Towards MSR-Bing Challenge: Ensemble of Diverse Models for Image Retrieval

Quan Fan, Hanqiu Xu, Ruowei Wang, Shengsheng Qian, Ting Wang, Jitao Sang, Changsheng Xu

Institute of Automation, Chinese Academy of Sciences

Chinese-Singapore Institute of Digital Media

October 07, 2013

Review of The Task

- Task: Develop a score system to assess the query-image relevance
  - For each image-query pair, output a floating score indicating how effective the query is used to describe the image.

Input			**Output**
$\boldsymbol{x}$ 	$\boldsymbol{q}$ barack obama		$\boldsymbol{s(q, x)}$

- Evaluation
  - For one specific query q, image rank list is generated by sorting the relevance scores $s(q,\cdot)$;
  - Ave. DCG@25 over all test queries is employed as the final evaluation metric.

Data Set

- Training Set:

image ID <tab> query <tab> click count

fall :113;fall pictures :85;fall leaves :48;fall

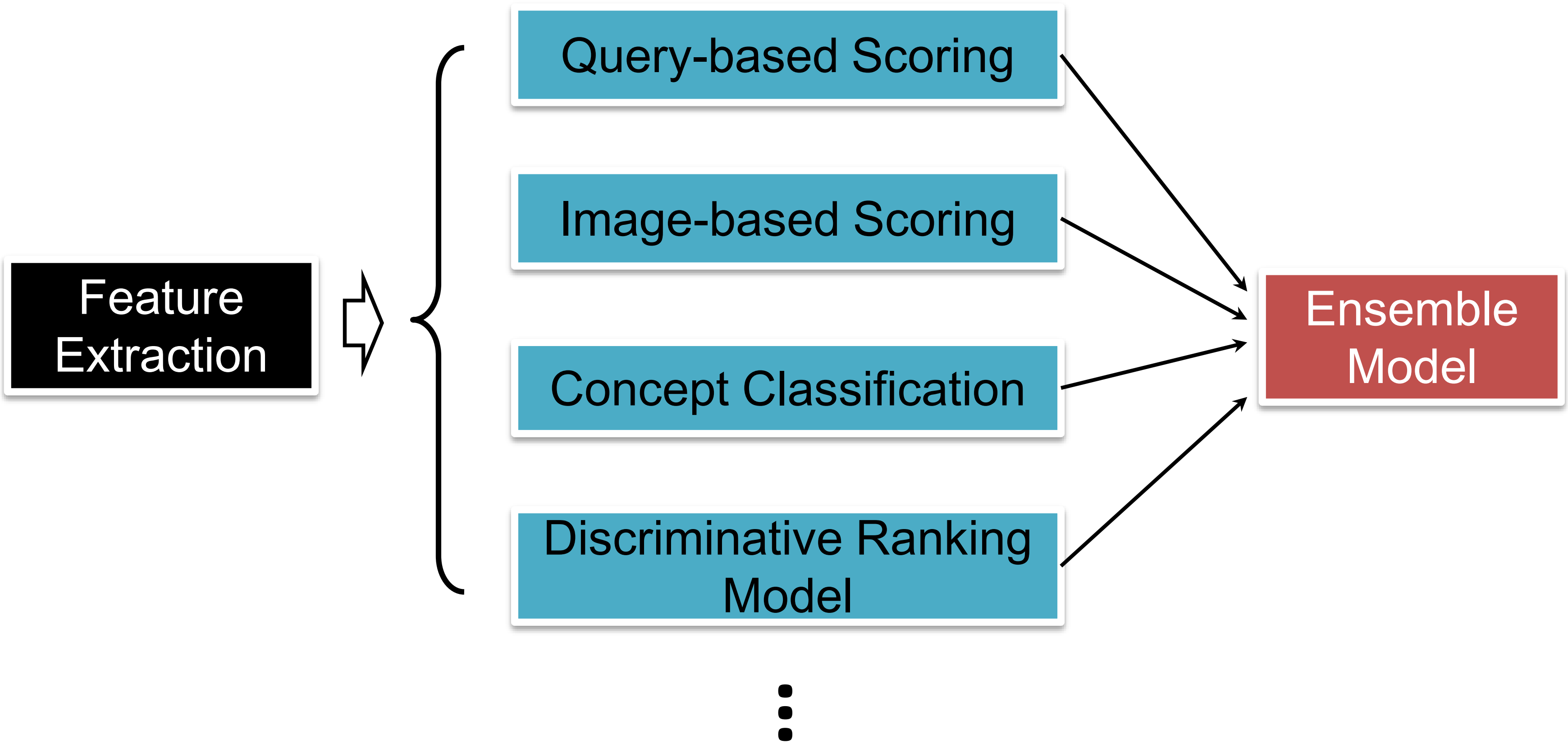
- Development Set:

