
SocialML: machine learning for social media video creators

Tomasz Trzcinski^{a,b}, Adam Bielski^b, Pawel Cyrta^b and Matthew Zak^b

^aWarsaw University of Technology ^bTooploox
firstname.lastname@tooploox.com

Abstract

In the recent years, social media have become one of the main places where creative content is being published and consumed by billions of users. Contrary to traditional media, social media allow the publishers to receive almost instantaneous feedback regarding their creative work at an unprecedented scale. This is a perfect use case for machine learning methods that can use these massive amounts of data to provide content creators with inspirational ideas and constructive criticism of their work.

In this work, we present a comprehensive overview of machine learning-empowered tools we developed for video creators at Group Nine Media - one of the major social media companies that creates short-form videos with over three billion views per month. Our main contribution is a set of tools that allow the creators to leverage massive amounts of data to improve their creation process, evaluate their videos before the publication and improve content quality. These applications include an interactive conversational bot that allows access to material archives, a Web-based application for automatic selection of optimal video thumbnail, as well as deep learning methods for optimizing headline and predicting video popularity. Our A/B tests show that deployment of our tools leads to significant increase of average video view count by 12.9%. Our additional contribution is a set of considerations collected during the deployment of those tools that can help to understand the challenges of applying machine learning methods in creative practice.

Tools

Assisted thumbnail selection. A video thumbnail is often the first information about the video a social media user can see. It is therefore critical to choose it wisely to maximize the chances of catching viewer's attention. Typically, the selection of the thumbnail is done based on publisher's intuition and experience. Using the consumption data of social media videos from the past, we built a machine learning algorithm to select the thumbnail for social media video that is likely to attract users' attention and recommend it to the publisher.

For the purpose of our training, we collected a dataset of 37'042 Facebook videos along with their thumbnails and titles from top publishers according to TubularLabs.com ranking. To take into account the influence of channel popularity on individual video popularity, we normalized number of views of each video by dividing it by the number of likes of a channel that the video was posted on. We then split the dataset into popular and unpopular class using median normalized view count. We trained a model for binary classification by fine-tuning the last layer of ResNet50 [1] model pre-trained on ImageNet dataset [2]. We used 80% of data for training, 10% for validation and 10% for testing. We reached 66% classification accuracy on the test dataset. In production, we apply the model to 40 frames extracted uniformly from a video. The frame evaluated with the highest score is recommended as a thumbnail. A screenshot of the application can be seen in Fig. 1.

Slack chatbot for researching archived materials. Quick and intuitive access to materials is a prerequisite for effective work of any video creator. In order to facilitate it, we implemented a responsive chatbot that returns lists of videos tagged with topics related to user's query. Materials are indexed using meta-data extracted from videos using image and text processing. More precisely, we use TextBoxes [3] to detect subtitles and other relevant text snippets in the videos, and then we recognize the text using CRNN [4]. Then we input video headline and text extracted from the video

into a neural network topic classifier consisting of dense layer, ReLU and a softmax layer for two tier category classes. As text embedding, we use 400-dimensional FastText vector representation [5]. Similar tagging is applied to audio transcription generated using speakers diarization segmentation [6] and speech recognition models.

After interviewing a sample of our users, we decided to deploy the service using Slack. User can ask our Slack bot several questions related to retrieving indexed content, e.g. related to user interaction statistics of videos with a given tag. We used standard decision tree rule-based and slot matching methods for the natural language understanding component of our bot. A screenshot of a sample conversation with Slack bot can be seen in Fig. 2 and Fig. 3.

Headline optimization. Video titles are short video descriptions displayed above the content and written to attract users' attention. To improve the effectiveness of those headlines, we trained a deep learning recurrent model that scores proposed title through indirect modeling of users' preferences. We used the dataset of Facebook videos described above along with their normalized view counts. Inspired by [7], we implemented a bi-directional LSTM network with attention [8] and trained it with video titles transformed with pre-trained GloVe embeddings [9]. The objective of our network was popularity classification and the model achieved 70% accuracy on the test dataset. Thanks to the attention mechanism we can visualize how specific words contribute to popularity of the video and guide publisher's creative process of improving it. We deployed the tool in production through the Slack bot described above. A screenshot of this functionality can be seen in Fig. 4.

