

Sequential Learning for Cross-modal Retrieval

Ge Song^{1,2,3} and Xiaoyang Tan^{1,2,3}

¹College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics

²MIIT Key Laboratory of Pattern Analysis and Machine Intelligence

³Collaborative Innovation Center of Novel Software Technology and Industrialization

{sunge, x. tan}@nuaa.edu.cn

Abstract

Cross-modal retrieval has attracted increasing attention with the rapid growth of multimodal data, but its **learning paradigm under changing environment** is less studied. Inspired by the recent achievement in the field of cognition mechanism on how the human brain acquires knowledge, we propose a new sequential learning method for cross-modal retrieval. In this method, a unified model is maintained to capture the common knowledge of various modalities but are learnt in a sequential manner such that it behaves adaptively according to the evolving distribution of different modalities, and needs no laborious alignment operations among multimodal data before learning. Furthermore, we propose a novel **meta-learning based** method to overcome the catastrophic forgetting encountered in sequential learning. Extensive experiments are conducted on three popular multimodal datasets, showing that our method achieves state-of-the-art cross-modal retrieval performance without any modal-alignment.

1. Introduction

Cross-modal retrieval, aiming to search instances in one modality that display similar content as the query from another modality, has gained increasing attention from both industrial and academic communities due to its wide usage, e.g., sketch-based image retrieval in the criminal investigation. The difficulty of the measurement of content similarity among data from different modalities, which is known as the **heterogeneity gap** [4], makes this task very challenging. Thus, bridging the heterogeneity gap between different modalities plays a key role in cross-modal retrieval.

Many methods [9, 21] have been developed to learn mapping different modalities into a shared feature space, such that the data of different modalities become computationally comparable. Due to the low storage costs and

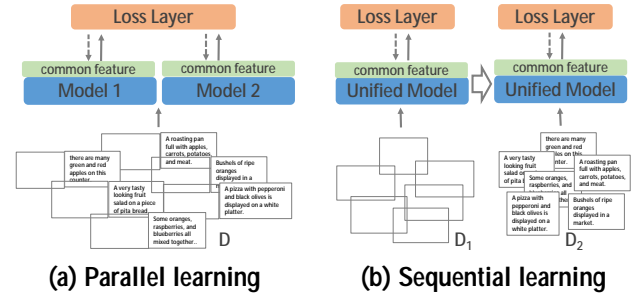


Figure 1. Illustration of the difference between two cross-modal learning paradigms: parallel and sequential. In the parallel paradigm, the whole architecture involves multiple individual sub-models with each responsible for one modality, and well-aligned multi-modal data (e.g., image-text pairs) are needed to jointly train them, while in the sequential paradigms, a single unified model is used to map all modalities into a common feature space, and the model is trained on different modalities sequentially.

the high computational efficiency of binary codes, **hashing-based methods** [19, 18] also have been extended to cross-modality retrieval by embedding the data of interest into a low-dimensional Hamming space. We observe that these methods are built in the same manner: developing individual sub-models for each modality and jointly learning them by aligned multi-modal data. We call this manner as **parallel cross-modal learning** (PCML), which is shown in Fig.1 (a). Despite the effectiveness of this parallel learning paradigm, it is unlikely to be adapted without retraining the whole system under a real-world environment when the underlying distribution of different modalities are gradually changing.

Recent work in cognitive science reveals that when a sequence of multimodal signals stimulates our brain, it is able to automatically integrate the elements from different modalities into one unitary representation [14]. In other words, our brain acquires knowledge or concepts across different modalities in a sequential learning manner (i.e., modality-by-modality). In contrast with PCML, this sequential manner is more practical in real-life scenario: 1)

~~~~~  
~~~~~

~~~~~  
~~~~~

~~~~~

~~~~~

~~~~~

~~~~~

~~~~~

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_













