UDC: 004.93 https://doi.org/10.15407/jai2024.04.041

S. Yakovlev[1], N. Shapoval[2]
[1,2]National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
37 Beresteysky Avenue, Kyiv, 03056
[1]se2001ya@gmail.com
[2]shovgun@gmail.com
[2]https://orcid.org/0000-0002-8509-6886

RECOMMENDATION OF HASHTAGS USING DEEP LEARNING METHODS BASED ON MULTIMODAL DATA

Abstract. Generating image text captions is an important task and aims to automatically generate a text description for an image. Recommendation of hashtags is a practical option for this task. Hashtags contribute to increasing the relevance of content for the audience and ensure better visibility of publications. The problem of choosing optimal hashtags becomes especially relevant for social platforms, where users generate huge amounts of content with different types of modalities — images, text captions, videos, etc. There are a number of challenges that need to be addressed when solving this problem. First, text captions for posts are often short or even absent. Secondly, multimodal algorithms often do not take into account the previous activity of the user, which can significantly limit the quality of recommendations. Third, the balance between the importance of textual and visual cues may vary depending on the nature of the publication. The purpose of this study is to develop a modified feature fusion algorithm for the task of multimodal hashtag recommendation, which is able to take into account the context of the user's previous history, adaptively evaluate the importance of textual and visual features, and improve the quality of recommendations in cases of the absence or weakness of textual description. As part of the study, a model was modified that, in addition to analyzing the image and text caption, takes into account part of the previous history of user interactions. The main contribution is a new feature fusion module that weights their importance depending on the context. This approach allows to improve the relevance of recommendations in situations where the textual modality is not informative enough, which is a common problem in real data. The experimental results confirmed that the proposed feature fusion module provided more accurate hashtag recommendations, especially for cases with short or missing text captions.

Keywords: neural networks, deep learning, multimodal data, hashtags, social networks.

Introduction

Today, when social networks such as X, Facebook, Instagram, etc. are gaining more and more popularity, a huge volume of information appears due to publications, which can be quite difficult for an ordinary person to understand. Therefore, so-called "hashtags" are available for use in the absolute majority of social networks. A hashtag is a keyword, or a set of words, other symbols or tokens, the content of which very succinctly reflects the main topic of the message. Their main purpose is to simplify the search or filtering of content according to the relevant topic or content. Therefore, automatic creation of hashtags is an important area of research in which companies and people are interested. Solving the task of recommending relevant hashtags in social networks allows you to automate the process of categorizing content in social networks. Which in turn will be useful for marketing or social research.

The recommendation can also be based on one type of publication data, such as only a caption, only an image, only a video, etc., or simultaneously on several types of data that characterize the publication, which are called multimodal data. This need is explained by the fact that multimodal data allow in combination to get a deeper understanding of the content of the publication.

As an example, consider a publication with the image shown in Fig. 1.

Fig. 1. Post image

This publication has the following caption: "You never know when the wind will change direction! Who is ready to go on a new adventure?". At the same time, the user assigned the following set of hashtags: #surfing, #sunset, #adventure. In this case, the hashtags #surfing and #sunset are defined only by the content of the image, while the hashtag #adventure is defined mostly by the content of the text. The image and the caption can display different information, but in combination can display a more detailed and complete content of the publication. That is why the task of analyzing multimodal data is so important.

This work is devoted to the development and modification of deep learning models for hashtag recommendation based on multimodal social network data, using the Instagram social network as an example.

Related work

When using deep learning models, the problem of "K-hashtag recommendation" is usually considered, in which predictions obtained by a neural network are ranked. Then, instead of simply choosing all the predicted hashtags, K of the most relevant ones, i.e. those with a degree of confidence above a certain threshold, are selected as a final recommendation.

It is also important to highlight two main approaches to solving the problem of generating hashtags:

– Multi-label classification, which assumes that several hashtags can be assigned to each object (for example, publications). Unlike multi-class classification, where each instance belongs to only one class, here the classes (hashtags) are not mutually exclusive. For example, an image with a dog in nature can receive the hashtags #dog, #nature, and #pet at the same time;

– Sequence generation is focused on creating an ordered sequence of hashtags, taking into account their relationship and context. For example, the hashtags #travel, #beach, #sunset have a logical sequence that reflects the content of the publication. This approach allows the model to take into account dependencies between hashtags, which can be especially useful in cases where the structure or order of hashtags is important for a better understanding of content.

