SIGIR 2020 E-Commerce Workshop Data Challenge Overview

Hesam Amoualian*
hesam.amoualian@rakuten.com
Rakuten Institute of Technology, Paris

Parantapa Goswami*
parantapa.goswami@rakuten.com
Rakuten Institute of Technology, Paris

Laurent Ach
laurent.ach@rakuten.com
Rakuten Institute of Technology, Paris

Pradipto Das
pradipto.das@rakuten.com
Rakuten Institute of Technology, Boston

Pablo Montalvo
pablo.montalvo@rakuten.com
Rakuten Institute of Technology, Paris

ABSTRACT

This paper presents an overview of the SIGIR 2020 eCom Rakuten Data Challenge. For this data challenge we make available a multi-modal dataset of 99 thousand product listings from Rakuten France catalog. Each product in the dataset contains a textual title, a (possibly empty) textual description and an associated image. Two tasks are proposed, namely large-scale multi-modal classification and cross-modal retrieval. Among the data sets, around 85 thousand products and their corresponding product type category are released as training data, around 9.5 thousand products and 4.5 thousand products are released as the test sets for the multi-modal classification and cross-modal retrieval tasks respectively. The evaluation is run in two stages to measure system performance, first on 10% of the test data, and then on the rest 90% of the test data. The different systems are evaluated using macro-F1 score for the multi-modal classification task and recall@1 for the cross-modal retrieval task. Sixteen independent teams submitted system outputs in the proposed tasks. The top performance obtained at the end of the second stage is 91.94% macro-F1 and 50.23% recall@1 for the two tasks respectively.

KEYWORDS

e-commerce datasets, multimodal classification, cross-modal retrieval

ACM Reference Format:
Hesam Amoualian, Parantapa Goswami, Laurent Ach, Pradipto Das, and Pablo Montalvo. 2020. SIGIR 2020 E-Commerce Workshop Data Challenge Overview. In *Proceedings of ACM SIGIR Workshop on eCommerce (SIGIR eCom'20).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Rakuten Multi-modal Product Data Classification and Retrieval challenge is organized by Rakuten Institute of Technology, the research and innovation department of Rakuten group. This challenge focuses on the topic of large-scale multi-modal (text and image) classification, where the goal is to predict each product's *type code* as defined in the catalog of Rakuten France, and cross-modal retrieval, aiming to retrieve the most relevant image of a product given the textual title and description.

The cataloging of product listings through some type of text or image categorization is a fundamental problem for any e-commerce marketplace, with applications ranging from personalized search and recommendations to query understanding. Manual and rule-based approaches to categorization are not scalable since commercial products are organized in many and sometimes thousands of classes. When actual users categorize product data, it has often been seen that not only the text of the title and description of the product is useful but also its associated images.

Advances in this area of research have been limited due to the lack of real large-scale multimodal product data from actual commercial catalogs. This data challenge presents several interesting research aspects due to the intrinsic noisy nature of the product labels and images, the size of modern e-commerce catalogs, and a highly skewed data distribution. We hope that by making the data available to the participants, these tasks will attract more research institutions and industry practitioners, who do not have the opportunity to contribute their ideas due to the lack of an actual commercial e-commerce catalog data.

2 PROBLEM DESCRIPTION

In the taxonomy of Rakuten France, products sharing the same product type code share the same exact array of attributes fields and possible values. Product type codes are numbers that match a generic product name, such as 1500 - `Watches`, 120 - `Laptops`, and so on. In that sense, the type code of a product is its category label.

For example, in the product catalog of Rakuten France, a product with a French title *Klarstein Présentoir 2 Montres Optique Fibre* is associated with an image and sometimes with an additional description. This product is categorized with a `product type code` of **1500**. There are other products with different titles, images and with possible descriptions, which are under the same `product type code`. Given these information on the products, like the example above, this challenge proposes participating teams build and submit systems that classify previously unseen products into their corresponding `product type codes`.

The main challenges for this task are as follows:

(1) **Multi-modal classification**. Given a training set of products and their `product type codes`, predict the corresponding `product type codes` for an unseen held out test set of

*Both authors contributed equally to this work.

