## 使用异步教学模型的逐位硬挖掘在线序列方法

## Asynchronous Teacher Guided Bit-wise Hard Mining for Online HashingAsynchronous Teacher Guided Bit-wise Hard Mining for Online Hashing

使用异步教学模型的逐位硬挖掘在线序列方法

**Abstract**

Online hashing for streaming data has attracted increasing attention recently. However, most existing algorithms focus on batch inputs and instance-balanced optimization, which is limited in the single datum input case and does not match the dynamic training in online hashing. Furthermore, constantly updating the online model with new-coming samples will inevitably lead to the catastrophic forgetting problem. In this paper, we propose a novel online hashing method to handle the above-mentioned issues jointly, termed Asynchronous Teacher-Guided Bit-wise Hard Mining for Online Hashing. Firstly, to meet the needs of datum-wise online hashing, we design a novel binary codebook that is discriminative to separate different classes. Secondly, we propose a novel semantic loss (termed bit-wise attention loss) to dynamically focus on hard samples of each bit during training. Last but not least, we design an asynchronous knowledge distillation scheme to alleviate the catastrophic forgetting problem, where the teacher model is delaying updated to maintain the old knowledge, guiding the student model learning. Extensive experiments conducted on two public benchmarks demonstrate the favorable performance of our method over the state-of-the-arts

流媒体数据的在线哈希最近吸引了越来越多的关注。然而，大多数现有的算法关注于批量输入和实例平衡优化，这些受限于单一数据输入的情况，并且不能匹配在线哈希的动态训练。此外，不断用新的样本更新在线模型将不可避免地造成灾难性的遗忘问题。在本文中，作者提出了一种新颖的在线哈希方法来共同处理上述问题，这个方法称为在线哈希的异步教师引导按比特位难例挖掘。首先，为了满足数据在线哈希的需求，作者设计了一种新颖的二进制码本，可以用来区分不同的类。其次，我们提出了一种新的语义损失（称为按位注意损失），用来在训练期间动态关注每个位的难例样本。最后但同样重要的是，我们设计了一种异步知识蒸馏方案来缓解灾难性遗忘问题，其中教师模型可以延迟更新来保持旧知识，指导学生模型学习。在两个公共基准上进行的广泛实验证明了我们的方法优于最先进的方法

To construct well-spreading codewords in the binary space, in this paper, we propose to generate codewords by utilizing the singular matrix, which avoids the dimension constraints. Since the singular matrix is mutually orthogonal among rows and columns, it naturally preserves good separability among different codewords

为了在二进制空间中构造传播良好的码字，在本文中，我们提出利用奇异矩阵生成码字，从而避免了维度约束。由于奇异矩阵在行和列之间是相互正交的，它自然保留了不同码字之间良好的可分离性

Another drawback of existing OH methods is that all bits of the hashing codes are equally treated, which neglects the fact that for each bit, the fitting degree evolves with time as shown in Fig 1. And for each streaming input, it should be dynamically re-weighted to reflect the accumulated fitting degree (e.g., a hard sample should be assigned with a bigger loss weight when it deviates from the ground truth binary code). To address this issue, we propose a bit-wise attention loss to pay more attention to the hard samples of each bit (especially for the well-fitted ones). This is reasonable since correcting the hard samples helps to generate a more compact distribution centered at the ground truth binary code. Specially, a gaussian distribution is calculated to simulate the fitting degree of each bit (centered at its ground truth binary code). In such a case, the standard deviation of the gaussian distribution can be utilized to indicate the fitting degree (e.g., smaller std reflects a better fitting degree). Then hard samples deviate from the center are affiliated with higher weights to improve the OH learning

