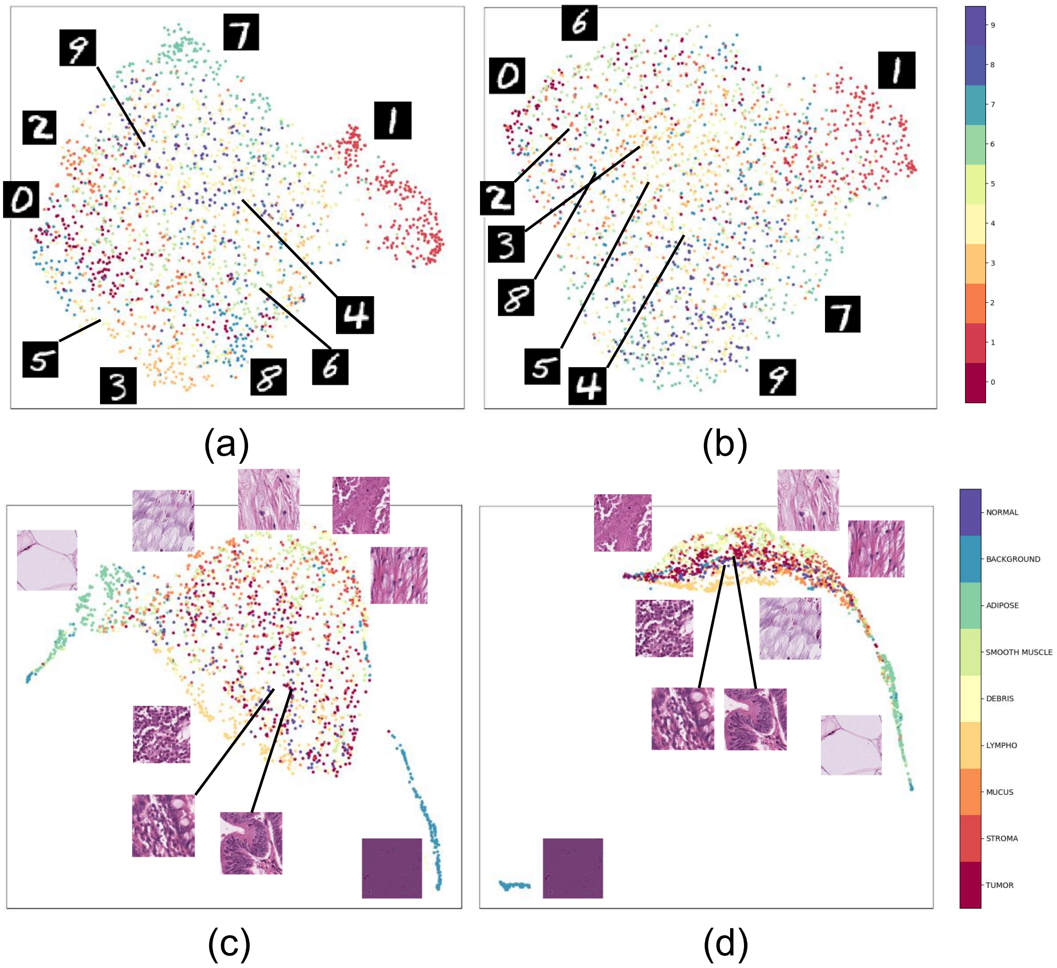

Fig. 1. 2D visualization of embeddings: (a) MNIST test set using BUT approach, (b) MNIST test set using BUNCA approach, (c) CRC test set using BUT approach, and (d) CRC test set using BUNCA approach.

C. *Visualization of Embedding Spaces*

The 2D visualization of spaces was performed using the Uniform Manifold Approximation and Projection (UMAP) [29] applied to the embedded data. Figure 1 illustrates the embedding of test sets of the MNIST and CRC data using the BUT and BUNCA sampling methods. As can be seen in this figure, the learned embedding spaces are interpretable. In embeddings of MNIST data, the similar digits, in the style of writing, fall close to one another. Closely embedded digits by BUT (see Fig. 1-a) are the digits 1 and 7, 7 and 9, 3 and 8, and 4 (second style of writing) and 9. Likewise, closely embedded digits by BUNCA (see Fig. 1-b) are the digits 0 and 6, 1 and 7, 7 and 9, 3 and 8, and 2 and 3 (because continuing the underneath curve of 2 results in 3).

