MTAF: SHOPPING GUIDE MICRO-VIDEOS POPULARITY PREDICTION USING MULTIMODAL AND TEMPORAL ATTENTION FUSION APPROACH

Ningrui Ou, Li Yu*, Huiyuan Li, Qihan Du, Junyao Xiang, Wei Gong

School of Information, Renmin University of China

ouningrui@ruc.edu.cn

icassp 2022 Singapore

Outline

- Introduction
- Methodology
- Experiments
- Conclusion

What is Popularity?

Online media

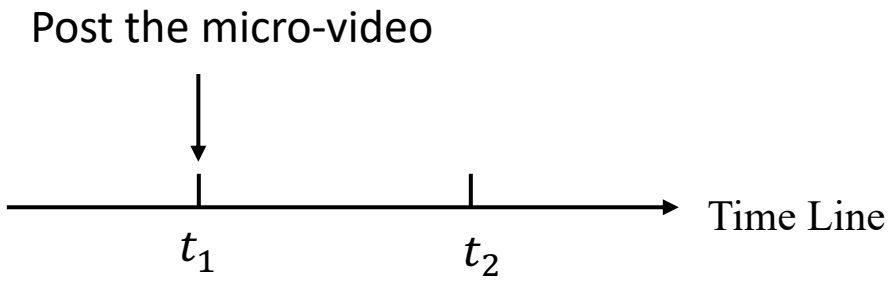
- Micro-blogging
- Music
- Pictures
- News
- E-commerce
- Micro-videos
- ……

Number of times

- Retweets
- Likes
- Views
- Comments
- Sales
- Downloads
- ……

Figure 1. Shopping guide micro-video comes with a product link (left) and KOL's individual store link to Taobao (right)

Existing Works



- Before posting the micro-video
  - Content-agnostic factors (Jia et al., WWW'17）
  - Content-based fusion (Jing et al., TKDE'18）
- After posting the micro-video
  - Popularity sequence (Vallet et al., CIKM'15）
  - Multimodal variational encoder-decoder (Xie et al., WWW'20)

Motivation

- Complex characteristics in shopping micro-videos
  - Rich information including anchor's voice emotion, product description, facial expression and social relationship.
  - Popularity trends that is affected by dramatic fluctuations in unexpected events.
- Develop a unified framework
  - Different viewers pay attention to different modality.
  - Applicable to various fields such as news, music, photo scenes.

Features Analysis

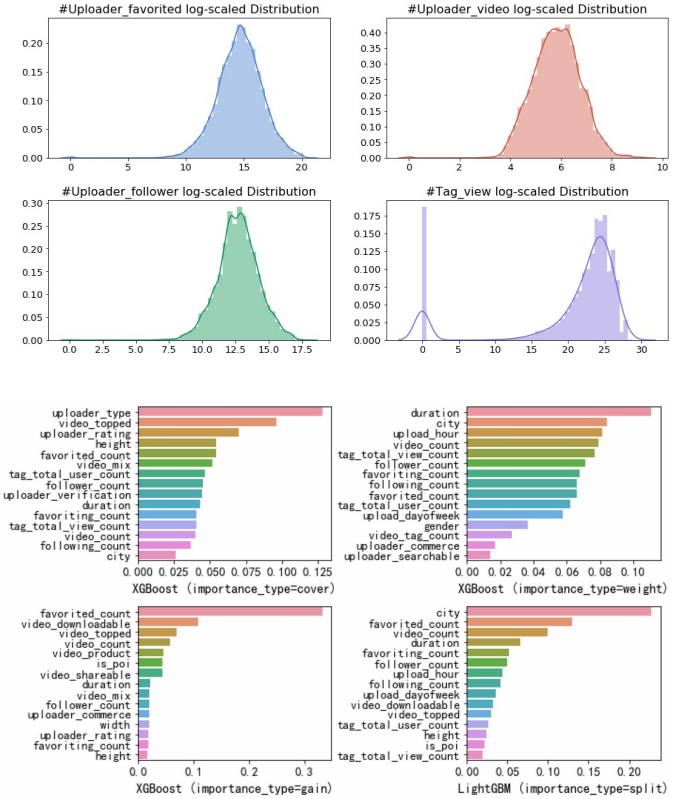
- Uploader
- Micro-video
- Hashtag
- Time
- Location

Table 1.The content-agnostic features in TikTok dataset.