现有 OH 方法的另一个缺点是哈希码的所有位都被平等对待，这忽略了一个事实，即对于每个位，拟合度随时间演变，如图 1 所示。对于每个流输入，它应该动态重新-加权以反映累积的拟合程度（例如，当硬样本偏离地面实况二进制代码时，应为其分配更大的损失权重）。为了解决这个问题，我们提出了按位注意损失来更多地关注每个位的硬样本（特别是对于拟合良好的样本）。这是合理的，因为纠正硬样本有助于生成以地面实况二进制代码为中心的更紧凑的分布。特别是，计算一个高斯分布来模拟每个比特的拟合程度（以它的ground truth二进制代码为中心）。在这种情况下，可以利用高斯分布的标准差来表示拟合度（例如，较小的标准差反映了更好的拟合度）。然后将偏离中心的硬样本关联到更高的权重以改善 OH 学习

Catastrophic forgetting is also demonstrated to severely influence the performance in an online setting (Weng and Zhu 2020), since the model is only updated with temporary input. The similarity relationship between the streaming input and existing database is introduced (Lin et al. 2019), which can not well adapt to the dynamic changes in online hashing, since hashing codes of database change with the model update. To address this problem, we introduce knowledge distillation (Hinton, Vinyals, and Dean 2014) into hashing and propose an asynchronous Teacher-Student training scheme.

灾难性遗忘也被证明会严重影响在线环境中的性能（Weng 和 Zhu 2020），因为模型仅使用临时输入进行更新。引入了流输入与现有数据库之间的相似关系（Lin et al. 2019），由于数据库的哈希码随模型更新而变化，因此不能很好地适应在线哈希的动态变化。为了解决这个问题，我们将知识蒸馏（Hinton、Vinyals 和 Dean 2014）引入散列，并提出了一种异步师生培训方案。

The proposed Asynchronous Teacher Guided Hashing is based on the assumption that hashing codes generated by different sets of codewords should share similar semantic relationships, and the asynchronous updating scheme can preserve data distribution on previous data to alleviate the catastrophic forgetting problem. Please note here, different from transferring instances (Ba and Caruana 2014) or local semantic relationships (Yu et al. 2019), our method focuses on the distillation of global semantic relationships.

所提出的异步教师引导散列是基于这样的假设，即由不同的码字集生成的散列码应该共享相似的语义关系，并且异步更新方案可以保留先前数据上的数据分布，以缓解灾难性遗忘问题。请注意，与迁移实例（Ba and Caruana 2014）或局部语义关系（Yu et al. 2019）不同，我们的方法侧重于全局语义关系的提炼。

Overall, in this paper, we propose a combo to jointly solve the problem of bit-wise hard sample mining and catastrophic forgetting in datum-wise online hashing, termed Asychronous Teacher-Guided Hard Mining for Online Hashing (ATHOH).

总的来说，在本文中，我们提出了一种组合来共同解决数据在线哈希中的逐位硬样本挖掘和灾难性遗忘问题，称为在线哈希的异步教师引导硬挖掘（ATHOH）。

The main contributions of ATHOH are three-folds:

We propose a singular-vectors-based method to generate discriminative codewords, which can be flexibly generalized to a flexible number of bits.

We propose a novel bit-wise attention loss to reweight each bit of each streaming input, aiming to dynamically lower weight easy samples and focus on hard samples considering their fitting differences.

We introduce knowledge distillation into the datum-wise online hashing to address the catastrophic forgetting problem, where global semantic relationships are exploited for better distillation, and asynchronous update strategy is utilized to preserve the previous knowledge

ATHOH 的主要贡献有三方面：

我们提出了一种基于奇异向量的方法来生成有区别的码字，可以灵活地推广到灵活的位数。

我们提出了一种新颖的按位注意损失来重新加权每个流输入的每一位，旨在动态降低简单样本的权重，并考虑到它们的拟合差异而关注困难样本。

我们将知识蒸馏引入数据在线哈希以解决灾难性遗忘问题，其中利用全局语义关系进行更好的蒸馏，并利用异步更新策略来保留先前的知识

We introduce the most related works from two main aspects: Online Hashing and Knowledge Distillation.