query <tab> image ID <tab> judgment (Excellent/Good/Bad)

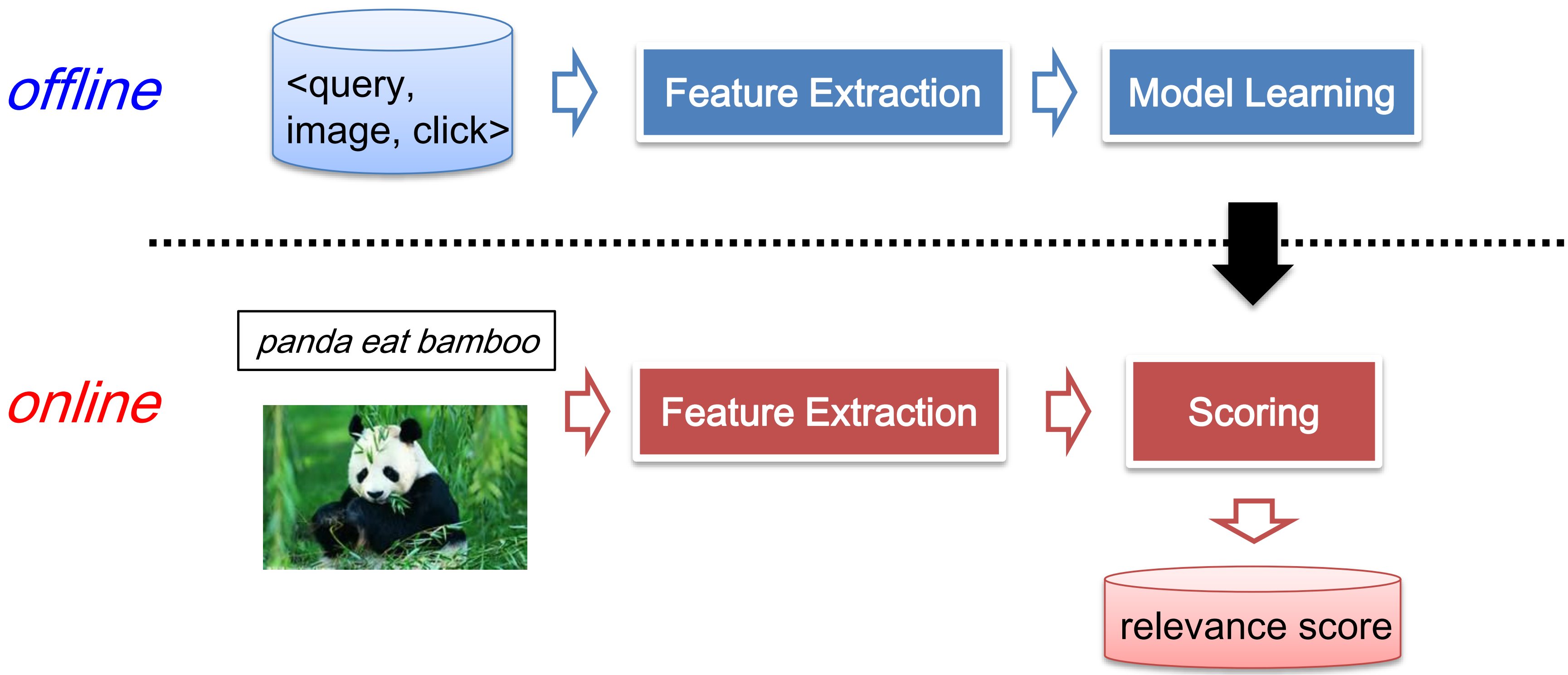
"katrina darling" **img1504** Excellent

"katrina darling" **img2817** Bad

Our Solution



System Illustration



Feature Extraction

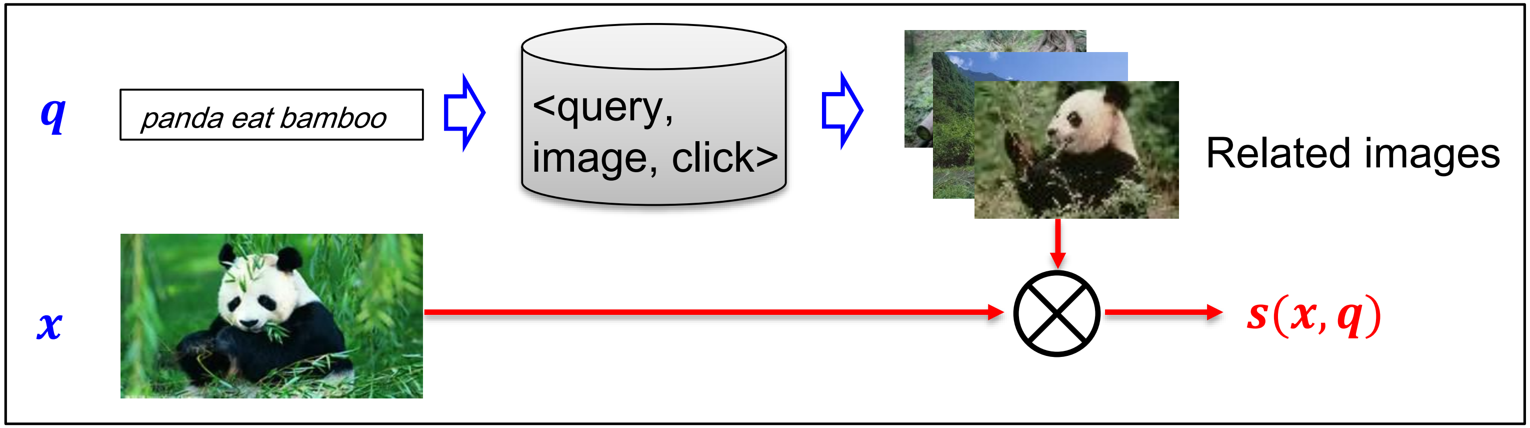
- Query Features
  - BoW representation: $q = (q_1, \dots, q_T) \in \mathbb{R}^T$, $T = 100{,}000$;
  - Feature value: word occurrence

- Image Features $(d = 22{,}312)$
  - Local Features
    - HOG+LLC+SPM
    - LBP
  - Global Features
    - Color moment
    - Edge histogram
    - Wavelet texture feature
    - GIST feature

#1 Query-based Scoring

- **Motivation**
  - Transfer to measure image-image visual similarity.
- **Solution**
  - Retrieve the related image set X by issuing the test query q into the training set;
  - Calculate query-image relevance by aggregating the visual similarities between test image x and the query-related images.