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

Table 1: Three samples in the X_train.tsv file

Integer_id	Title	Description	Image_id	Product_id
2	Grand Stylet Ergonomique Bleu Gamepad ...	PILOT STYLE Touch Pen ...	938777978	201115110
40001	Drapeau Américain Vintage Oreiller ...	Vintage American Flag Pillow Cases ...	1273112704	3992402448
84915	Gomme De Collection 2 Gommes Pinguin ...	NaN	684671297	57203227

Figure 1: Images of the three example products shown in Table 1.

(a) `image_938777978_product_201115110.jpg`; Category: Entertainment

(b) `image_1273112704_product_3992402448.jpg`; Category: Household

(c) `image_684671297_product_57203227.jpg`; Category: Books

products. The systems are free to use the available textual titles and/or descriptions whenever available and additionally the images to allow for true multi-modal learning.

(2) **Cross-modal retrieval**. This task is more challenging than a classification task since systems have to predict the correct image for a product given its textual content.

The difficulty in solving the tasks stems from the following observations:

- Highly imbalanced number of samples within the classes.
- Length of titles can vary – they can sometimes consist of one or two words.
- Descriptions, when present, may be a verbose representation of the product rather than a very specific one with precisely defined attributes for the product.
- Images may not be "clean". Some images could be of low quality, while some images may have text in them as often found in a banner.

3 DATA DESCRIPTION

Rakuten France has released approximately 99K product listings in `tsv` format, including a training (84,916) and two test sets (9,372 samples for the classification and 4,440 samples for the retrieval task). The training and test splits have been obtained using stratified sampling over categories. Each product in the dataset consists of product titles, product descriptions, product images and their corresponding *product type code*. The dataset is distributed over 27 unique product type categories.

The complete catalog of products of Rakuten France is much larger than 99 thousand listings and contains much more than 27 product type codes. Among all the available product type codes, initially 27 are manually identified based on how often the products belonging to these type codes need to be categorized and how much GMV[1] they generate. This choice makes it more challenging to do the classification on only one modality. For each of these identified type codes, a 10% of the product listings are randomly sampled among all the products that it contains.

The training data file is in a tab-separated values (`tsv`) format where each line contains a product title, (possibly empty) description, product id, id of the associated image and its corresponding product type code. Additionally an image folder is supplied containing all the training images. One can use the image id and product id to obtain the image files from the image folder. The test data file for multimodal classification contains all the fields as training data file except the product type code, and similarly image id and product id can be used to obtain the corresponding image files from the test image folder. The test data file for the cross-modal retrieval task contains only product title, (possibly empty) description and the product id. Image files for the products for this test set is also provided, but the link between the products and their corresponding image files are not provided in this case.

Table 1 displays three different lines of the training file, and Figure 1 shows the corresponding images for these three products. The examples are selected from the head, torso, and tail of the distribution, where two of which have descriptions and one without.

Also, `catalog_english_taxonomy.tsv` is a tab-separated file containing the correspondence between each `product type code`

[1]Gross Merchandise Volume (GMV) is the total monetary value for merchandise sold through a particular marketplace over a certain time frame.

(abbreviated `Prdtypecode`) and its top level category in English. For example:

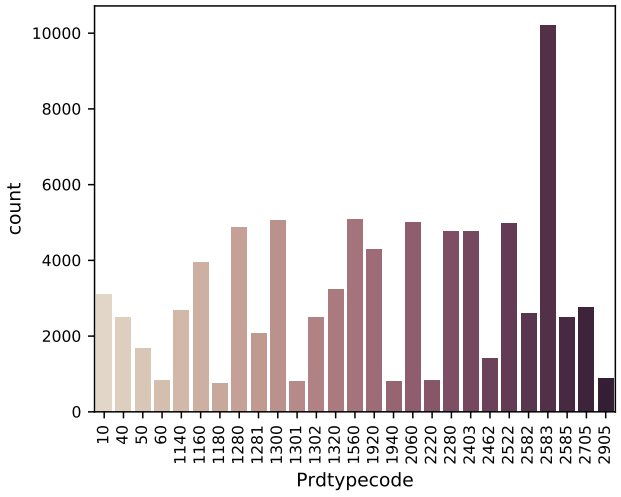
`Prdtypecode`	`Top level category`
`2280`	`Books`
`1280`	`Child`

It should be noted that the product titles and descriptions are for the vast majority written in French (99%), although, one can find some outlying samples related to other languages like English, German, and Spanish. Almost 35% of the products contain an empty description. The images are all squares of dimensions 500 x 500 px^2, which can have white or black borders included.