The embedding spaces for the histopathology data are also meaningful. The histopathology patches with similar patterns have been embedded close to each other as expected. In embedding using the BUT approach (see Fig. 1-c), the patches are embedded from smoothest to busiest patterns in a circular manner. These patches, with smoothest to busiest patterns, are adipose (with thin stripes of fat), mucus, smooth muscle, debris, stroma, tumor, normal, and lympho (with many dots). Moreover, the background patch with no pattern (but purple because of stein normalization) is separated from the tissues, as expected. In embedding using the BUNCA approach (see Fig. 1-d), the patches with a considerable amount of crowdedness are embedded closely. For example, adipose, mucus, stroma, and smooth muscle, which are smoother, fall close to each other while tumor, normal, lympho, and debris, with diverse patterns, are embedded close to each other. Again, the background patches are embedded far from the tissue types. The meaningfulness of the learned embedded spaces shows the effectiveness of the proposed BUT and BUNCA approaches.

D. *Query and Retrieval*

For the evaluation of the embedding space, one can see the embedded instances as a database where nearby cases can be retrieved for a query instance. The retrievals are extracted using the nearest neighbors in the embedding space. Because of representation learning, the retrievals are expected to be similar to the query in terms of pattern. In Fig. 2, we illustrate the top ten retrievals for query examples in both MNIST and

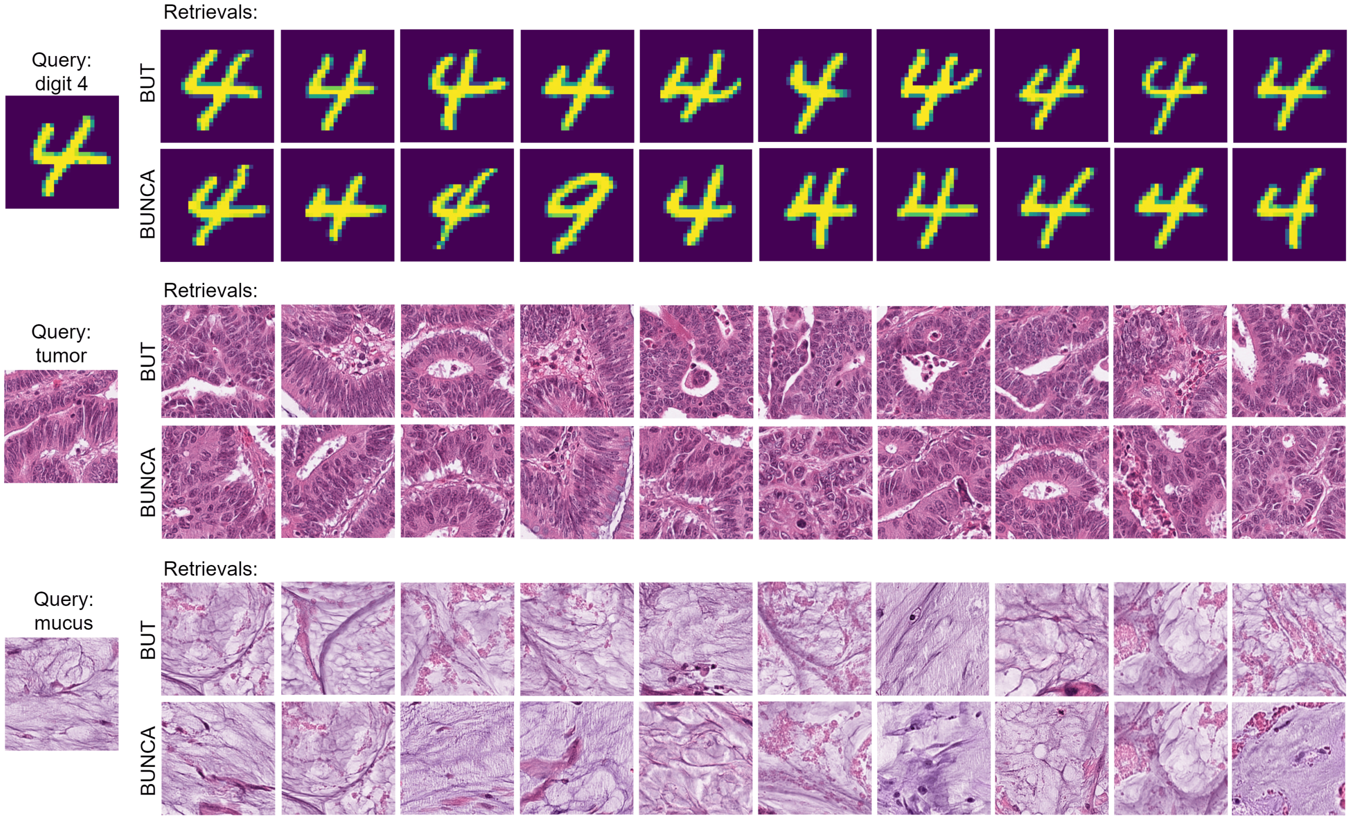

Fig. 2. The retrieval of images for a query image in the embedded spaces learned using the BUT and BUNCA sampling approaches. The retrievals are sorted from left to right. The top two rows correspond to the MNIST data, and the other rows are for the histopathology data.

