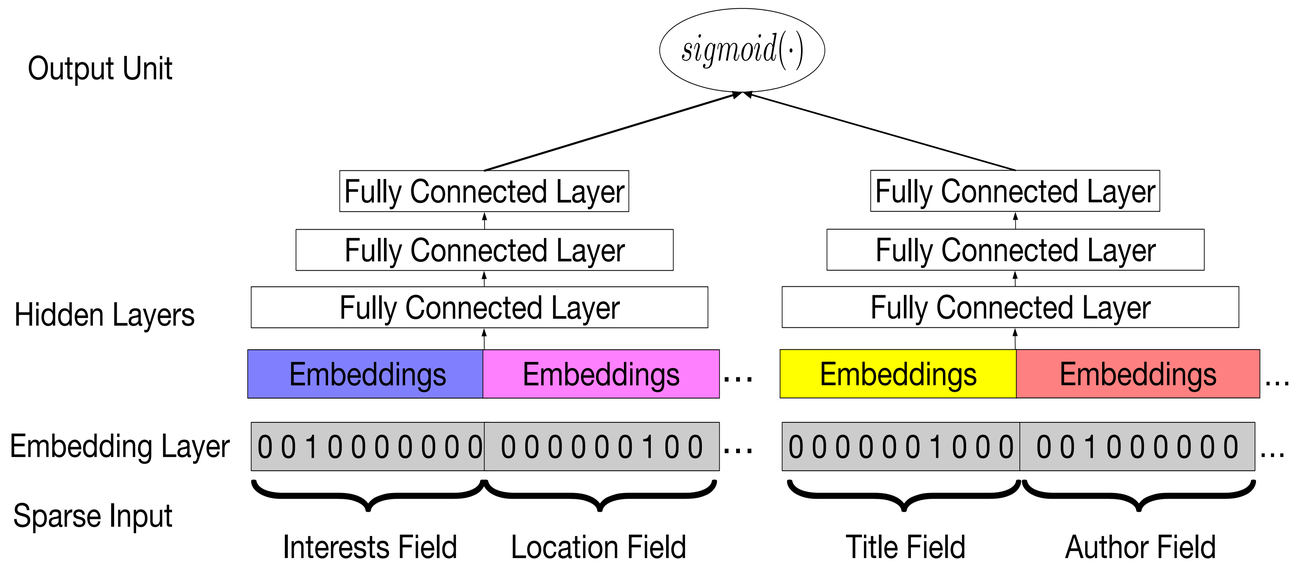
Saec: Similarity-Aware Embedding Compression in Recommendation Systems

Xiaorui Wu[1], Hong Xu[1], Honglin Zhang[2], Huaming Chen[2], and Jian Wang[2]
(1. City University of Hong Kong, 2. Tencent)

Motivation

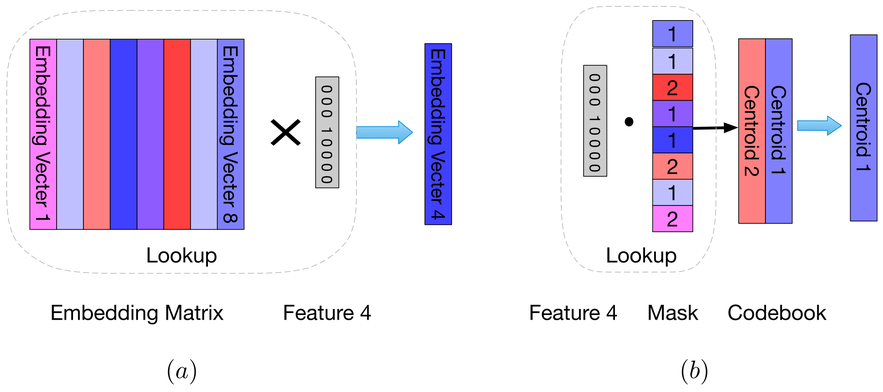
- **Billions of embeddings are needed in practice to represent each unique feature. a billon usually occupies about 360 GB memory**
- **New features are extracted continuously in order to reflect the latest trends and popularities, and the number of embeddings is constantly increasing.**
- **Large number of embedding vectors may cause communication bottleneck in serving.**

Background



In a recommender system, embedding methods are widely used as a critical part of the click-through rate (CTR) prediction model. The main task of the model is to predict the probability of a user clicking on a given item. To do this, it takes as input a set of features about both the user (e.g. gender, interests, etc.) and the item (e.g. title, brand, etc.).

Saec Design



Basic Field-Aware Clustering Framework

Features and the corresponding embeddings are organized into individual fields based on the attributes they reflect. Saec generates a codebook for each field, the features in a field shares the same codebook.

Fast Clustering

Clustering billions of embeddings may take well over an hour which is the typical training interval for recommender systems in practice. We develop a fast clustering method that exploits the feature characteristics to overcome this challenge. .

System Overview

Saec trains the model periodically (say every one hour) with newly generated feature data, and then applies model compression to the initial model. In order to recover the prediction performance, Saec retrains the compressed model before serving the online service.

Evaluation

* Please refer to our paper for details

- Figure 1 shows the performance between the compressed model and original model
- Figure 2 shows reduction ratio in the number of embeddings for the last 24 epochs