Success Prediction on Crowdfunding with Multimodal Deep Learning

Chaoran Cheng, **Fei Tan**, **Xiurui Hou** and **Zhi Wei** *
Department of Computer Science, New Jersey Institute of Technology, USA
{cc424, ft54, xh256, zhi.wei}@njit.edu

Abstract

We consider the problem of project success prediction on crowdfunding platforms. Despite the information in a project profile can be of different modalities such as text, images, and metadata, most existing prediction approaches leverage only the text dominated modality. Nowadays rich visual images have been utilized in more and more project profiles for attracting backers, little work has been conducted to evaluate their effects towards success prediction. Moreover, meta information has been exploited in many existing approaches for improving prediction accuracy. However, such meta information is usually limited to the dynamics after projects are posted, e.g., funding dynamics such as comments and updates. Such a requirement of using after-posting information makes both project creators and platforms not able to predict the outcome in a timely manner. In this work, we designed and evaluated advanced neural network schemes that combine information from different modalities to study the influence of sophisticated interactions among textual, visual, and metadata on project success prediction. To make pre-posting prediction possible, our approach requires only information collected from the pre-posting profile. Our extensive experimental results show that the image features could improve success prediction performance significantly, particularly for project profiles with little text information. Furthermore, we identified contributing elements.

1 Introduction

Crowdfunding platforms, like Kickstarter (kickstarter.com) and GoFundMe (gofundme.com), are emerging portals for designers, artists, startups, small businesses and entrepreneurs to raise funds for their projects through the internet. Such platforms provide opportunities for all the people having creative ideas to pitch a campaign to gather capital and bring their ideas into reality. The fundraisers typically seek funding and in return provide certain rewards either in the type of future tangible products, experience, service, or acknowledgment of donation. Moreover, the audience demographics, aesthetic design, and terminologies are entirely different across varied fields spreading from arts to sciences on crowdfunding platforms. For example, the owner of a local restaurant started a campaign to rebuild their business which was devastated by Superstorm Sandy, and in return the backers will have their names listed on restaurant's website or a free dinner for family as acknowledgement[1]; a technician posted a profile explaining his innovative product, monetary goal, and timeline to deliver the product which the backers could have after the project is completed[2].

Project success prediction in crowdfunding is very challenging. There are increasing recent efforts to elucidate contributing factors to make a project to be funded successfully. The major existing studies consider only dynamic factors after projects are posted. Several factors relevant to success have been identified including number of backers [Etter *et al.*, 2013; Zhang *et al.*, 2015], promotion on social media [Etter *et al.*, 2013; Li *et al.*, 2016], comments and replies [Zhao *et al.*, 2017], or fund pledged dynamic during campaign [Zhao *et al.*, 2017]. All those postlaunch factors show powerful predictive potential. Nevertheless, both project creators and platforms can predict the outcome only after the project is posted for a certain period of time. It is desired if we can predict project outcome even before a project is posted. For platforms, they can reserve spaces for the projects more likely to succeed; for creators, they may revise their profiles proactively and save their time of going through predicted failure.

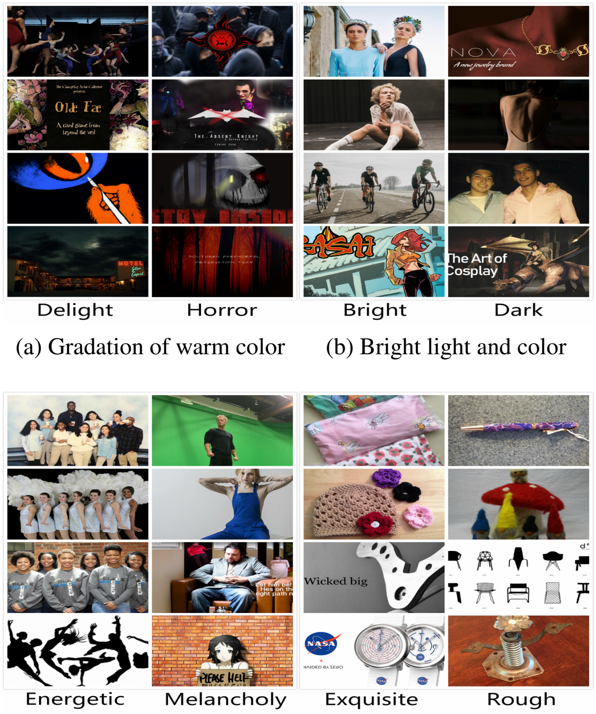
Moreover, most existing researches focus on primarily the text in profiles. Visual images as one important modality in pre-launch profile have not been studied yet, to the best of our knowledge. In most contexts, images are used to deliver ideas in a more effective way than text. For example, the product designers have to describe their concrete idea about prototype by images, while artists need to demonstrate their artifacts by showing visible sketch. Some funding campaigns would illustrate their blueprint merely by images instead of words[3]. With the huge collection of images available, it's appealing