3.1 Data Characteristics

The product listing distribution in this dataset over the 27 product type code classes is highly imbalanced. Figure 3 shows the distribution of product frequency in the training dataset across all the product type codes. The largest class contains 12% of the products in the entire training dataset, whereas the smallest one contains 0.9%.

Figure 3: Product type code distribution



The average product title length in the training dataset is 12 tokens with the maximum title length of 40 tokens. The maximum length of concatenated title and description is 512.

4 EVALUATION

4.1 Evaluation Metric

Since in this challenge, we are dealing with many classes with highly asymmetric number of samples, an item weighted metric used to rank the participants will not reveal the deficiencies of the classification algorithms.

Task 1 We will use the **macro-F1 score** to evaluate product type code classification on held out test samples. The score is understood as the arithmetic average of per-product type code F1 score.

Task 2 For the cross-modal retrieval task, the systems will be evaluated on **recall at 1 (R@1)** on held out test samples. 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.

4.2 Evaluation Phases and Timeline

This data challenge has been held in two phases which includes model building and model evaluation. In each phase, there is a separate test set for each task.

Stage 1 - Model Building Participants built and tested models on the training data. The models are evaluated on a 10% subset of the test set. This phase was open for a little under three months.

Stage 2 - Model Evaluation The models are evaluated on the remaining 90% of the test set. This phase was open for eight days.

4.3 Scoreboard

System performance of each task is shown in different scoreboards as shown in Fig. 5. The overall scoreboard is divided to display performance for stage 1 and stage 2. Percentage scores (macro F1 scores for task 1 and recall@1 for task 2) are shown for the **latest** submissions from each participating team. The corresponding file submission time is shown as well so participants can refer the scores to which submission it belongs to. Each portion of the scoreboard is sorted by submission score.

5 RIT-PARIS BASELINE METHOD

In this section, we introduce the baseline algorithm for the classification task. Rakuten Institute of Technology (RIT) benchmark model, dubbed `RIT-Paris Baseline` is based on the Multimodal Bitransformers [4], which has been recently a very popular approach for multi-modal classification task where the textual dataset is accompanied by other modalities such as images. Recent methods in representation learning such as BERT [1] for natural language processing, have gained dramatic improvements in classification and other text-only related tasks.

This model combines two pre-trained networks. For the image network, it utilizes a ResNet-152 [2] with average pooling of $K \times M$ grids in the image, yielding $N = KM$ output vectors of 2048 dimensions for each image. Input images are normalized, center-cropped, and resized at 224x224. For language representation, the model uses a bidirectional transformer architecture with pre-trained BERT. The input of this network is contextual embeddings and each contextual embedding is computed by the sum of separate D-dimensional token, segment, and position embeddings from the text. The model then performs an affine transformation $I_n = W_n f(img, n)$ to map each of the N image embeddings to the D-dimensional token input embeddings. Here $f(, n)$ is the n^{th} output of the image encoder and $W_n \in \mathbb{R}^{2048 \times D}$ the network weights to be learned.

As an example, for a single text and single image input, they assign token inputs to one segment ID and image embeddings to another. This architecture can be generalized to a desired number of modalities. Since pre-trained BERT itself has only two segment embeddings, in those cases we initialize additional segment embeddings as $s_i = (s_0 + s_1)/2 + \epsilon$ where s_i is a segment embedding for $i \geq 2$ and $\epsilon \sim N(0, 1e^{-2})$. For the classification task, the output of

the final layer of bi-transformer serves as an input for a classification network. For multi-class task, this layer is a softmax on top of the logits and will be trained with a regular cross-entropy loss.