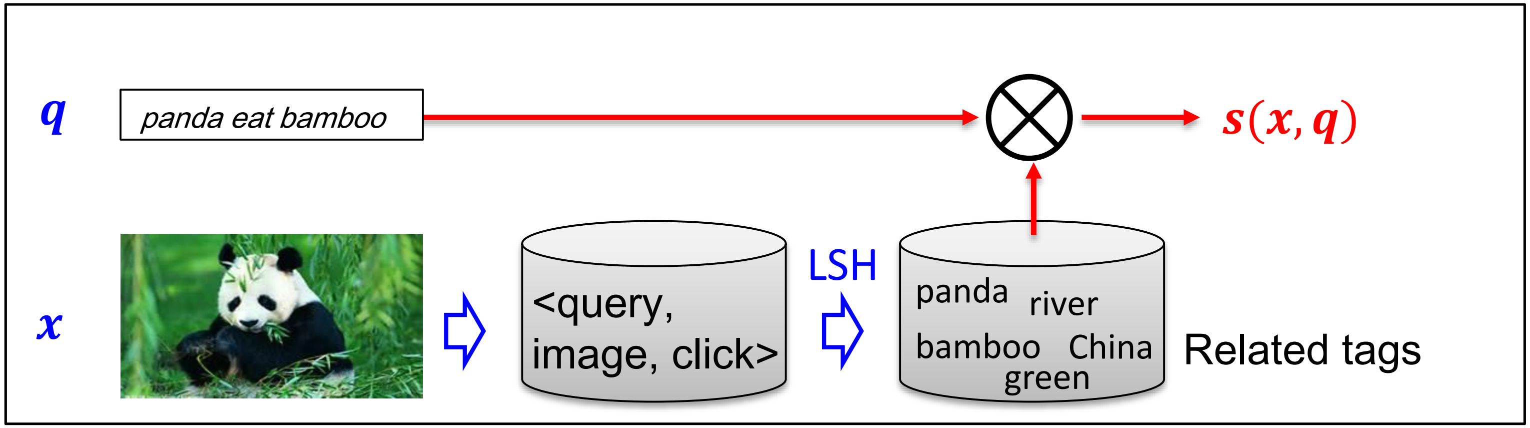
$$s(x,q) = \frac{1}{|X|} \sum_{x_k \in X} K_\sigma(x - x_k), K_\sigma(x - x_k)$$



#2 Image-based Scoring

- **Motivation**
  - Transfer to measure query-tag textual similarity.
- **Solution**
  - Retrieve the related tag set H by issuing the test query x into the training set via locality sensitive hashing (LSH);
  - Calculate query-image relevance by aggregating the textual similarities between test query q and the image-related tags.

$$s(x, q) = \sum_{(x_k, q_k) \in (X, Q)} e^{-l_k} R_k$$



#3 Concept Classification

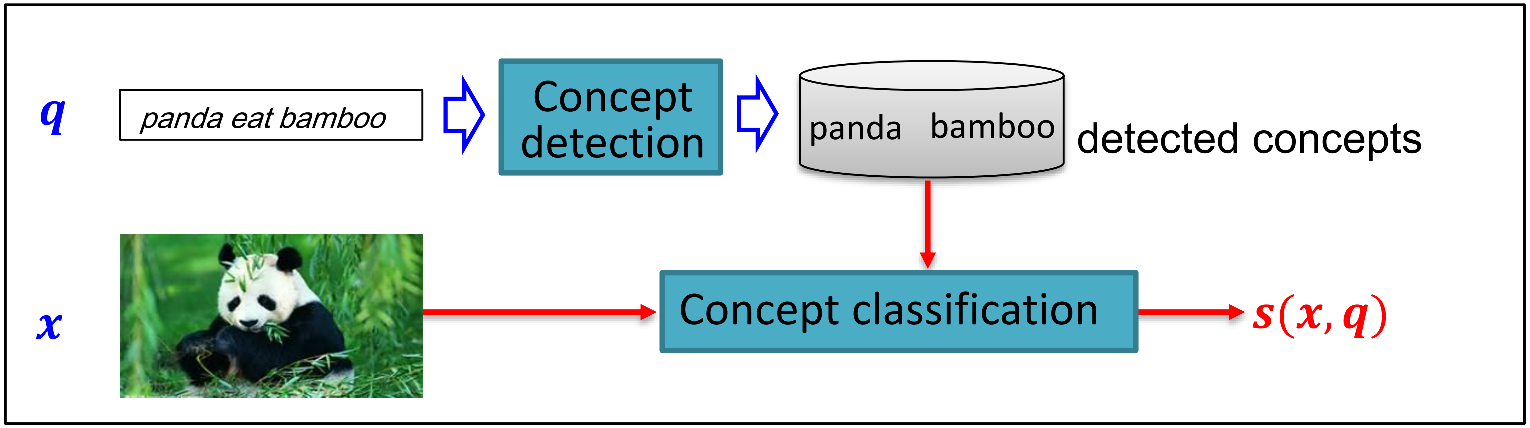
- Motivation
  - Transfer to an image classification problem;
  - Classification confidence as query-image relevance.
- Solution
  - Concept set
    - Concept refers to a salient term or phrase
    - Construct 249,527 concept vocabulary from training queries;
    - Using OpenNLP toolbox.