Thus, in the paper [1] Attention-Based Multimodal Neural Network Model for Hashtag Recommendation (AMNN), the task of hashtag recommendation is considered as a sequence generation task, and the encoder-decoder architecture is used. First, the features from the image and caption are extracted separately, then concatenated, which helps to obtain the final representation of the post, after which a sequence of hashtags is generated using the GRU recurrent network.

Another option is the Co-Attention (CoA) model [2], where hashtag recommendation is treated as a multi-label classification task. First, a text-based image attention mechanism is applied, and then the resulting representation is used in an image-based text attention mechanism, after which the obtained feature set is already fed to the output layer of the network for the final prediction.

A more interesting option is the Memory Augmented Co-attention Model (MACoN) [3], which uses the parallel co-attention mechanism [4] for the image and caption at the same time in order to take into account the influence of each of them on each other. And it additionally uses the previous history of the user regarding the use of hashtags, thereby adding an element of personalization to the recommendations.

A rather specific solution to this problem is the Triplet-Attention Graph Network (TAGNet) [5], where the so-called "visual similarity graph" is built based on the assumption that similar images will have the same hashtags. The features of each node are the features of the current caption combined with the influence vector of the corresponding user based on the history of his previous publications, after which the Triplet-Attention module is used, which calculates the influence of the image, text, and user data on each other.

The multimoDal pErSonalIzed hashtaG recommeNdation (DESIGN) model [6] combines the key features of almost all the models mentioned above. In addition to the text caption and image of the corresponding

publication, it also uses a certain number of previous user posts to add an element of personalization. DESIGN also combines the above two approaches to solving the given problem - multi-label classification and sequence generation - to aggregate results and provide more relevant results. However, the given model has the following drawback: when forming the final vector of features of the content of the publication, the features of both modalities are equally taken into account, although they are not always equally informative.

A kind of revolution in the task of recommending hashtags, as well as in most other, if not all tasks of natural language processing, was made by GPT models (Generative Pre-trained Transformer), which can be used in particular for multimodal recommendation of hashtags. But for this task, it has a significant drawback, which is that it can make too uniform recommendations, such as generating hashtags that are generally very popular but already widely used by others, which can reduce the uniqueness of the user's content and limit its visibility among competitors' publications, which can be critical, for example, when distributing advertising in social networks.

Therefore, it was decided to take the DESIGN model as a model for further research.

Statement of the problem

It is necessary to create a model that will be able to offer a set of K relevant hashtags for the corresponding publication based on the textual, visual composition of the publication and a certain number of previous publications of the user in the Instagram social network. The resulting hashtags should be, on the one hand, sufficiently comprehensive, and on the other hand, not redundant. Then analyze the results and draw conclusions about which value of K is generally optimal.

Generation of hashtags

First of all, let's briefly describe the methodology of the basic DESIGN model. As mentioned earlier, DESIGN is a model that combines the features of all other models to solve the given problem. Its general architecture is shown in Fig. 2:

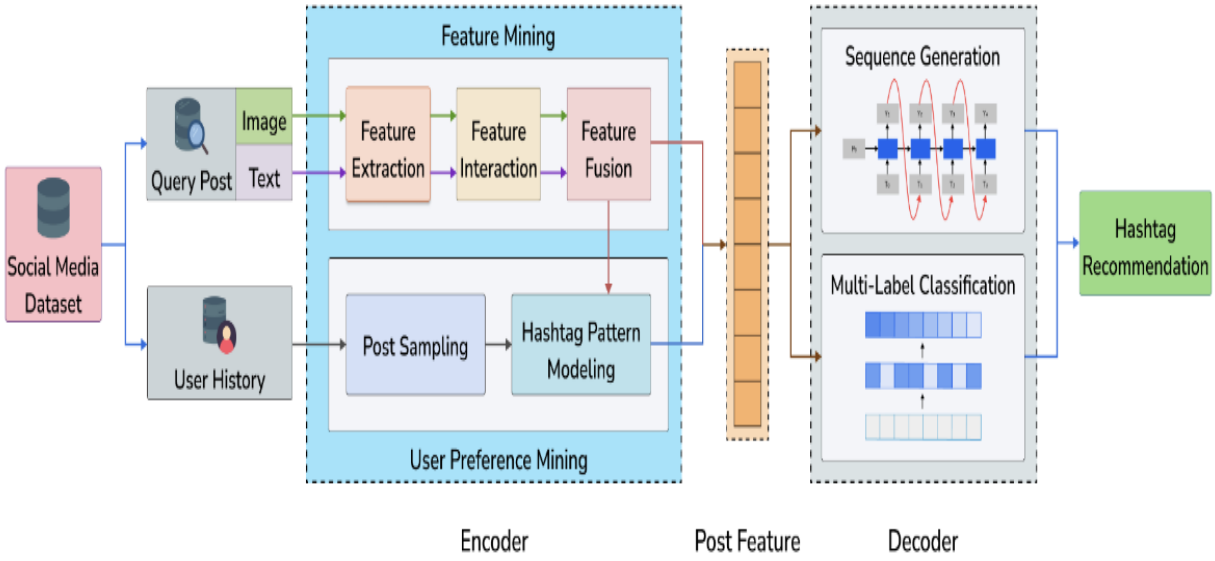

Fig. 2. General architecture of the DESIGN model

An important component is the feature extraction module from the content of the publication (Feature Mining), which receives an image and a caption as input, and returns a continuous vector of features, taking into account the information of both modalities.

This module consists of: 1. feature extraction block from each modality separately

(Feature Extraction); 2. interaction analysis block (Feature Interaction); 3. Feature Fusion block. The block of interaction analysis with the help of mechanisms of word-level attention and parallel joint attention takes into account the influence of the textual and visual components of the publication on each other. The feature fusion block based on vectors of textual and visual features forms the final vector of features of the publication's content.

Suppose that a vector of text features is supplied to the input of the feature fusion module t and a vector of visual features v. Then the feature fusion module simply sums the vectors of visual and textual features obtained in the previous step. That is, the vector of features of the content of the publication is equal $p = t + v$. In this case, the textual and visual components are mostly taken into account equally.

Within the framework of this study, the use of a modified feature fusion module is proposed, which will weigh the importance of each publication modality. This modified merge module works according to the following algorithm:

1. $f = concatenate(t, v)$;
2. $\hat{f} = BatchNorm(f)$;
3. $o = LeakyRelu(Dense(\hat{f}))$;
4. $\hat{o} = BatchNorm(o)$;
5. $w = Sigmoid(Dense(\hat{o}))$;
6. $p = (1 - w) \cdot t + w \cdot v$.

Here $f -$ concatenated vector of features of both modalities, $\hat{f} -$ the vector obtained after applying batch normalization to the vector f, $o -$ intermediate vector after vector submission $\hat{f}$ on a fully connected layer, $\hat{o} -$ vector o after applying normalization, $w -$ weights vector.

We get that the vector of features of the content of the publication is calculated not simply as the sum of the vectors of textual and visual features, but as their coordinate-weighted average, which allows us to additionally take into account the importance of the features of each of the modalities for influencing the content of the publication.

Also, in the future, to extract text features, as in the original article, we will use the multilingual BERT model [7]. Regarding the extraction of visual features, two convolutional neural network models – VGG-16 [8] and ResNet50 [9] – were studied in the original work, and it was found that the VGG-16 model is the better solution out of the two, so we will take it, and we will additionally take the basic model of the vision transformer (ViT) [10].

Metrics for evaluating the quality of models

Evaluating such models is quite a difficult task, due to the factor of subjectivity, when each user assigns hashtags to a post subjectively, guided by certain beliefs and motives, due to which the number of hashtags to a post can vary significantly, while a fixed number is usually recommended number.

However, there are several standard solutions for quality assessment metrics, mostly as a recommendation task.

To define them, we will introduce such notations as Gh - the set of ground truth assigned hashtags for this specific publication, Rh - set of recommended hashtags for the given publication, and Ch - a set of common hashtags, defined as the intersection of the two previous sets.