Video popularity prediction based on its frames. Incrementing a video view count after first few seconds and setting Facebook videos to auto-play by default leads to increased importance of the first few seconds of the video published on this platform. We therefore leverage past consumption data to help publishers improve the quality of the opening scene and hence lead to higher video popularity and viewer retention. To that end, we implemented a deep neural network architecture that analyses first frames of a video and predicts its future popularity.

To train our model, we used the dataset of Facebook videos described above and extracted 18 evenly distributed frames from the first 6 seconds of a video. For each frame, we extract features using penultimate layer of the ResNet50 [1] model pre-trained on ImageNet [2] dataset. We follow [10] and train a neural network model with attention mechanism to classify video as popular or not. The resulting model achieved over 68% classification accuracy on the test dataset. Inspired by [11], we wanted to visualize what parts of the image contribute to popularity score. Hence, we used GradCAM [12] to create heatmaps overlaid on the original frames. This way we can identify parts of frames that contribute the most to video popularity, as in Fig. 5. We deployed the tool as a Web service. Fig. 6 shows a screenshot of a working application along with a popularity heatmap.

Applying machine learning in creative practices

When developing our tools, we faced several difficulties related to socio-cultural aspects of using machine learning in creative practice, which can be summarized in the following questions:

How to deploy machine learning algorithms to assess content quality without the risk of offending its creators? Journalists often consider creativity as their most valuable talent and they are reluctant to allow machines to quantify it. We therefore introduced our tools for popularity prediction as an automatic system to alert the editor if their video was expected to be particularly unpopular. In practice, the deployed implementation analyses every video uploaded through a Web user interface and if the normalized popularity score falls below 20-th percentile score, the user is alerted through a pop up window. This significantly increased the acceptance rate of the tool within our user base.

How to validate the impact of a machine learning system in creative practice? To better understand the impact of our toolkit on the work of video creators, we ran A/B tests on user groups working with and without our tools. We then analyzed content popularity metrics, such as video views, shares and comments, for the videos created by both groups. Although using these metrics as content quality estimators has several shortcomings (such as sensitiveness to clickbaity news titles), they remain main evaluation metrics used by creators themselves and, hence, we adopted them too.

How to avoid the situation where all the creative content converges on one topic that is intrinsically linked to higher popularity? This bias of a popularity estimation algorithm, related closely to the bias present in the training dataset, also exists in other machine learning applications [13]. The solution we proposed to avoid it in our application is to categorize the content and normalize its popularity scores according to the median category value. By using this additional normalization step, our method is more likely to concentrate on the features related to content quality instead of its topic.

Acknowledgment

The authors would like to thank Group Nine Media Inc. for enabling this research. The work was partially supported as RENOIR Project by the European Union Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 691152 (project RENOIR) and by Ministry of Science and Higher Education (Poland), grant No. 34/H2020/2016.

References

- [1] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.
- [3] M. Liao, B. Shi, X. Bai, X. Wang, and W. Liu, “Textboxes: A fast text detector with a single deep neural network,” in *AAAI*, 2017.
- [4] B. Shi, X. Bai, and C. Yao, “An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition,” *CoRR*, vol. abs/1507.05717, 2015.
- [5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *arXiv preprint arXiv:1607.04606*, 2016.
- [6] P. Cyrta, T. Trzcinski, and W. Stokowiec, “Speaker diarization using deep recurrent convolutional neural networks for speaker embeddings,” in *International Conference on Information Systems Architecture and Technology*, pp. 107–117, Springer, 2017.
- [7] W. Stokowiec, T. Trzcinski, K. Wolk, K. Marasek, and P. Rokita, “Shallow reading with deep learning: Predicting popularity of online content using only its title,” *CoRR*, vol. abs/1707.06806, 2017.
- [8] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *CoRR*, vol. abs/1409.0473, 2014.
- [9] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.
- [10] T. Trzcinski, P. Andruszkiewicz, T. Bochenski, and P. Rokita, “Recurrent neural networks for online video popularity prediction,” *CoRR*, vol. abs/1707.06807, 2017.
- [11] Z. Bylinskii, N. W. Kim, P. O’Donovan, S. Alsheikh, S. Madan, H. Pfister, F. Durand, B. C. Russell, and A. Hertzmann, “Learning visual importance for graphic designs and data visualizations,” *CoRR*, vol. abs/1708.02660, 2017.
- [12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: Visual explanations from deep networks via gradient-based localization,” in *The IEEE International Conference on Computer Vision (ICCV)*, 2017.
- [13] T. Tommasi, N. Patricia, B. Caputo, and T. Tuytelaars, “A deeper look at dataset bias,” *CoRR*, vol. abs/1505.01257, 2015.