Table 2: The statistics of our extracted concepts

#Term	132,416	#Name	30,962
#Chunk	78,860	#Location	5,289
#Query	2,000		

#3 Concept Classification

- Solution
  - Concept classifier training
    - ➢ Large-margin classifier (SVM, boosting, etc.);
    - ➢ Positive v.s. Negative sample collection.
  - Test
    - ➢ Concept detection from test query;
    - ➢ Calculate the classification confidences of the test image to each detected concept;
    - ➢ Sum. or Ave. fusion of confidences as the final relevance score.



#4 Discriminative Ranking

- Motivation
  - Learn a discriminative model that both reserves ranked relationship in the training set and boosts ranking performance on new data.
- Solution
  - Based on model from [1].
  - Learn a mapping function f_θ from image space to text space:

$$s(x,q) = q \bullet f_\theta(x)$$

  - f is optimized towards minimizing the supervised loss for image-query ranking in the training set:

$$\min_\theta \sum_{i=1}^{N^+} \sum_{j=1}^{N^-} \max(0, 1 - s(x_i, q) + s(x_j, q)) + \frac{\lambda}{2} \|\theta\|^2$$

  - Generalization capability is guaranteed by SVM-alike formulation.

[1] David Grangier, Samy Bengio: **A Discriminative Kernel-Based Approach to Rank Images from Text Queries.** PAMI=30(8): 1371-1384 (2008)

Ensemble Model

- Ranking SVM-based ensemble
  - Ensemble on score level
  - Supervised learning to obtain optimal fusion weight on the development set.
- Ensemble schemes
  - Two Model Fusion
    - Concept Classification + Discriminative Ranking
  - All Model Fusion
    - Image-based Scoring
    - Query-based Scoring
    - Concept Classification
    - Discriminative Ranking

Evaluation Results

Table 4: Performance of the individual models and ensemble models on the test set

Model	Public Leaderboard DCG@25
Concept Classification	0.4937
Query-based Scoring	-
Image-based Scoring	-
Discriminative Ranking Model	0.4962
Two Models Fusion	0.5017
Ensemble of All Models	0.5033

Discussion

- What’s the most difficult part in this challenge?
  - Textual query complexity (noise, multiple words, etc.).
- What did you spend most of your time on?
  - Implement and compare between different models.
- How did you handle system scalability?
  - Model-based && preprocessing.
- What would you do if you do it again?
  - Explicitly analyze the word relations within test query.
- What would you do if the data size increases to 40 M?
  - Most of the examined models are expected to scale well.
- What else can we do with this dataset?
  - If extended by the user dimension, tasks of personalized image retrieval is enabled.

Q & A ?

