# Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Cross-Modal Retrieval

Barbara Rychalska
Synerise
Warsaw, Poland
barbara.rychalska@synerise.com

Jacek Dąbrowski
Synerise
Warsaw, Poland
jack.dabrowski@synerise.com

## ABSTRACT

The recently proposed EMDE (Efficient Manifold Density Estimator) model proved competitive in multimodal settings, achieving state-of-the-art results in session-based and top-k recommendations. In this work we explore its application to Rakuten Data Challenge Task 2: Cross-Modal Retrieval. The aim of the challenge is to match item titles and descriptions with corresponding item photos. We achieve state-of-the-art results in this task using a 2-stage scheme. First, we use a simple OCR-based approach to match text extracted from images to item titles and descriptions. Next, we apply EMDE to match the remaining items which need a more elaborate matching scheme. This approach proves competitive, covering both fine-grained as well as more generalized notion of text-image similarity.

## CCS CONCEPTS

• **Information systems** → **Information Retrieval**.

## KEYWORDS

Rakuten Data Challenge, deep learning, neural networks, cross-modal retrieval

**ACM Reference Format:**
Barbara Rychalska and Jacek Dąbrowski. 2020. Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Cross-Modal Retrieval. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20).* ACM, New York, NY, USA, 4 pages.

## 1 INTRODUCTION

**Rakuten Data Challenge Task 2: Cross-Modal Retrieval.** The aim of the challenge is to link item textual representations to corresponding visual representations. Given a set of product items with their titles and (possibly empty) descriptions, the goal is to predict the best image from among a set of available images, each of which corresponds to some item in the dataset.

The released train set contains 84916 items. The test set is released in two stages: Stage 1 data contains 444 items while Stage 2 test set (used for determining the final results) contains 3995 items. Test sets are drawn from the same distribution as the train set. The difficulty in handling the challenge data consists in high levels of noise, missing data, and variable text languages (mainly French, but also English and German).


*SIGIR eCom'20, July 30, 2020, Virtual Event, China*


The gold standard assignments for the test set are not public and the models are evaluated via an online leaderboard. The metric used for performance evaluation is recall at 1 (Recall@1). The score is understood to be the average of the per-sample scoring of 1 if the image returned matches the title and 0 otherwise.

**Our solution.** Our contributions are as follows: 1) We establish state-of-the-art results for this task with a system composed of two complementary methods: Optical Character Recognition (OCR) and EMDE [5]. We do not exploit any external data sources, relying only on the challenge train set. 2) We analyze the effectiveness and challenges of our method.

**Overall challenge results.** Our approach wins the challenge in Stage 1 by a significant margin - 50.23% or 47.16% (depending on configuration - see Section 4.2) versus next-best contender score of 46.85%. The final results of Stage 2 depend on the ability of all systems to gracefully scale to a larger test set. The overall results are bound to be lower as models have to choose from a much broader selection of item images. Our system wins the challenge in Stage 2 with 34.28% Recall@1, against the next best contender at 31.93% Recall@1.

## 2 RELATED WORK

A recurring challenge in cross-modal retrieval is finding a common representation space, in which the samples from different modalities can be compared directly. [15] exploit information stemming from classification task to make the model learn modality-invariant representations of visual and textual data. [7] use a form of Generative Adversarial Network with a generator for each modality, pitted against a discriminator to eliminate cross-modal differences in representations. Our EMDE approach allows to bypass this problem using analogous sketch representations both for text and image.

A problem related to image-text cross-modal learning is caption generation for images [14] [1]. In our approach we exploit a model for reverse captioning: learning image features from textual representations [4].

## 3 SYSTEM

We apply a two stage approach:

(1) **Optical Character Recognition (OCR).** Many output images are book or newspaper covers which contain author names and titles. This textual data is often straightforwardly repeated in item titles or descriptions. We devise a simple matching procedure between texts detected in images and textual item data. This way, we solve cases which would be very hard to detect for our 2nd stage system (EMDE).