Figures

Thumbnail selection assistance

Upload your video

Select Upload

Predicted popularity score:

6.44

0 - the least popular, 10 - the most popular



Scenes with corresponding popularity








Scene	Predicted popularity
	5.07
	5.62
	5.38
	4.55
	4.94
	5.19
	6.44

Figure 1: Web application for thumbnail selection assistance. Subsets of analyzed frames with corresponding scores are shown. Our model uses fine-tuned convolutional neural network to analyse a set of uniformly sampled frames from a video and outputs a corresponding popularity score. Recommended thumbnail is selected as the frame with the highest score returned.


Diagram labels for the query: Source (points to the video title), Data Subject (points to the video title), Tags (points to the hashtag), and Time (points to the date range).

John 6:18PM
/wiz list of ***NowThis*** \$videos\$ tagged #politics, media, president# posted [May 29 - June 28]

WizardBot 6:18PM
Hi John, I found 3 videos, metrics below

Video Title	Views	Reactions	Comments	Reach
Media and Politics in the Age of Trump	983,384	124,443	1023	1,394,293
Trump just issued a direct threat to the media	983,384	124,443	1023	1,394,293
Nearly half of voters think media concocts Trump stories	983,384	124,443	1023	1,394,293


Figure 2: Example of a query question posed at the Slack bot for indexing archive materials. As a result, the bot returns social media interaction statistics related to all videos with similar tags. The tagging is done offline using textual input from both headline and text displayed in the video frames.

Paweł Cyrta  < 1 minute ago
@wizardbot Please, Give me videos tagged cats

Wizardbot App APP < 1 minute ago
Found 1568 videos 📊
Total views: 3778999451
Total video impressions: 20156904405
Top 5 videos:

Link	Views
https://facebook.com/thedodosite/videos/1281395508661788/	64685407
https://facebook.com/thedodosite/videos/1281395508661788/	64596858
https://facebook.com/thedodosite/videos/1281395508661788/	64459339
https://facebook.com/thedodosite/videos/1281395508661788/	64321377
https://facebook.com/thedodosite/videos/1281395508661788/	64187805

Did it help?

Paweł Cyrta  6 minutes ago
@wizardbot Show me now a brand-videos for Pepsi on Dodo channel.

Wizardbot App APP < 1 minute ago
Found 5 videos for brand **pepsi** on **The Dodo**
with at least one of tags: **Beverages**, **Food**, **Interest**, **Manufacturing**, **Marketing**, **New York**, **PepsiCo**, **Snack food**, **Snacking**.
5 most popular are below:

This cat has dinner with his mom every single night – and loves posing with their meals 🍽️

Post video views	Post impressions
7,510,500	40,446,754

Content tags
Animal Lovers, I Love Animals, Cats, "Food", Gastronomy

This dog used to live on the streets – now he's living (and eating) like a king 🍽️

Post video views	Post impressions
2,716,743	11,629,581

Content tags
Animal Lovers, "Food", Dogs, I Love Dogs, Rescue dog, Dogs Are Awesome

This cat has dinner with his mom every single night – and loves posing with their meals 🍽️

Post video views	Post impressions
2,659,781	15,017,689

Content tags
I Love Animals, Cats, "Food", Gastronomy

Here's an easy recipe for bunny treats 🐰

Post video views	Post impressions
1,966,955	9,968,574

Content tags
Animal Lovers, I Love Animals, Rabbits, "Food", "Snack food"

This dog used to live on the streets – now he's living (and eating) like a king 🍽️

Post video views	Post impressions
1,369,679	5,423,419

Content tags
"Food", Rescue dog, Dogs Are Awesome

Did it help?