histopathology data. The retrievals in the embedding spaces using both BUT and BUNCA approaches are shown.

1) Retrieval of Digit Images: In Fig. 2, the retrievals for a digit 4 with the second style of writing are depicted. As expected, the retrievals are very similar to the pattern of the query image. Compared to the last retrievals, the first retrievals are more similar to the query as expected. For this query example in the BUNCA approach, one of the retrievals is wrong, but it is interpretable. The second writing style of digit ”4” is very similar to digit ”9” and can be converted to it by a slight change.

2) Retrieval of Histopathology Patches: Query-retrieval is very common for histopathology data in hospitals where similar patches are extracted from the database. The type of disease or tissue can be found out by a majority vote amongst the retrievals [30]. Figure 2 shows retrievals for two different tissue types, which are tumor and mucus. The former has more complex patterns, in contrast to the latter one. As the figure shows, the retrievals are very similar to the pattern of query patch, and this validates the learned embedding spaces using the BUT and BUNCA approaches.

E. Comparison with Baseline Methods

In Tables I and II, we compare the proposed BUT and BUNCA approaches with the existing triplet mining methods in the literature. These tables report the Recall@k (R@k) and Top-k Accuracy (TkA) metrics on the embedded test data, for different k values. The baseline approaches, which we compare with, are BA [10], BSH [2], BH [11], EP [12], DWS [9], NCA [7], and PNCA [8]; these methods were briefly introduced in Section I. Among these methods, DWS is a sampling method that samples from the existing instances in the mini-batch in contrast to our proposed approaches, which sample from the distribution of data.

TABLE I
COMPARISON OF THE PROPOSED TRIPLET MINING APPROACHES WITH THE BASELINES ON THE MNIST DATASET.

	R@1/T1A	R@4	R@8	R@16	T3A	T5A
BA [10]	79.31	93.53	96.55	98.21	91.44	94.62
BSH [2]	78.95	92.61	96.09	98.17	90.68	93.93
BH [11]	85.75	95.31	97.43	98.63	94.09	96.16
EP [12]	73.34	90.09	95.08	97.68	87.69	92.19
DWS [9]	76.44	91.35	95.72	97.68	89.38	92.79
NCA [7]	85.40	95.48	97.46	98.76	94.09	96.25
PNCA [8]	83.71	94.69	97.31	98.55	93.36	95.71
BUT	88.03	96.25	98.15	99.09	95.33	97.02
BUNCA	78.67	92.44	95.77	98.02	90.22	93.70

Table I reports the results for the MNIST dataset. The proposed BUT approach outperforms all other methods. Moreover, BUNCA performs better than EP and DWS, where DWS is also a sampling approach for triplet mining. The results for

TABLE II
COMPARISON OF THE PROPOSED TRIPLET MINING APPROACHES WITH THE BASELINES ON THE CRC DATASET.

	R@1/T1A	R@4	R@8	R@16	T3A	T5A
BA [10]	38.54	66.76	80.64	89.97	60.52	71.50
BSH [2]	30.85	60.39	77.73	90.33	53.32	66.17
BH [11]	79.09	92.60	96.00	97.95	90.71	93.93
EP [12]	69.94	87.88	93.20	96.38	84.97	89.86
DWS [9]	76.06	91.31	95.34	97.58	88.95	92.83
NCA [7]	77.87	92.25	95.92	98.01	90.48	93.70
PNCA [8]	78.85	92.24	95.80	97.78	90.31	93.54
BUT	79.14	92.32	95.60	97.65	90.33	93.59
BUNCA	78.67	92.28	95.64	97.71	89.92	93.56

the CRC histopathology data are reported in Table II. On this data, the performance of BUNCA is closer to BUT. In most cases, BUT has the best performance against all the baseline approaches. On this dataset, BUNCA performs better than BA, BSH, EP, DWS, NCA, and is comparable with PNCA. Overall, these two tables demonstrate the effectiveness of the proposed mining approaches for triplet training.