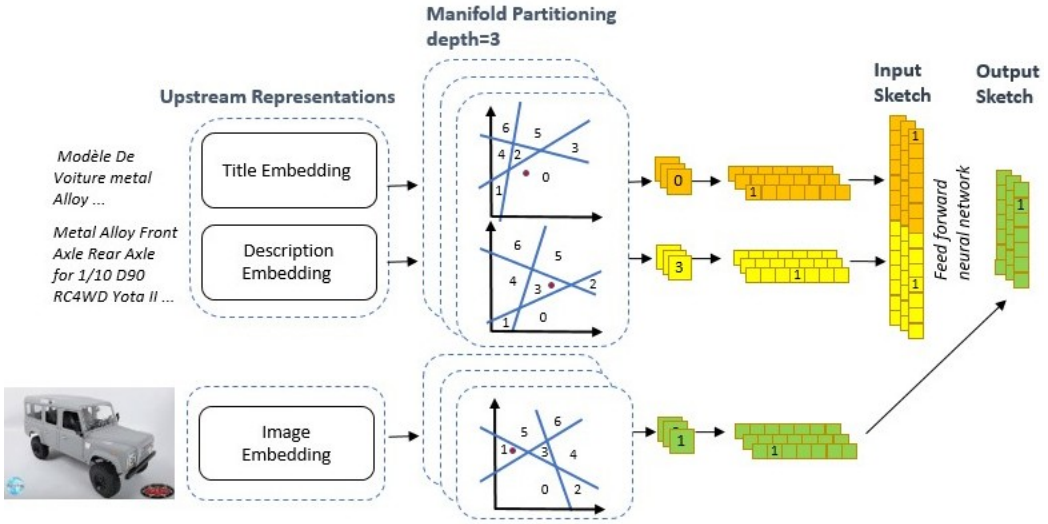


**Figure 1: EMDE with Title and Description sketches at input and Image sketch at output.**

(2) **EMDE.** EMDE is a multimodal density estimator based on intuitions from Local Sensitive Hashing and Count-Min Sketch [3] algorithms. It recently achieved competitive results in recommender systems, reframing recommendations as a density estimation problem.

Efficient Manifold Density Estimator (EMDE) introduced in [5] is a probability density estimator inspired by Count-Min Sketch algorithm (CMS) and local sensitive hashing (LSH). Input data represented by vectors embedded on a manifold obtained by metric representation learning methods is partitioned via a data-dependent LSH method (DLSH). A region of the partitioning corresponding to a region of the vector manifold is analogous to a hash-bucket in CMS. While a single region is large (typically 64-256 regions form a single partitioning covering the whole manifold), multiple independent partitionings allow to obtain a high resolution map of the manifold via intersection or ensembling.

Intuitively, EMDE works by dividing the item embedding space into regions and assigning items to specific buckets based on similarity of their embedding vectors. An encoded representation of a particular item is thus comprised of specific 'buckets' this item fell into, which can then be combined into a joint item representation. Such structures, called *sketches*, are used at both input and output of a simple feed forward neural network. Sketches are prepared independently for each modality (e.g. text, image, interactions) of each item. A schema of EMDE in the setting of the challenge task is displayed in Figure 1.

The retrieval of items from the encoded sketch representation is done at the prediction stage. In our current problem, the following retrieval procedure allows to locate the most probable output image:

(1) Encode all item image representations into sketches (this is done just once).
(2) Identify relevant buckets for each item image in the output.
(3) Select the values from the output which correspond to the relevant buckets.
(4) Count the score as a geometric mean of the values from relevant buckets instead of *min* function in Count-Min Sketch.
(5) Pick the item image with the highest score as the most likely prediction.

The advantages of using EMDE are numerous, especially for large datasets. Representations produced by EMDE have constant size irrespective of the number of samples and original embedding dimensions. This allows to arbitrarily adjust the size of the downstream model. Flexibility comes from the ability to combine various modalities of input data without the necessity to create a joint embedding model. EMDE retrieval procedure allows us to efficiently solve the problem of picking the right image from even very large image sets.