我们从两个主要方面介绍最相关的作品：在线哈希和知识蒸馏。

Online Hashing has become a hot topic in recent years since it merits in updating the hash functions efficiently by using the streaming data online. Existing OH work can be further categorized into either supervised methods or unsupervised ones. Compared with the unsupervised methods (Leng et al. 2015; Chen, King, and Lyu 2017), supervised methods leverage the label information to learn hashing codes. OKH (Huang, Yang, and Zheng 2013) proposes a structured similarity loss function and learns the hash functions via an online passive-aggressive strategy (Crammer et al. 2006). Similar to OKH, BSODH (Lin et al. 2019) further propose a balanced similarity to reweight positive and negative pairs. However, these methods take pairs or batches as input. To consider the case of an extreme input (a single datum). Cakir et al. propose the online supervised hashing (OSH) which is inspired by the Error-Correcting Output Codes (ECOCs) (Dietterich and Bakiri 1994). Very recently, HCOH (Lin et al. 2018) introduce the Hadamard into online hashing to boost retrieval performance due to the orthogonality of the Hadamard. However, the strict dimension constraint of Hadamard limits its performance in some cases.

Knowledge Distillation is propose by Hinton et al. (Hinton, Vinyals, and Dean 2014) based on a teacher-student framework, which can be further subdivided into two categories: the absolute teacher (Zhou et al. 2018) and the relative one (Liu et al. 2019; Yu et al. 2019). The absolute teacher-based methods focus on transferring instance features from the teacher to the student, including regressing logits before the Softmax layer (Ba and Caruana 2014), the instance features of intermediate layers (Romero et al. 2015), the attention maps (Zagoruyko and Komodakis 2017) and so on (Lee, Kim, and Song 2018). However, the absolute teacher (Koratana et al. 2019) requires student networks has very similar architectures with their corresponding teacher, which is a crucial limitation. Very recently, some relative teacher based methods (Tung and Mori 2019; Tian, Krishnan, and Isola 2020) are proposed to be able to transfer semantic relationships to the student. However, these methods utilize pairwise similarity for semantic relationships, which is captured from a local perspective. Different from the above-mentioned methods, we propose a global semantic relationship-based knowledge distillation loss, where feature proximities calculated on the teacher model are transferred to guide the learning process of the student model.

在线哈希近年来已成为一个热门话题，因为它可以通过在线使用流数据来有效地更新哈希函数。现有的 OH 工作可以进一步分为有监督的方法或无监督的方法。与无监督方法（Leng et al. 2015; Chen, King, and Lyu 2017）相比，监督方法利用标签信息来学习哈希码。 OKH (Huang, Yang, and Zheng 2013) 提出了一种结构化的相似性损失函数，并通过在线被动攻击策略学习散列函数 (Crammer et al. 2006)。与 OKH 类似，BSODH (Lin et al. 2019) 进一步提出了一种平衡相似性来重新加权正负对。但是，这些方法将成对或批次作为输入。考虑极端输入（单个数据）的情况。卡基尔等人。提出受纠错输出代码 (ECOC) 启发的在线监督哈希 (OSH) (Dietterich and Bakiri 1994)。最近，由于 Hadamard 的正交性，HCOH (Lin et al. 2018) 将 Hadamard 引入在线哈希以提高检索性能。然而，Hadamard 的严格维度约束在某些情况下限制了它的性能。

In this paper, we present a novel Asychronous TeacherGuided Bit-wise Hard Mining for Online Hashing (ATHOH) method. The ATHOH aims to jointly address three challenging problems in datum-wise online hashing including the discriminative codebook design, online bit-wise hard mining, and catastrophic forgetting. Specifically, singular vectors is utilized to generate well-spreading codebook for datum-wise hashing. Bit-wise Attention Loss is introduced to reweight each sample to pursue more compact fitting to the ground truth during training. Last but not least, the asynchronous scheme is proposed to alleviate the catastrophic forgetting by delaying update the teacher model. Extensive experiments demonstrate the effectiveness of our method (ATHOH). Acknowledgements This work was supported by the National Natural Science Foundation of China under Project No. 61772158, U1711265 and by National Key Research and Development Project, Ministry of Science and Technology, China under Project No. 2018AAA01013