Multimedia Computing group

http://nlpr-web.ia.ac.cn/mmc/

National Lab of Pattern Recognition

Institute of Automation, Chinese Academy of Sciences

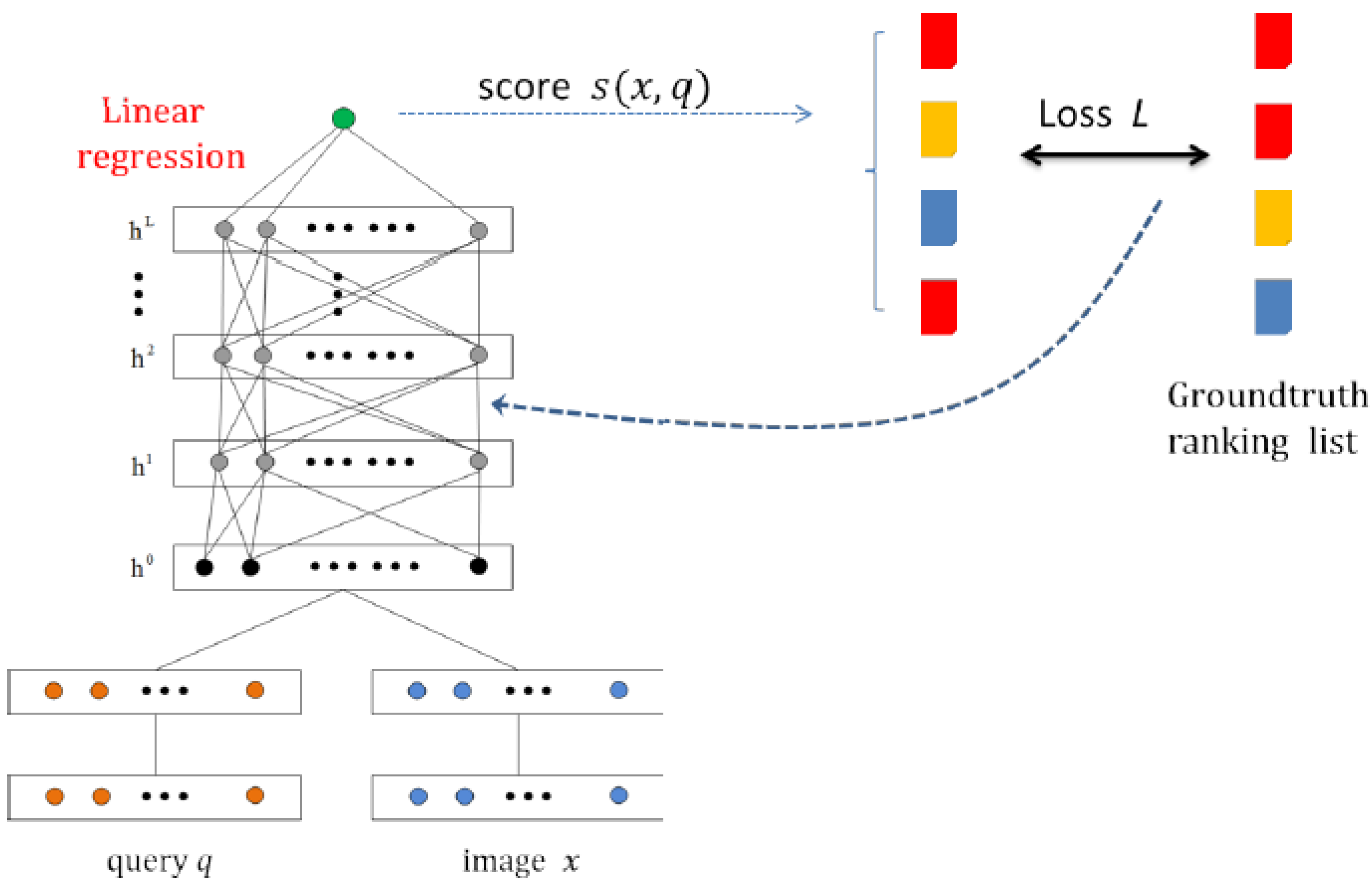
#5 Matrix Factorization-based Scoring

- Motivation
  - Assumption: similar images are relevant to similar queries;
  - Transfer to a recommendation problem.
- Solution
  - Analogous to collaborative filtering
    - Image-query relevance as the confidence of recommending the image to the query
  - Factorization Machine (FM [2]) model

$$s(x,q) = w_0 + \sum_{j=1}^{A} w_j \alpha_j + \sum_{j=1}^{A} \sum_{k=j+1}^{A} < p_j, p_k > \beta_j \beta_k$$

[2] Steffen Rendle: **Factorization Machines with libFM.** ACM TIST 3(3): 57 (2012)

#6 Multimodal Deep Learning



Results on Development Set

Model	Development set
Concept Classification	0.6955
Query-based Scoring	0.6759
Image-based Scoring S1	0.6794
Image-based Scoring S2	0.6815
Image-based Scoring S3	0.6802
Image-query-based Scoring	0.6785
Multimodal Deep Learning	0.6842
Matrix Factorization	0.6732
Discriminative Ranking Model	0.6976