For the implementation of this model, we use the `MMBT` library from Pytorch-Transformers [2] [9]. We set Multilingual DistilBert [8] as the language pre-trained model and ResNet-152 for the image part. In this challenge, we try to categorize 27 different imbalanced classes of text and associated images (Figure 1). For training and fine-tuning the model, we split the training set to 90% train and 10% development and train the model on a machine with 4 GPU cards. *All model components in the baseline model have been used as released by their respective authors without any changes in default parameter values and no tuning has been performed.*

Following are the macro-F1 scores obtained using the `RIT-Paris Baseline` model for the test set of the first phase and the second phase of challenge on only the multimodal classification task:

	Phase1	Phase2
Score	0.8705	0.8536

Figure 4 shows that the baseline model can achieve a very high score for most of the categories. However, not all classes have been easier to classify by the model. According to the results, products related to Child (1280,1281) and Entertainment (1180) top-level categories, have the worst score (0.73, 0.59, and 0.66 accordingly).

6 SYSTEM DESCRIPTIONS

The team names, affiliation and corresponding system description paper title available on the workshop website is summarized in Table 2. In total 8 papers were submitted in this data challenge. Among which 6 papers from 5 teams are accepted. Following are the brief descriptions of these systems submitted by 5 teams.

Team Synerise AI submitted two systems for the two tasks of this challenge. For the classification task, a 2-stage scheme (separate pre-training of each modality using a trainable Efficient Manifold Density Estimator + multimodal fusion) is used yielding a 89.78 macro F1 score. For the cross-modal retrieval task also, the team employs a 2-stage scheme (OCR + Efficient Manifold Density Estimator). This methodology is able to yield a 34.28 recall@1 score and put the team on the first position for this task.

Team Beantown submitted a system which first fine-tunes feature extractors from text (CamemBERT) and image modalities (BiT) respectively, then applies Highway network based fusion to obtain multimodal features. These features are then used to train a classifier for the classification task and in a similarity search method to retrieve product images from their text titles for the cross-modal retrieval task. This system resulted in 90.22 macro F1 score for the first task and 23.3 recall@1 score for the second task.

Team pa_curis submitted a system which uses pre-trained CamemBERT for text and pre-trained ResNet152 for image modality to learn unimodal features and then deploys late decision level fusion to combine the modalities. Using different versions of text and image classifiers and fusion techniques 12 classifiers are obtained and finally an ensemble strategy of majority voting is applied. This final classifier ended up in the first position for the multimodal classification task with 91.44 macro F1 score.

Team Alto submitted a system where ResNet is used to extract image embeddings, a combination of BERT-based transformer and biLSTM is used to encode text and finally a co-attention block is used to jointly reasons between words and images. The learned image and text vectors are then concatenated and passed through a softmax layer for classification. Furthermore an ensemble is created by stacking multimodal models with different base architecture and then using an machine learning method that leveraged each individual model's strengths. This strategy yielded 90.87 macro F1 score for the classification task.

Team Transformers proposes a deep multimodal multi-level boosted fusion learning framework. Text features are extracted using two different transformer models, namely CamemBERT and FlauBERT. Images are extracted using SE-ResNeXt. Then these features are combined using addition, concatenation, and attention maps. Finally boosted late-fusion is used to learn a combination of weights of single-modality models and the multi-modal model.

6.1 Results

Altogether fifteen teams submitted system results in this data challenge, although the number of registrations has been around one hundred. The submitted systems are scored against the gold standard using the metrics as defined in Section 4.1. The final results are summarized in Figure 5. The left hand side of Figure 5 shows the scoreboard in phase 1 and the right hand side shows phase 2 results. Section 4.2 describes the evaluation phases.

The ranks for the multimodal classification task (task 1) does not change much between phase 1 and phase 2, except for team pa_curis who gained 5 rank positions (from 6^{th} to 1^{st}), and team Alto who gained 1 rank position (from 3^{rd} to 2^{nd}). Similarly for the cross-modal retrieval task (task 2) the ranks mostly remain unchanged, except for team Beantown who gained 1 rank position (from 4^{th} to 3^{rd}), and team Alto who gained 1 rank position (from 5^{th} to 4^{th}).