在本文中，我们提出了一种新颖的异步教师引导的在线哈希（ATHOH）方法。 ATHOH 旨在共同解决数据在线哈希中的三个具有挑战性的问题，包括判别码本设计、在线逐位硬挖掘和灾难性遗忘。具体来说，使用单一向量来生成用于数据散列的良好传播码本。引入了按位注意损失来重新加权每个样本，以在训练期间追求更紧凑的拟合地面实况。最后但同样重要的是，提出了异步方案，通过延迟更新教师模型来缓解灾难性遗忘。大量实验证明了我们的方法（ATHOH）的有效性。致谢这项工作得到了国家自然科学基金项目 No. 61772158、U1711265 和中国科技部国家重点研发项目项目 No. 2018AAA01013 的支持

### 方法论

### 问题定义

给定n个图像样本X={x1，...，xn}，其对应的标签Y={y1，...，yn}，哈希方法旨在将给定的数据点编码为k位二进制代码b∈{-1，1}，这可以保留其语义信息。按照（Cakir, Bargal, and Sclaroff 2017; Lin et al 2019），我们采用基于线性投影的散列函数，其定义为。



其中W = [w1, w2, - -, wk] ∈Rd×k是要学习的投影矩阵，wi∈R1×d负责第i个哈希位，d表示特征的维度。符号函数sgn(x)在输入变量x>0时返回+1，否则返回-1。对于在线学习问题，数据是以流的方式出现的。当前的映射矩阵W是在第t轮输入流数据X t = {x , ..., xt nt }上学习的。在本文中，对于单数据输入的情况，nt被固定为1。

### 基于奇异向量的编码本

所提出的算法是一种两步的散列方法。(1)对于每个类别，我们为其分配一个码字。(2)每个具有相同标签的新输入被编码以适应这个码字。对于多标签的情况，输入数据的目标编码是其相关类标签的码字的平均值。目标函数可以重写为：。

其中，我们表示D为编码本，Yt为xt的类标签集，Dyt为编码本的第y t列。

对于这种方法，正如在介绍中所说明的那样，一个好的编码本对于散列学习是很重要的，而辨别性的编码本可以大大地提高性能。为了产生有效的码表，我们建议通过设计一个单元矩阵来构建码表。单元矩阵的特性是每个单元矩阵的行向和列向都是相互正交的，因此码字在二进制空间中分布良好。单元矩阵的详细构造将在下文中介绍。

在本文中，单元矩阵是通过随机生成的矩阵的辛格值分解得到的。我们首先初始化一个随机矩阵M∈Rm×m，其中m=max(k, c)，c是类的数量。然后我们按行将这个矩阵归一化。并通过M的SVD得到单元矩阵U。最后，我们通过以下方式得到D。



其中U是M的奇异矩阵。给定在公式3中定义的编码本D，我们旨在优化目标函数如下。SGD优化可用于迭代更新公式3。为了使SGD可行，我们放松了非凸和非光滑的符号函数sgn()，如下。



在t轮中，l(x t , W)相对于wi的偏导可以推导为:



讨论。在本节中，我们简要总结了现有的码字生成方法，分为两种类型：基于随机抽样的方法和基于正交性的方法。第一种类型是在期望值中最大限度地增加码字之间的相互距离。编码的每一位都是从对称分布中抽样，如伯努利分布、均匀分布、标准高斯分布和贝塔分布，其公式为:



其中U是一个从对称分布中采样的随机矩阵，u0是这个分布的平均值。我们可以很容易地证明，这些码字之间的距离在期望值中是K/2，其中k是散列码的长度.

在第二类中，码字之间的正交性有助于产生具有鉴别力的编码本。例如，一个相互正交的二进制矩阵是一个最佳编码本。在这种情况下，码字之间有一个固定的距离k/2。当矩阵的维度被严格限制在2的幂级时，Hadamard（Lin等人，2018）是一个特殊的情况，所提出的方法属于这一类。然而，与现有的方法（如Hadamard矩阵）不同，我们利用一个单元矩阵来保留正交性，这避免了维度限制.