V. CONCLUSION AND FUTURE DIRECTION

Different triplet mining approaches have been proposed since the appearance of triplet networks. In this paper, we proposed a triplet mining method which considers a multivariate normal distribution for the embedding of every class through sampling the triplets from these distributions rather than from the existing instances in the mini-batch. By Bayesian updating, the distributions are dynamically updated using the received stream of mini-batches. This approach makes use of the stochastic information of the embedding space, rather than being restricted to the existing instances, for better discrimination of classes. The proposed BUT and BUNCA approaches of the dynamic triplet sampling were validated by experiments on two datasets. As a possible future work, one can explore a mixture of Gaussian distributions for every class of data using expectation maximization.

REFERENCES

[1] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2006, pp. 1735–1742.

[2] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[3] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[4] F.-J. Chang and R. Nevatia, "Image set classification via template triplets and context-aware similarity embedding," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 231–247.

[5] B. Ghojogh, M. Sikaroudi, S. Shafiei, H. Tizhoosh, F. Karray, and M. Crowley, "Fisher discriminant triplet and contrastive losses for training siamese networks," in *2020 international joint conference on neural networks (IJCNN)*. IEEE, 2020.

[6] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6210–6219.

[7] J. Goldberger, G. E. Hinton, S. T. Roweis, and R. R. Salakhutdinov, "Neighbourhood components analysis," in *Advances in neural information processing systems*, 2005, pp. 513–520.

[8] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 360–368.

[9] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl, "Sampling matters in deep embedding learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2840–2848.

[10] S. Ding, L. Lin, G. Wang, and H. Chao, "Deep feature learning with relative distance comparison for person re-identification," *Pattern Recognition*, vol. 48, no. 10, pp. 2993–3003, 2015.

[11] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, 2017.

[12] H. Xuan, A. Stylianou, and R. Pless, "Improved embeddings with easy positive triplet mining," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2474–2482.

[13] J.-Y. Jaffray, "Bayesian updating and belief functions," *IEEE transactions on systems, man, and cybernetics*, vol. 22, no. 5, pp. 1144–1152, 1992.

[14] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*. CRC press, 2013.

[15] K. P. Murphy, "Conjugate Bayesian analysis of the Gaussian distribution," University of British Colombia, Tech. Rep., 2007.

[16] M. I. Jordan, "The conjugate prior for the normal distribution," University of California, Berkeley, Tech. Rep., 2010.

[17] A. K. Gupta and D. K. Nagar, *Matrix variate distributions*. CRC Press, 2018, vol. 104.

[18] K. Mardia, J. Kent, and J. Bibby, *Multivariate analysis*. AcadeInic Press, Londres, 1979.

[19] D. von Rosen, "Moments for the inverted Wishart distribution," *Scandinavian Journal of Statistics*, pp. 97–109, 1988.

[20] P. Theodossiou, "Financial data and the skewed generalized t distribution," *Management Science*, vol. 44, no. 12-part-1, pp. 1650–1661, 1998.

[21] I. Papastathopoulos and J. A. Tawn, "A generalised Student's t-distribution," *Statistics & Probability Letters*, vol. 83, no. 1, pp. 70–77, 2013.

[22] M. Hazewinkel, "Central limit theorem," *Encyclopedia of Mathematics, Springer*, 2001.

[23] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[24] J. N. Kather, J. Krisam, P. Charoentong, T. Luedde, E. Herpel, C.-A. Weis, T. Gaiser, A. Marx, N. A. Valous, D. Ferber *et al.*, "Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study," *PLoS medicine*, vol. 16, no. 1, 2019.

[25] M. Sikaroudi, A. Safarpoor, B. Ghojogh, S. Shafiei, M. Crowley, and H. Tizhoosh, "Supervision and source domain impact on representation learning: A histopathology case study," in *2020 International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2020.

[26] E. W. Teh and G. W. Taylor, "Learning with less data via weakly labeled patch classification in digital pathology," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2020, pp. 471–475.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] R. Caruana, S. Lawrence, and C. L. Giles, "Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping," in *Advances in neural information processing systems*, 2001, pp. 402–408.

[29] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[30] S. Kalra, H. Tizhoosh, S. Shah, C. Choi, S. Damaskinos, A. Safarpoor, S. Shafiei, M. Babaie, P. Diamandis, C. J. Campbell, and L. Pantanowitz, "Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–15, 2020.