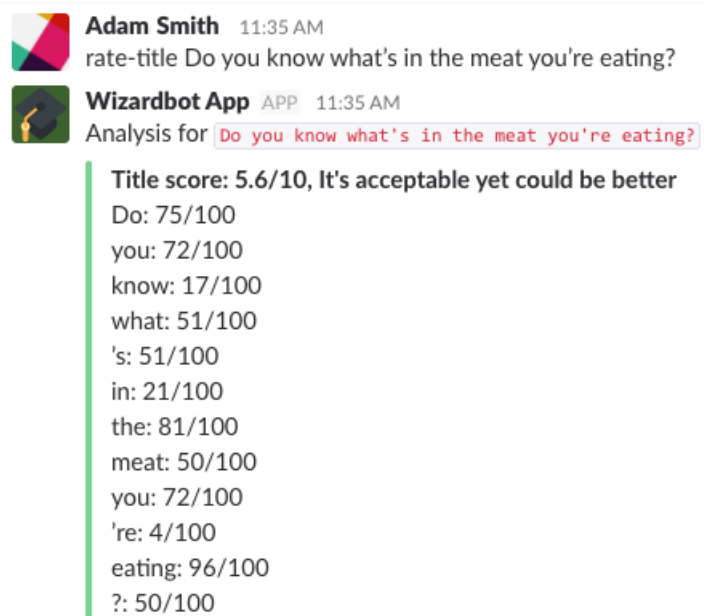
Reply... 

Figure 3: Sample conversation with a Slack bot using natural language understanding module. The bot is able to retrieve various statistics related to material published in the past.



Adam Smith 11:35 AM
rate-title Do you know what's in the meat you're eating?

Wizardbot App APP 11:35 AM
Analysis for `Do you know what's in the meat you're eating?`

Title score: 5.6/10, It's acceptable yet could be better

- Do: 75/100
- you: 72/100
- know: 17/100
- what: 51/100
- 's: 51/100
- in: 21/100
- the: 81/100
- meat: 50/100
- you: 72/100
- 're: 4/100
- eating: 96/100
- ?: 50/100

Figure 4: Sample use case of headline optimization model API deployed as a Slack bot functionality. Using bi-directional LSTM neural network model with attention, we are able to estimate popularity of a video as well as determine contributions of individual words.