## 4 EXPERIMENTS

### 4.1 Data Preparation

Most title and descripton texts are in French, but English is also common. As the first stage of preparation of the textual data, we apply Huggingface [13] to detect the language of item descriptions (as they are usually longer than titles) and we translate the whole textual data of the detected English items into French.

**Titles.** We tokenize the titles by simple splitting on white characters (in order to preserve integrity of chunks composed of special characters which could be meaningful, such as *N°987*) and use NLTK's [11] `ngrams` package to obtain word unigrams, bigrams, and trigrams.

**Descriptions.** Descriptions are often very verbose or even literary, so useful knowledge must be extracted from them. We strip descriptions of HTML residues (e.g. *&amp*). Then, we proceed to tokenize descriptions using French SpaCy tokenizer [6]. We drop all tokens which do not belong to significant part of speech classes within SpaCy: NOUN, PROPN, VERB, ADJ, X. Then, we proceed to filter the tokens to include just the tokens which also appear in titles to further dispose of unnecessary information.

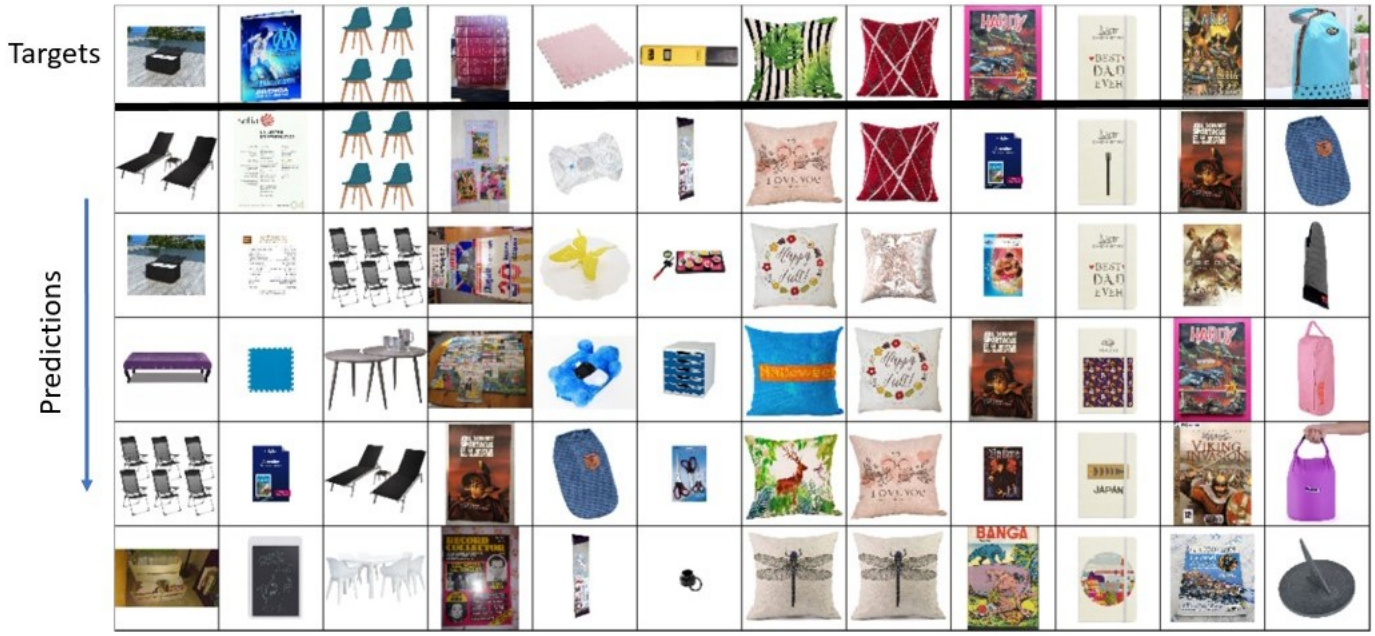


**Figure 2: Examples of EMDE predictions. Top row presents the target images, lower rows show the predicted images at Recall@5. Predictions are arranged in columns. The topmost predicted image is the most probable one, the bottom image is the least probable one.**

**Images.** Although noisy, images are not transformed in any way in our final solution. We experimented with background removal [12] and cropping the white frames around items, but we have not observed any decisive performance gains from these attempts.