The first metric is a standard solution for the recommendation problem, which is called Hit rate and is determined by formula 1:

$$Hit\ rate = min\{1, |Ch|\}, \qquad (1)$$

where $|\cdot| -$ the number of elements in the corresponding set.

For each specific post, the Hit rate is an indicator that at least one hashtag was recommended correctly, while the value over the entire training sample is calculated by averaging over each post, and will give the proportion of posts for which at least one hashtag was correctly selected.

The second metric is precision, which is a generalization of the appropriate metric for the binary classification problem and is defined as:

$$Precision = |Ch|/|Rh| \quad (2)$$

For each individual post, this is a measure of the extent to which the recommendations consist of valid hashtags.

An alternative to accuracy is completeness, defined by formula 3:

$$Recall = |Ch|/|Gh| \quad (3)$$

In contrast to precision, completeness is a measure of how well valid hashtags are covered by recommendations.

To find a compromise between accuracy and completeness, the F1 measure is used, which is essentially a harmonic mean between the values of the corresponding metrics, and is determined by formula 4:

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (4)$$

Data set preparation

For experiments, a new data set was collected, which in the final version includes 41873 publications for 5900 users, with an average number of publications per user – 7.09. Low-frequency hashtags were removed, resulting in a total of 1,734 unique hashtags and an average number of hashtags per post of 8.28. In addition, we note that in this case, the publications reflect the specifics of the Instagram social network, where the text caption can be very long, quite short, or absent at all.

Experiment

Let's proceed directly to conducting experiments. To do this, we will consider model configurations depending on which network is used to extract visual features, namely, consider the VGG-16 model as the best solution in this aspect from the original article, and additionally consider the basic model of the visual transformer, as well as depending on of the feature fusion module used, where as alternatives we will consider the basic one, which includes a simple summation of feature vectors, and the proposed modification, which additionally weighs the importance of the features of each of modalities.

Graphs of precision and recall values depending on the number of recommended hashtags K for all models can be seen in Fig. 3, while graphs of Hit rate and Fig. 4:

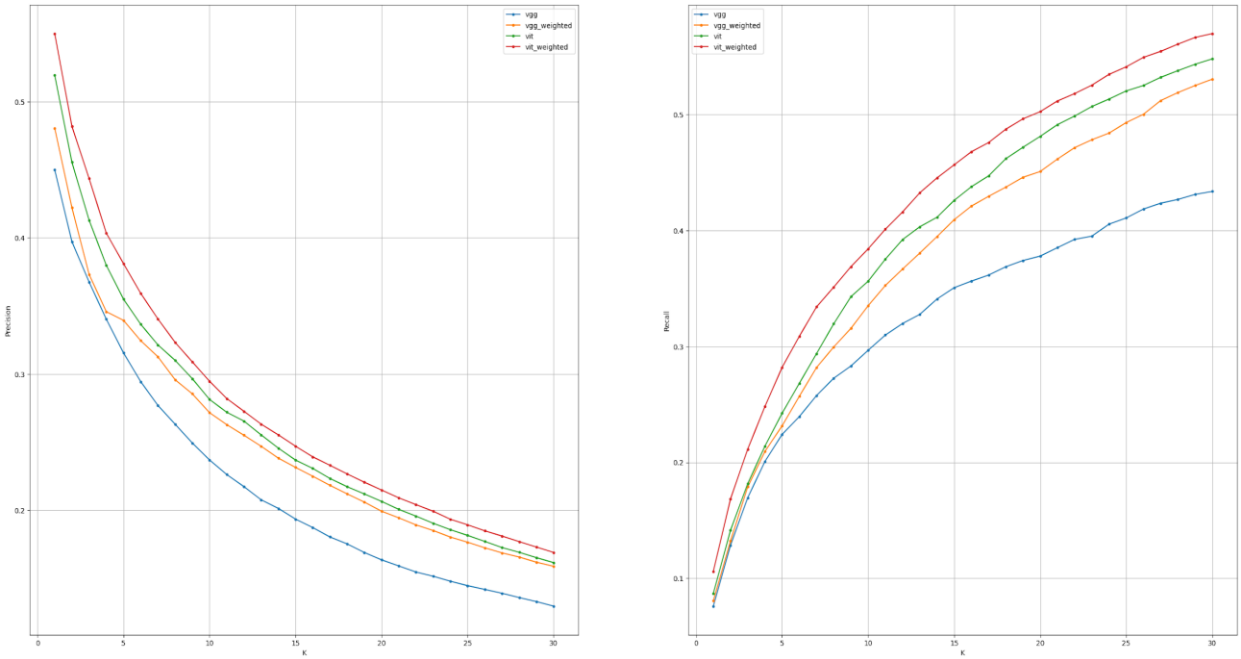

Fig. 3. Results for Precision and Recall

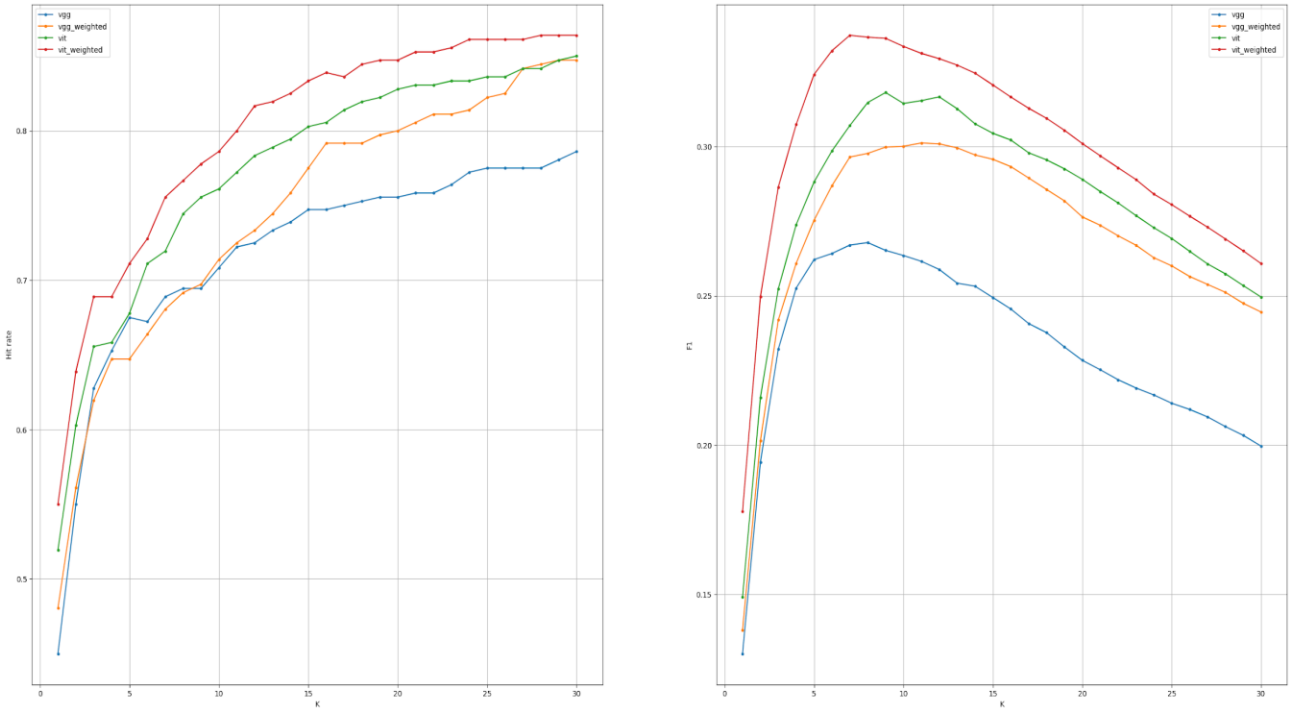

Fig. 4. Results for Hit rate and F1

We see that according to the values of all metrics, the best solutions for each value of K are the model using the visual transformer for the extraction of visual features and the modified feature fusion module.