为了比较上述这些码本的质量，我们对码本进行了104次合成，然后计算出码字对之间的哈明距离。图3左边是所有码本对的相应距离直方图，图3右边是距离累积直方图。可以看出，基于奇异向量的码字更稳定，更有鉴别力.

### 位数注意损失

新的动态注意力损失被设计为在训练过程中根据每个比特的拟合程度对其进行硬样本挖掘。在单标签情况下，通过将公式4分解为位级，我们可以得到与第i位相关的目标为:



其中，Dyti是码字Dyt的第i位

从公式7中，我们可以看出，在训练过程中，每个即将到来的样本都被平均加权。正如在介绍部分所分析的那样，应该更加关注偏离地面真实二进制代码的硬样本。为了解决这个问题，我们提出了一个拟合度引导的权重来更好地指导学习过程。

在本文中，我们计算累积的统计数据来反映每个比特的拟合程度。在训练阶段，平均数被设定为其基础真值的二进制代码，而标准差则被计算出来。对于即将到来的输入，如果它的计算结果（sgn(W x)）偏离了基础真相，那么它就被视为该位上的硬样本。在这种情况下，它的密度分数会很小，因此我们利用1-p(bi |y)作为损失权重，对硬样本进行巨大的惩罚。下面给出了信息感知损失的细节。首先，我们假设比特遵循基于其标签的高斯分布，其表述为：



其中bi表示b的第i位，Dyi表示码字Dy的第i位，σ表示标准。σyi的值在训练阶段被更新，其计算方法如下。



其中n表示在第t轮中带有标签y的样本数量。 在实际实验中，方差的更新间隔为.

然后我们得到第t次迭代时第i位的损失权重的表述如下：

其中T>0是一个常数，它在[0，1]范围内重新调整权重值，因为概率密度的值在某些极端情况下可能非常大。

最后，比特-维西的注意力损失被表述为。



其中αi是平衡权重.

注意力损失的梯度表述为:



其中η是学习率.

对于多标签的情况，我们计算x在其类别标签集Y下的权重，然后通过这些值的最小值得到总体权重:



在多标签的情况下使用最小值可以帮助散列函数更多关注码字之间的共享位。

### 异步知识蒸馏

对于在线任务，以前的样本不能用来重新训练模型，这导致了灾难性的遗忘。为了解决在线散列中的灾难性遗忘问题，我们通过引入一种新的延迟教学策略，提出了一个异步的师生框架，如图2所示。特别是，我们首先为教师和学生分别生成不同的编码簿。然后，我们在这些编码簿的指导下训练散列函数。为了保留以前的知识，教师模型在两次更新之间被更新和固定，时间间隔为.

至于从教师到学生模型的知识转移，现有的工作集中在绝对知识上，如实例特征或局部语义关系（如成对的相似度）。然而，在数据化的在线散列情况下，由于单一数据的输入，局部语义知识是不可用的。而在我们的方法中，不同的代码书会导致教师和学生之间的输出不一致，使绝对的知识转移失效。因此，我们提出了一种新的知识转移策略，它对全局语义关系进行建模。图4说明了所有提到的蒸馏损失之间的差异。

所提出的基于语义关系的知识提炼允许学生与教师有不同的网络架构，使其更加灵活。全局语义损失被定义为：



其中r和r分别是输入数据x t的散列码与学生和教师模型预先生成的编码簿之间的距离:



其中，Dstu和Dtea分别为教师和学生的编码本。.



其中λ是权重参数.

### 优化

拟议的ATHOH的整体框架是一个教师-学生框架。教师模型由公式4中定义的目标损失来训练，并通过以下方式更新

其中,为权重向量，定义于公式10。- 代表元素相乘。

学生模型是由公式4中定义的客观损失和公式14中定义的新型全局相对知识蒸馏损失来训练的，其更新方式是。

其中，和在公式15中定义，的定义与相同。整个学习过程总结于Alg.1。