7 FINAL COMMENTS

In this data challenge the two tasks were designed to have gradation in difficulty. The classification task is a much easier task and the submission from team `Beantown` shows a macro-F1 score of 87.2% by just using CamemBERT on the textual modality. They also show that using both modalities is helpful in the final classification.

The majority of submissions have used pre-trained models that serve as better priors for initialization of embedding vectors. CamemBERT [7] and FlauBERT [6] have been on the forefront of the such model choices due to the underlying French corpora on which these models have been trained. Similarly the majority of model choices for image modeling has been using ResNet and its variants [3] in addition to some very recent models such as Google's Big Transfer (BiT) model [5] for pre-trained initialization. Finally the top scoring

[2] https://huggingface.co/transformers/summary.html

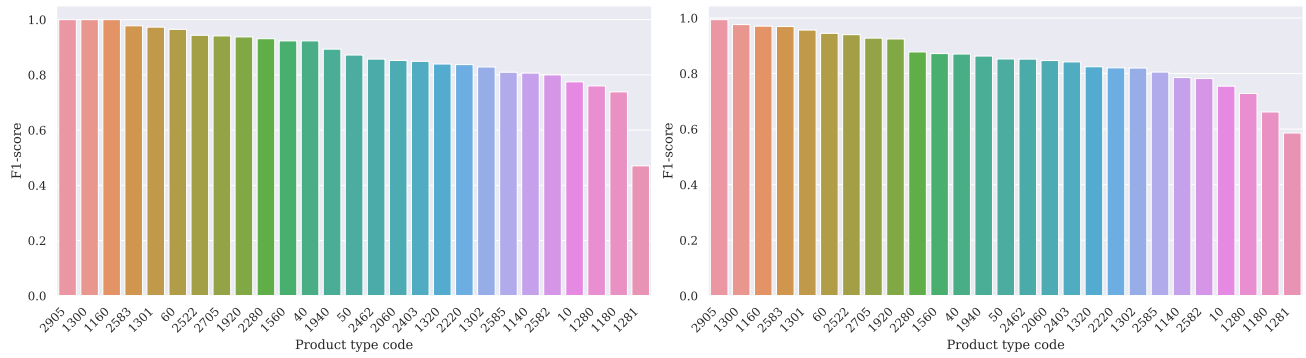


Figure 4: F1 scores from `RIT-Paris Baseline` model on the test sets from phases 1 (left) and 2 (right) for the classification task.

Team Name	Affiliation	Title of System Description Paper
Synerise	Synerise AI	Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Cross-Modal Retrieval
Beantown	Rakuten USA	CBB-FE, CamemBERT and BiT Feature Extraction for Multimodal Product Classification and Retrieval
pa_curis	Ping An Technology	A Multi-Modal Late Fusion Model for E-Commerce Product Classification and Retrieval
Alto	Xerox PARC	Large Scale Multimodal Classification Using an Ensemble of Transformer Models and Co-Attention
Synerise	Synerise AI	Synerise at SIGIR Rakuten Data Challenge 2020: Efficient Manifold Density Estimator for Multimodal Classification
Transformers	Walmart Labs	Deep Multi-level Boosted Fusion Learning Framework for Multi-modal Product Classification

Table 2: List of participants who submitted system description papers.

system has shown some novelty in fusing the outputs from the unimodal models using co-attention, highway networks and gradient boosted trees.

In conclusion, we hope that this dataset can be a de-facto resource for multi-modal classification and cross-modal retrieval on e-commerce data. We will release the dataset using proper channels on https://rit.rakuten.co.jp/data_release/.