[*] Corresponding author

[1] www.kickstarter.com/projects/573995669/rebuild-a-better-bait-and-tackle

[2] www.kickstarter.com/projects/jalousier/flipflic

[3] www.kickstarter.com/projects/414768297/keepers-of-the-moonandsun-english-edition



(a) Gradation of warm color (b) Bright light and color

(c) Energetic gesture (d) Exquisitely crafted objects

Figure 1: Examples of profile images on crowdfunding platform Kickstarter. For each case, images in the left column are from the successful projects, while the ones in the right column are from the failed projects. There is clear difference in visual style between successful and failed projects.

to ask: can we improve project success prediction accuracy by leveraging image information? There are two major confronted challenges, however. First, as examples shown in Fig. 1, image content understanding in the context of crowdfunding is far from trivial. Second, there may exist complex interactions among different modalities (image, text, animation, etc) in a profile that work together to deliver success. We need to investigate and evaluate different multimodal representations and find a principled way to integrate the heterogeneous textual and visual information, as well as other modalities.

In this work, we proposed a Multimodal Deep Learning (MDL) model to predict the success of crowdfunding projects[4]. The major contributions of this paper can be summarized as follows.

- We designed the multimodal feature representation for the profile with textual, visual contents, and metadata. We investigated fusion schemes with different modalities and evaluated our multimodal architecture on the real crowdfunding dataset.
- We systematically investigate the contribution of images to project success. Our extensive experiments proved the effectiveness of images on promoting the outcome prediction from different aspects.
- Our approach requires only the information collected from the pre-launched profile, which makes early predicting project outcome possible. Such early prediction will benefit to both platforms and creators.

2 Related Work

2.1 Project Success Prediction

The booming trend of crowdfunding has drawn much attention from academia. The conventional approach was to build a machine learning classifier like SVM based on meta features from campaign profile. Greenberg *et al.* [2013] showed an improvement by utilizing various decision tree algorithms and SVM trained with features such as whether the video was present, the sentence count in profile, project goals, project duration, and other possible additional factors like creators' demographic attributes. Some advanced approaches utilize textual description while certain models additionally exploit dynamic information by monitoring social media or crowdfunding campaign. Mitra and Gilbert [2014] analyzed the linguistic features with 59 other common features to predict project success. Yuan *et al.* [2016] proposed a text analytic topic framework to predict the fundraising success by extracting latent semantics from the text description with a combination of common numerical features. Etter *et al.* [2013] and Zhao *et al.* [2017] studied dynamic time-series factors by tracking the social media and monitoring the dynamic features like backer and money pledged status during the campaign. Zheng *et al.* [2014] found the degree of the creator's social network is positively associated with project success since creators can broadcast their crowdfunding projects to more audiences through the social network. Li *et al.* [2016] formulated the success prediction problem from the aspect of censored regression and achieved better performance utilizing temporal features from pledging and social media dynamics.

From the above descriptions, we can observe that most of the previous works focused on textual profile and postlaunch information which makes both project creators and platforms not able to predict the outcome in a timely manner. Therefore, to make pre-posting prediction possible, our approach focuses on the joint analysis of the textual and visual information collected from the pre-launch stage, which has not been fully explored yet in previous studies.

2.2 Multimodal Analysis

The multimodal approaches using text and image joint analysis have been explored in social media quite a while, e.g., multimodal semantic analysis [Roller and Im Walde, 2013], multimedia market evolution monitor [Zhang *et al.*, 2015], and multimodal news analysis [Ramisa Ayats, 2017]. To encode visual information, most of the earlier approaches relied on hand-crafted features combined with methods to aggregate manually engineered descriptors before the rise of CNNs. The Bag-of-Visual-Words (BoVW) [Csurka *et al.*, 2004] model was the common choice of image feature representation. It would collect codewords from feature descriptor like Scale-Invariant Feature Transform (SIFT) [Lowe, 2004]

[4] The data and code will be released at https://github.com/andrea1980s/Kickstarter

, and then learn the codebooks from unsupervised learning. In more recent works, hybrid architecture was introduced to leverage the best of the BoVW and deep neural networks. One approach was to combine the pre-trained deep features with BoVW [Zhang *et al.*, 2015], while some other approaches projected the hand-crafted feature descriptors to lower dimensionality and feed them to neural networks [Perronnin and Larlus, 2015].