Variable	Description
Uploader verification	Is the uploader authenticated by the platform
Uploader type	Type of the uploader, such anchor, star, etc.
Uploader rating	Rating of the uploader
Uploader commerce	Commerce level of the uploader
Uploader video count	Number of videos published by the uploader
Uploader follower	Number of followers of the uploader
Uploader following	Number of the uploader follows others
Uploader favorited	Number of likes for uploader's all videos
Uploader favoriting	Number of the uploader favorite other videos
Uploader searchable	Is the uploader searchable to others
Uploader visible	Is the uploader visible to others nearby
Uploader gender	Gender of the uploader
Uploader age	Age of the uploader
Video duration	Length of the video, in seconds
Video favorite count	Number of times the video was 'favorited'
Video comment count	Number of times the video was commented
Video share count	Number of times the video was shared
Video quality	Width and height of the video
Video topped	Is the video topped in the video list
Video downloadble	Is the video allowed to be downloaded
Video shareable	Is the video allowed to be shared
Video commentable	Is the video allowed to be commented
Video mix	Is the video a part of a collection
Video product	Is there a product link attached to the video
Video tag count	Number of the tags assigned to the video
Tag total view count	Number of times the tag was viewed
Tag total user count	Number of uploaders who used the tag
Upload dayofweek	What day of the week the video was uploaded
Upload hour	What hour of the day the video was uploaded
Upload city	City where video was uploaded
Upload POI	Is a point of interest in the location uploaded

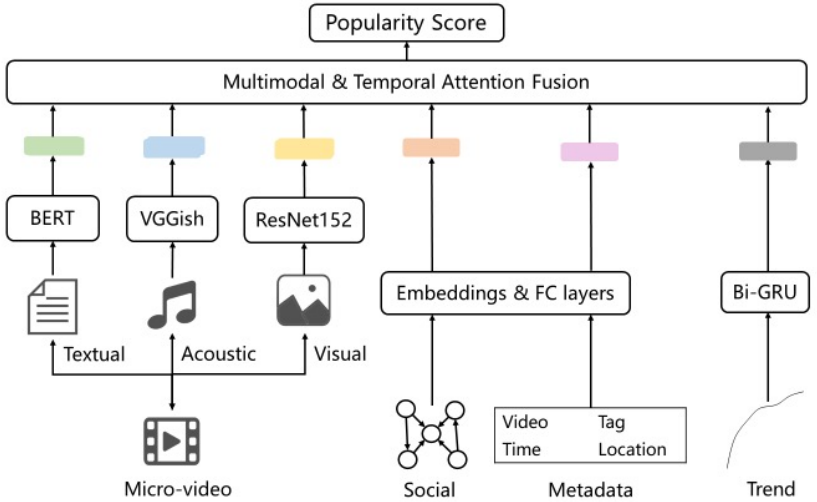
Features Analysis

- Most of the numerical features are approximately a logarithmic distribution.
- The hashtag itself has a certain popularity.
- The spatiotemporal information of the posted videos is crucial to the popularity.



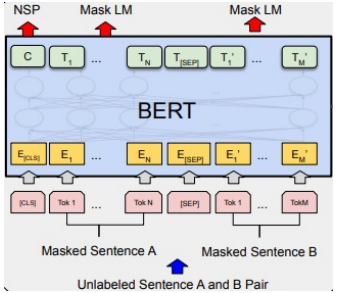
MTAF Framework

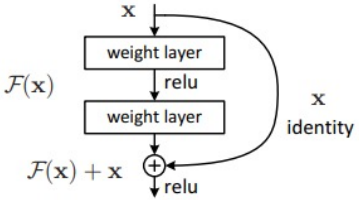
- Encoding Layer
  - Multi-modal Content Representations
  - Temporal Trend Representations
  - Content-agnostic Representations
- Attention Fusion Layer
- Prediction Layer



Methodology

- Encoding Layer
  - BERT to extract a deep semantic representation. (Devlin et al., arXiv'18)
  - VGGish to obtain a deep signal information. (Hershey et al., ICASSP'17)
  - ResNet152 to capture a deep visual features. (He et al., CVPR'16)
  - Bi-GRU to learn "rich-get-richer" phenomenon and dramatic fluctuations.
- Attention Fusion
- Popularity Prediction





$$\hat{y} = \sum_{j=1}^{N} P(x)_j * X \qquad \sum_{i=1}^{N} P(x)_j = 1$$

$$L(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Experiments

- Dataset

  7599 active users, 20445 shopping micro-videos, 122670 records

- Research Questions
  - Q1：What is the performance of proposed MTAF model?
  - Q2：Do multimodal content features have a significant impact?
  - Q3：How effective are early popularity trend in predicting task?