**Train/Valid Split.** We do not use any additional data sources apart from the challenge train set. We create our own validation set by sampling out 5000 examples from the train set. We divide this sample into 10 smaller validation sets of size 444, to match the size of the test dataset on Stage 1 leaderboard. We test our system on all 10 small validation subsets and average their results to get an approximate overall result. The variablity of results on each subset can be quite high (up to 4 pp difference between maximum and minimum result), but the average of 10 subsets is empirically found to give a good approximation of final performance.

## 4.2 Optical Character Recognition

OCR is the first step of our system and its results are treated as final, i.e. its predictions are immediately returned as detected matches and removed from the pool of available items for the second stage (EMDE). We test two approaches to OCR:

- **Commodity OCR (`c-OCR`)**. We apply Textsnake [10] to detect regions with text in images. We cut out these regions and run Tesseract [8] on them. We test both English and French Tesseract models with Engine Mode (oem) set to 1 (LSTM neural network) and Page Segmentation Mode (psm) set to 6 (assuming a single uniform block of text). We filter and tokenize titles, descriptions, and texts detected with OCR as in 4.1. We match OCR-detected phrases with title and description texts using 2 simple, complementary strategies:
  - **Overlaps.** We count the maximum overlaps (the percentages of common tokens) between OCR texts and images and descriptions. The overlap threshold of 60% is found to give high accuracy of text-image matching.
    The overlap method works because `c-OCR` is able to locate mainly the biggest, most pronounced texts in images which are most often titles, thus accidentally achieving a form of importance selection.
  - **Inverted Index.** We create an inverted index out of each OCR token and the IDs of items whose titles and descriptions contain the token. We match the image with the item where there are 2 unique matches in the inverted index (the OCR token appears only in a single title/description). In Stage 2 (due to a greater size of the dataset and less matches), we loosen this restriction to 1 unique match.
    This approach works because OCR texts often contain rare tokens such as author surnames which can be uniquely matched in a small dataset.
- **High quality OCR (`hq-OCR`)**. Since `c-OCR` proved effective, we venture on to verify the maximum gain achievable with this approach with the usage of a very high quality OCR. We apply Google OCR to the images and repeat the Inverted Index procedure to match items. We do not use overlap counts since the large amount of discovered texts introduces too much noise. With Google OCR, our system achieves its maximum performance at 50.23 Recall@1. However, note that

**Table 1: Ablation Study displaying R@1 scores across both testing stages. Results with asterisk (*) were estimated on the validation set.**

| Testset | Bare EMDE | EMDE+c-OCR | EMDE+hq-OCR |
|---|---|---|---|
| Stage 1 | 39.61%* | 47.16%* (+7.55pp) | 50.23% (+10.62pp) |
| Stage 2 | 23.29%* | 32.56%* (+9.27pp) | 34.28% (+10.99pp) |

as shown in Table 1, we achieve a state-of-the-art result of 47.16 Recall@1 even without application of Google OCR.

### 4.3 EMDE

**Input Embeddings.** To embed text we use the simple embedding method from [5] with embedding length 1024 and 4 iterations (each token is treated as graph node). We obtain image embeddings from the last layer of Virtex [4]. This ResNet-based model learns to generate visual features from captions, so we suspected that its feature vectors will capture cross-modal image-text relations nicely. We train Virtex from scratch on our train set, using the default parameters from original paper. We feed item images as input, and item titles as captions. The titles are preprocessed by Virtex with SentencePiece [9].

**Configuration.** We encode all embeddings with sketches of depth 42 and width 256. Each modality is encoded separately. In addition to sketches prepared specifically for this task, we also use a precomputed sketch of item titles embedded by Camembert from our other paper on multimodal classification, also from Rakuten Data Challenge 2020 (see [2]). Note that this sketch was precomputed with a trainable version of EMDE, which is also introduced in [2].