The pre-trained CNN in large databases, like ImageNet, can be used as off-the-shelf feature extractors. The transfer feature learning from existing deep convolutional neural networks showed promising results in varied researches [Sharif Razavian *et al.*, 2014; Kiela and Bottou, 2014; Ma *et al.*, 2016; Wang *et al.*, 2016]. Nevertheless, this approach has not been proved to be an effective corollary in crowdfunding project success prediction. Moreover, the adoption of existing approach was hindered by the heavier sparsity of image fusion which need to be carefully addressed in this problem.

3 Problem Formulation

As a crowdfunding project example illustrated in Fig. 2, the typical structure of the campaign page includes a goal, a project description (red box at the bottom) embedded with tens of figures (red box in the middle), perk structure (right column of the webpage), links to creator's social media platforms, and some other metadata (red box on the top) like category, location.

Let $\mathcal{D}$ represent the crowdfunding dataset with N profiles. The definition of success is that the creators can reach their initial goal by the end of the limited campaign period. Then $y_k \in \{\pm 1\}$ specifies the ground-truth outcome whether project $k \in [1, N]$ is successful or failed. Our goal is to learn a multimodal feature map $F(X)$ for given $\mathcal{D}$ to predict the success outcome $\mathbf{Y}$. The feature mapping is defined as:

$$F(X) = f(W_F \cdot X + b_F) \tag{1}$$

where $X = \phi(X_T) \oplus \phi(X_I) \oplus \phi(X_M)$, the symbol $\oplus$ means concatenation, $f(\cdot)$ is a non-linear activation function such as rectified linear unit (ReLu), and W_F, b_F are weight and bias, respectively. In this equation, $\phi(\cdot)$ can be considered as a feature mapping of modalities for text X_T, images X_I, and metadata X_M.

In our work, the training objective function is based on cross entropy:

$$\mathcal{L} = -\sum_{k=1}^{N} [\delta(y_k = 1)\log(p_k) + \delta(y_k = -1)\log(1 - p_k)] \tag{2}$$

In the above, $\delta(\cdot)$ is the indicator function, and $p_k \in [0, 1]$ is the estimated probability for the class with label $y_k = 1$. And our implementation of loss layer combines the sigmoid operation for computing p_k given $F(X)$.

4 Joint Fusion of Heterogeneous Features

Fig. 2 illustrates how our system computes multimodal representations. As shown in Fig. 2, our MDL model has three branches: (1) bottom branch for encoding textual input $\mathcal{T}$; (2) middle branch for encoding visual content of images $\mathcal{I}$; (3) top branch for encoding meta information $\mathcal{M}$. Each branch is composed of either CNN subnet or fully connected hidden layers. In general, each branch can have a different number of layers, and the inputs for the 3 branches could be produced by their own upstream networks, such as word embedding or pre-trained ImageNet. At the end of each branch, the feature maps from three streams are concatenated into one feature map.

Textual Feature. We applied two popular feature representations for the textual input: Bag of Words (BoW) with Term Frequency-Inverse Document Frequency method (TF-IDF), and word embedding. In BoW, the text in a given profile is encoded as a histogram of weights for the words appeared in the $\mathcal{T}$, and the weights normally would collect from TF-IDF weighting scheme. Despite the generated feature vector ignores the order and semantics of the word, the BoW model shows magnificent power in tons of varied NLP applications. In contrast with the sparse BoW representation, word embedding encodes text as the continuous distributed representation of a short fixed-size vector. It's semantically compatible to represent both word and its related context. The embedding model we used is GloVe [Pennington *et al.*, 2014] with 300-dimensional vectors.

Visual Feature. The pre-trained ImageNet is used to extract feature map for each individual image. Specifically, we use a pre-trained 16-layer VGG model [Simonyan and Zisserman, 2014] and take its output from the fully-connected layer (fc6). For any project k, $I_k \subseteq \mathcal{I}$ as its input of images, and additionally $\ell_\imath \in I_k$ in which $\imath$ is the index of images in profile k which may contain multiple images with varied size n. Given an image ℓ_i, it's rescaled to 224 × 224 and represented by a 4096-dimensional vector extracted from VGG16 model. Image feature maps for any $\ell_i, i \in [1, n]$ in I_k are used as visual input for profile k. Then we applied two popular approaches to generate the visual representations for profiles: BoVW and CNNs. To work with BoVW, the pre-trained deep features are used as descriptors and clustered using mini-batch k-means model to generate the codebook. Each profile is then represented by a bag of visual words. While for CNNs, those pre-trained deep features are aggregated in different ways, like averaging, flattening, and stacking. Kornblith *et al.* [2018] suggested ResNets are the best feature extractor for transfer learning tasks. However, we observed better performance using VGG16 model. Considering our feature vectors are generated from large group of images with heavy sparsity, and the style of images on crowdfunding platforms differ from ImageNet-like dataset as shown in Fig. 1, this hypothesis may justify the contradictory observation in our experiments to some extent.