The values of the metrics for the case of K=7, in which the highest value of F1 is achieved, are given in Table. 1:

Table 1. Results for K = 7

Model configuration	Hit rate	Precision	Recall	F1
VGG	0.689	0.277	0.258	0.267
VGG+Weighted	0.681	0.313	0.282	0.296
ViT	0.719	0.321	0.294	0.307
ViT+Weighted	0.755	0.34	0.334	0.337

Demonstration of results

To demonstrate the operation of the model, let's give an example of its operation for one publication from the test sample, which is shown in Fig. 5:

Fig. 5. An example of a test publication

The sets of recommended hashtags for each of the models are given in Table. 2.

Table 2. Results of the recommendation

Model configuration	Recommended hashtags
VGG	#fitnessmotivation, #fitness, #gay, #fit, #food, #bodybuilding, #fitfam
VGG+Weighted	#fitnessmotivation, #fitnessaddict, #fit, #fitness, #food, #inspiration, #sport
ViT	#fitnessmotivation, #fitness, #selfie, #fitnessaddict, #running, #motivation, #sport
ViT+Weighted	#fitnessmotivation, #fitness, #fitnessaddict, #fit, #run, #training, #fitfam

Note that hashtags that were actually assigned to publications and were recommended by the model are marked in green, hashtags that were not actually assigned to publications, but were recommended by the model and are relevant, are marked in blue and hashtags that were recommended by the model but not actually intended for publication and are not relevant to its content are marked in red.

Conclusions

In summary, we can say that the use of the proposed modification of the feature fusion module in DESIGN model showed an improvement in the results of key metrics with both models of visual feature extraction. Especially promising results were achieved when using the proposed feature fusion module in combination with the vision transformer,

which indicates the possibility of using the given model configuration in the application for personalized recommendation of hashtags in the Instagram social network.

References

1. Q. Yang, G. Wu, Y. Li, R. Li, X. Gu, H. Deng, and J. Wu, "AMNN Attention-based multimodal neural network model for hashtag recommendation," IEEE Transactions on Computational Social Systems, vol. 7, no. 3, pp. 768–779, 2020.
DOI: https://dx.doi.org/10.1109/TCSS.2020.2986778

2. Q. Zhang, J. Wang, H. Huang, X. Huang, and Y. Gong, "Hashtag recommendation for multimodal microblog using co-attention network." in IJCAI, 2017, pp. 3420–3426.
DOI: http://dx.doi.org/10.24963/ijcai.2017/478

3. S. Zhang, Y. Yao, F. Xu, H. Tong, X. Yan, and J. Lu, "Hashtag recommendation for photo sharing services," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, 2019, pp. 5805–5812.
DOI: http://dx.doi.org/10.1609/aaai.v33i01.33015805

4. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical Question-Image Co-Attention for Visual Question Answering. 2016. arXiv:1606.00061.
DOI: http://dx.doi.org/10.48550/arXiv.1606.00061

5. Y. -C. Chen, K. -T. Lai, D. Liu, and M. -S. Chen, "Tagnet: Triplet-attention graph networks for hashtag recommendation," IEEE Transactions on Circuits and Systems for Video Technology, 2021.
DOI: https://dx.doi.org/10.1109/TCSVT.2021.3074599

6. Bansal, Shubhi & Gowda, Kushaan & Kumar, Nagendra. (2022). A Hybrid Deep Neural Network for Multimodal Personalized Hashtag Recommendation. IEEE Transactions on Computational Social Systems. DOI: http://dx.doi.org/10.1109/TCSS.2022.3184307.

7. Ashish Vaswani. et al. Attention Is All You Need. 2017. arXiv:1706.03762.
DOI: https://doi.org/10.48550/arXiv.1706.03762.

8. Karen Simonian, Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. 2014.
DOI: https://doi.org/10.48550/arXiv.1409.1556.

9. Kaiming He, Xianghu Zhang, Shaoqing Ren, Jian Sun. Deep Residual Learning for Image Recognition. 2015. DOI: https://doi.org/10.48550/arXiv.1512.03385

10. Alexey Dosovitskiy et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. arXiv:2010.11929v2.
DOI: https://doi.org/10.48550/arXiv.2010.11929

The article has been sent to the editors 29.11.24.
After processing 10.12.24.
Submitted for printing 30.12.24.