8 LEGAL NOTICE

By express derogation from any preexisting or future contractual documents and/or terms and conditions pertaining to the Rakuten Data Challenge occurring on the occasion of the SIGIR 2020 Workshop on eCommerce ("Rakuten Data Challenge"), the participant ("Participant") acknowledges that the study data ("Study Data") uploaded by Rakuten France (the "Provider") on the occasion of the Rakuten Data Challenge is strictly considered as confidential information. The Participant unreservedly undertakes to (i) hold in strict confidence and not disclose to any third party all or part of the Study Data, (ii) use the Study Data for the sole purpose of the good performance of the Rakuten Data Challenge (the "Purpose"), (iii) not use, apply, reveal, report, publish, extract all or part of the Study Data or otherwise disclose all or part of the Study Data in any circumstances for a purpose other than the Purpose, excluding notably commercial or technical use of any kind. As of the termination of the Rakuten Data Challenge, the Participant shall immediately cease any use of the Study Data unless otherwise agreed by the Provider. The present specific terms shall remain in full force and effect until the termination of the Purpose and for a period of two (2) years following the termination date of the Purpose.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.
[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
[4] D. Kiela, S. Bhooshan, H. Firooz, and D. Testuggine. Supervised multimodal bitransformers for classifying images and text, 2019.
[5] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby. Big transfer (bit): General visual representation learning, 2019.
[6] H. Le, L. Vial, J. Frej, V. Segonne, M. Coavoux, B. Lecouteux, A. Allauzen, B. Crabbé, L. Besacier, and D. Schwab. Flaubert: Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 2479–2490. European Language Resources Association, 2020.
[7] L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.
[8] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2019.
[9] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771, 2019.

Figure 5: Phase 1 scoreboard (left) and phase 2 scoreboard (right).

Task 1: Multimodal Classification

Rank	Team Name	Last Submission	Macro-F1 score
1	Transformers	2020 Jul 13 06:53:43	91.94
2	zenit84	2020 Jun 27 17:33:34	91.63
3	Alto	2020 Jul 14 15:40:57	91.63
4	Beantown	2020 Jul 15 23:47:03	90.89
5	Synerise AI	2020 Jul 07 13:45:38	89.72
6	pa_curis	2020 Jul 15 17:38:26	89.65
7	RIT-Paris Baseline	2020 Jul 15 10:48:14	87.05
8	tester	2020 Jun 24 17:53:50	86.94
9	testers	2020 Jul 08 16:19:24	85.87
10	MMG_AI_TEAM	2020 Jul 15 11:27:08	84.81
11	DeepData	2020 Jun 18 09:11:58	84.32
12	overfiTTers	2020 Jul 12 12:57:20	81.9
13	Team MLG	2020 Jul 15 05:01:11	65.8
14	qrudraksh	2020 May 27 20:03:26	58.0
15	7ate9	2020 Jun 10 04:39:36	53.29

Task 1: Multimodal Classification

Rank	Team Name	Last Submission	Macro-F1 score
1	pa_curis	2020 Jul 21 10:25:24	91.44
2	Alto	2020 Jul 23 21:35:59	90.87
3	Transformers	2020 Jul 23 16:23:52	90.53
4	zenit84	2020 Jul 23 22:22:36	90.39
5	Beantown	2020 Jul 22 03:58:44	90.22
6	Synerise AI	2020 Jul 17 04:42:21	89.78
7	MMG_AI_TEAM	2020 Jul 22 13:22:00	86.94
8	RIT-Paris Baseline	2020 Jul 18 09:19:53	85.36
9	Team MLG	2020 Jul 17 01:22:23	64.48

Task 2: Cross-modal Retrieval

Rank	Team Name	Last Submission	Recall@1 score
1	Synerise AI	2020 Jul 01 12:30:48	50.23
2	changer	2020 Jul 12 11:14:59	46.85
3	pa_curis	2020 May 27 09:34:52	41.89
4	Beantown	2020 Jul 15 20:38:29	38.96
5	Alto	2020 Jul 03 01:49:35	38.29
6	MMG_AI_TEAM	2020 Jul 15 11:29:07	27.25
7	kenneth	2020 Jun 16 20:12:10	1.35

Task 2: Cross-modal Retrieval

Rank	Team Name	Last Submission	Recall@1 score
1	Synerise AI	2020 Jul 17 19:04:22	34.28
2	changer	2020 Jul 21 16:59:52	31.93
3	Beantown	2020 Jul 22 17:05:28	23.3
4	Alto	2020 Jul 20 23:40:02	19.99
5	pa_curis	2020 Jul 23 13:19:58	19.74
6	MMG_AI_TEAM	2020 Jul 20 12:39:02	15.77