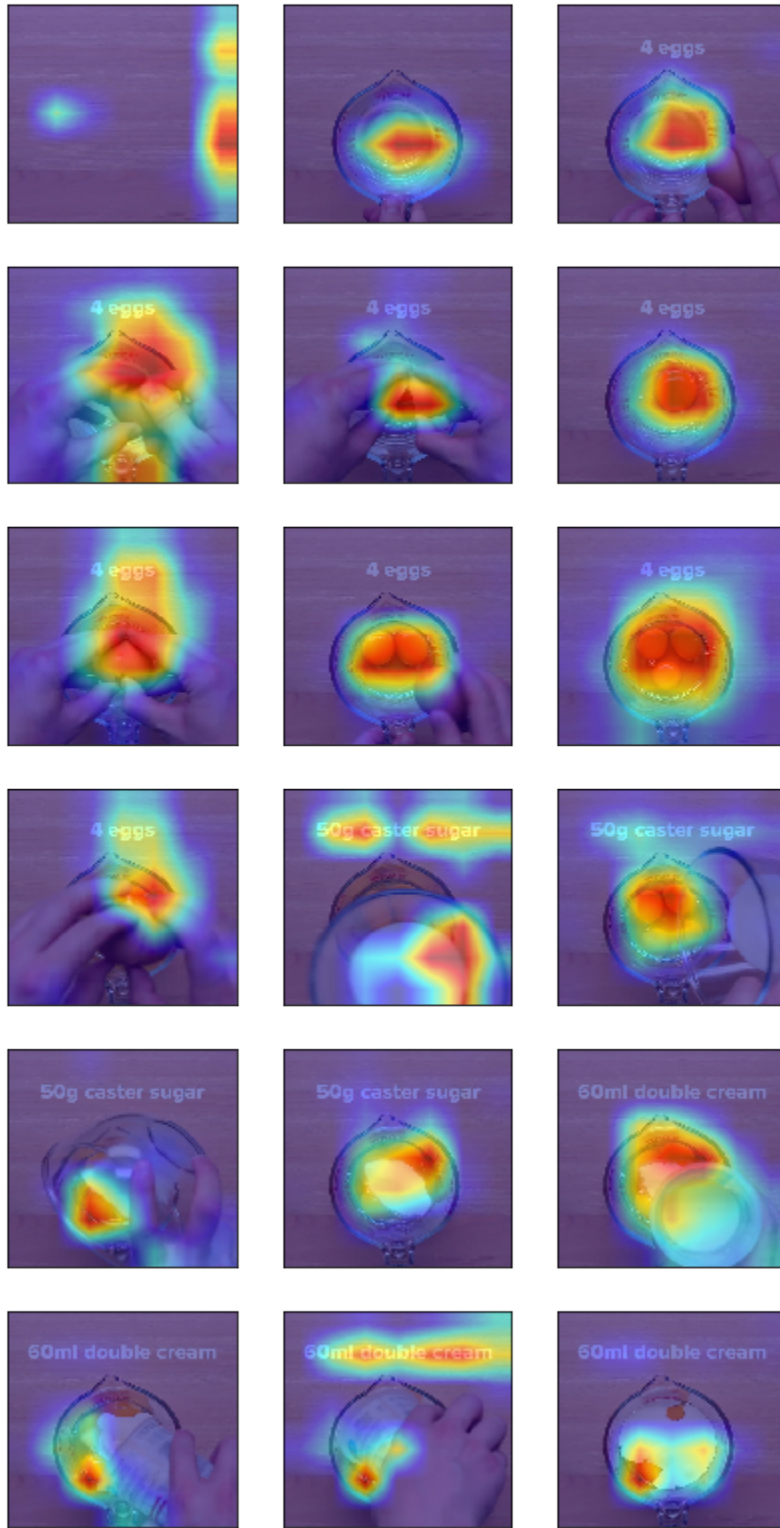


Figure 5: Sample GradCAM visualizations of a video opening scene frames generated using deep neural network trained for popularity classification. The visualizations can be used by video editors to analyze the importance of specific frame parts to social media users.

Rate video popularity

Upload your video

Select

Upload

Popularity score: 0.768

Left: original image, middle: heatmap for class *popular*, right: heatmap for class *not popular*

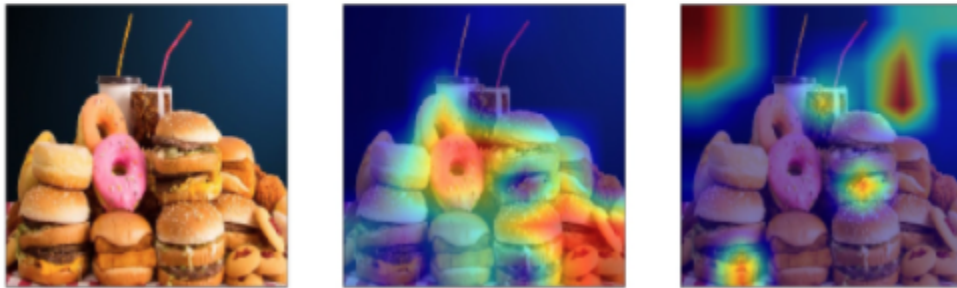


Figure 6: A working Web-based application for predicting future popularity of a video using deep visual features. Our model uses deep visual features extracted from penultimate layer of the ResNet50 model to classify content as popular or unpopular. The probability of class 'popular' is used as popularity score displayed in the application. Furthermore, GradCAM visualisation is used to identify parts of the frame contributing to the popular (middle image) and unpopular (right image) class.