For mapping input sketches to output sketches we use a simple neural network. The neural network is composed of 2 feed forward layers of 12,000 neurons with batch normalization. The loss function is regular cross entropy. We use Adam optimizer with a learning rate of $1e^{-4} \cdot 0.8^{epoch}$. Additionally we apply a weight decay of $1e^{-4}$.

**Output Handling.** At output, we require that no single image can be returned for two or more items, excluding the image after it has been matched to an item via maximum score selection over the whole item-image score matrix.

### 4.4 Results

Our final results are presented in Table 1. We find that EMDE alone achieves 39.61/23.29% Recall@1. Due to its usage of squashed output image representations it is practically unable to represent such fine-grained features as image text, focusing on larger concepts of similarity. Indeed, with book/magazine cover matching we observe that EMDE usually predicts another book/magazine image, but the choice is random. Thus, OCR scores do not have to be combined with EMDE scores for a single item, as both solve disjoint parts of the dataset. On the other hand, EMDE is able to accurately model item classes, differentiating between furniture types (tables, closets, desks), or decorative items (cushions, carpets, blankets). The most common errors committed by EMDE consist in selecting an item with a wrong style (e.g. a Christmas-styled cushion instead of a cushion with a leaf pattern), or matching of uncommon items. Figure 2 shows examples of predictions picked randomly from our validation set.

## 5 SUMMARY AND FURTHER WORK

In this paper we present a 2-stage system which achieves state-of-the-art results in SIGIR Rakuten Data Challenge Task 2: Cross-Modal Retrieval. The system uses EMDE to model high-level similarity, aided by OCR to distinguish fine-grained similarity features in the special case of book/magazine covers. In future work we aim to introduce image segmentation to be able to encode image segments into sketches instead of whole image. In this way we will give EMDE the ability to perform well on tasks requiring large attention to detail in images.

## 6 ACKNOWLEDGMENT

Barbara Rychalska was financially supported by grant number 2018/31/N/ST6/02273 funded by National Science Centre, Poland.

## REFERENCES

[1] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal Coherence Modeling for Caption Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 6525–6535.

[2] Dominika Basaj, Barbara Rychalska, Jacek Dąbrowski, and Konrad Gołuchowski. 2020. Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Multimodal Classification.

[3] Graham Cormode and S. Muthukrishnan. 2004. An Improved Data Stream Summary: The Count-Min Sketch and Its Applications. In *LATIN 2004: Theoretical Informatics*, Martín Farach-Colton (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 29–38.

[4] Karan Desai and Justin Johnson. 2020. VirTex: Learning Visual Representations from Textual Annotations. *arXiv preprint arXiv:2006.06666* (2020).

[5] Jacek Dąbrowski, Barbara Rychalska, Michał Daniluk, Dominika Basaj, Piotr Babel, and Andrzej Michałowski. 2020. An efficient manifold density estimator for all recommendation systems. https://arxiv.org/abs/2006.01894

[6] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[7] Peng Hu, Dezhong Peng, Xu Wang, and Yong Xiang. 2019. Multimodal adversarial network for cross-modal retrieval. *Knowledge-Based Systems* 180 (2019), 38 – 50. https://doi.org/10.1016/j.knosys.2019.05.017

[8] Anthony Kay. 2007. Tesseract: An Open-Source Optical Character Recognition Engine. *Linux J.* 2007, 159 (July 2007), 2.

[9] Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. 66–71. https://doi.org/10.18653/v1/D18-2012

[10] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. 2018. TextSnake: A Flexible Representation for Detecting Text of Arbitrary Shapes. In *The European Conference on Computer Vision (ECCV)*.

[11] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1* (Philadelphia, Pennsylvania) *(ETMTNLP '02)*. Association for Computational Linguistics, USA, 63–70. https://doi.org/10.3115/1118108.1118117

[12] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar Zaiane, and Martin Jagersand. 2020. U2-Net: Going Deeper with Nested U-Structure for Salient Object Detection. *Pattern Recognition* 106, 107404.

[13] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[15] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep Supervised Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.