- Baselines

  SVR, LR, RFR(MM'18), XGBR(MM'19), Bi-GRU

- Evaluation Metrics

  MAE, MSE, Coefficient of Determination (R^2),

  Spearman Rank Correlation Coefficient (SRCC),

  Normalized Discounted Cumulative Gain (NDCG)

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$SRCC(R_1, R_2) = 1 - \frac{6 \times \sum_{i=1}^{N}(R_{1,i} - R_{2,i})^2}{N \times (N^2 - 1)}$$

$$DCG@K = \sum_{i=1}^{k} \frac{r_i}{\log_2(i+1)} \qquad NDCG@K = \frac{DCG@K}{iDCG@K}$$

Experiments

Table 2. Comparison of performance between our model MTAF and the several baseline methods.

Methods	MAE	MSE	R^2	SRCC	NDCG@10
SVR	1.518	3.631	0.130	0.363	0.106
LR	1.032	1.735	0.584	0.789	0.241
RFR	0.894	1.303	0.688	0.837	0.218
XGBR	0.877	1.275	0.695	0.841	0.134
MTAF-Att	0.864	1.249	0.701	0.849	**0.390**
MTAF	**0.822**	**1.139**	**0.727**	**0.860**	0.380

- The proposed model outperforms the advanced machine learning-based models.
- The attention-based approach is better than the methods that not use it.

Experiments

Table 3. Evaluation of robustness when some modality is missing in features fusion stage.

Methods	MAE	MSE	R^2	SRCC	NDCG@10
MTAF-N	1.363	2.978	0.286	0.538	0.078
MTAF-T	0.881	1.288	0.691	0.846	0.339
MTAF-A	0.870	1.259	0.698	0.849	0.380
MTAF-V	0.864	1.243	0.702	0.850	0.364
MTAF	**0.822**	**1.139**	**0.727**	**0.860**	**0.380**

- When content-agnostic features are missing, the performances of the model are worse dramatically.
- While textual, acoustic and visual modalities are in decreasing order of influence. Visual features are more significant compared to other modalities.

Experiments

Table 4. Evaluation of early popularity trend.

Methods	MAE	MSE	R^2	SRCC	NDCG@10
Bi-GRU	0.455	0.368	0.916	0.986	0.988
Bi-LSTM	0.439	0.389	0.911	0.986	0.988
MTAF	**0.083**	**0.019**	**0.995**	**0.991**	**0.992**

- Performances are substantially improved when combined with temporal features, which may be related to the fact that the popularity sequence is monotonically growing.
- Compared with only employ time-series models, a combination of multimodal and sequence representations can be effectively enhanced.

Conclusions

- Main contributions
  - A unified framework to efficiently represent and fuse multimodal content.
  - We explore the important factors that influence the popularity of shopping micro-videos.
  - The model is easy to extend and deploy.
- Future works
  - Consider knowledge graph to enrich representations.
  - Cross-modal perception

Main References

Jia, Adele Lu, et al., "An analysis on a YouTube-like UGC site with enhanced social features," *in International Conference on World Wide Web Companion*, pp. 1477-1483, 2017.

P. Jing, Y. Su, et al., "Low-Rank Multi-View Embedding Learning for Micro-Video Popularity Prediction," *in IEEE Transactions on Knowledge and Data Engineering*, pp. 1519-1532, 2018.

D. Vallet, S. Berkovsky, et al., "Characterizing and Predicting Viral-and-Popular Video Content," *in International Conference on Information and Knowledge Management*, pp. 1591-1600, 2015.

J. Xie et al., "A Multimodal Variational Encoder-Decoder Framework for Micro-video Popularity Prediction," *in International Conference on World Wide Web*, pp. 2542-2548, 2020.

K. Xu, et al., "Multimodal Deep Learning for Social Media Popularity Prediction with Attention Mechanism," *in ACM International Conference on Multimedia*, pp. 4580-4584, 2020.

Thanks

13 May, 2022