Meta Feature. Meta data in our experiment is composed by campaign category and funding goal extracted from the profile. Funding category could be converted to the one-hot encoder directly while funding goal is converted with the binning transformation. We used binning transformation to group numerical funding goal values into a set of customized discrete ranges, and then assigned each numerical value to a range bin. Specifically, we summarized the data and used (1) fewer bins to encode numeric values falling inside the lower

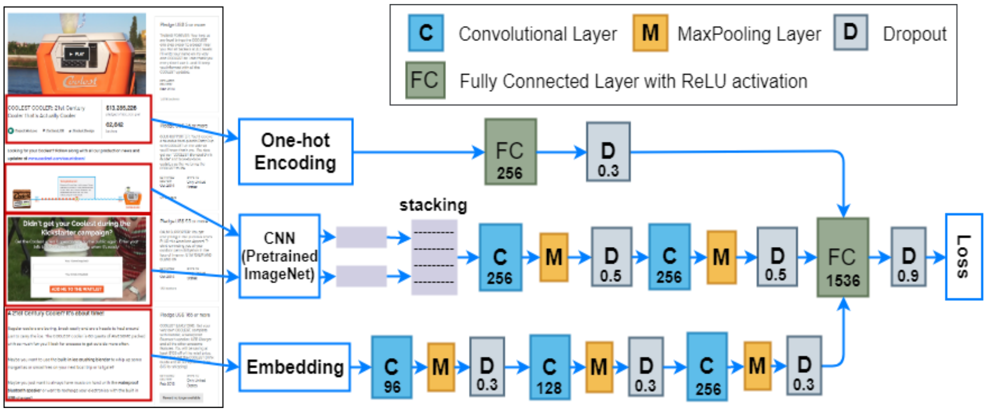


Figure 2: Framework Overview of Our Multimodal Deep Learning Framework. The input is from pre-posting profile features. There are three pipelines in framework: (1) top branch for meta modality; (2) middle branch for visual modality; (3) bottom branch for textual modality.

and higher quantile; (2) more bins for other values belonging to the in-between quantile. Then accordingly, each numeric value of the funding goal would be assigned to a binary vector. More specifically, we set 3 bins for the lower 0 - 0.3 quantile, 4 bins for the upper 0.9 - 1 quantile, and 50 bins for the values in between.

5 Experiment

5.1 Data Set

Data Collection. The evaluation of our MDL framework is done on the dataset scraped from Kickstarter. Kickstarter is the largest crowdfunding platform in the U.S. since it was launched in 2009. The campaign profile may be modified during the funding stage and thus not exactly the same as its prelaunch profile. Yet it's likely to be served as a broadly useful model for examining crowdfunding efforts. We crawled the data using the seed from webrobots[5]. We ran the scraping script for 2 weeks to collect the data. For each campaign webpage, we extracted the textual profile, images, category, funding goal, funding start date, and end date. The textual profile includes the title, summary, funding description, and risk/challenges. Visual contents include all the jpg format images but not the png or gif format. Furthermore, the campaigns before 2015 were discarded.

Preprocessing. Before feature extraction, we carried out the following data preprocessing procedures:

1. *Project status*. The status of the funding campaign on Kickstarter include successful, failed, canceled, suspended, removed, live, and others. We only kept the funding projects in the status of successful and failed.
2. *Textual profile*. We removed the stop words, punctuation and digits from the text, and further removed the project with text length less than 12 words.
3. *Visual profile*. We removed small images whose row dimension size is less than 200. We observed the small images function as banner since most of them are the aesthetic shape or section header. It will increase the aesthetic value of the profile for sure, but bring excessive extra computational cost.
4. *Category*. We also removed the projects in the skewed categories. If the projects listed in a given category is less than 1/10 of the largest category in our data, we removed those categories and projects.

Data Splitting. To train our model, select best parameters and evaluate the performance, we split the dataset into 3 parts as shown in Table 1 which describes the statistics of our dataset grouped by their campaign year and funding outcome. The reported overall success rate of Kickstarter is 43% at 2011[6] , and increased to 75% at 2017[7] . The discrepancy in our data could be explained by our scraping procedure and data cleaning process. To evaluate the predictive potential on shifting distribution, we set campaigns launched at 2015 and 2016 as the training set, campaigns at 2017 as the validation set, and campaigns at 2018 as the testing set.

Data Set	Year	No. of Samples			No. of
		Success	Fail	Total	Images
Training	2015/16	8705	9806	18511	135404
Validation	2017	4655	2687	7342	65907
Testing	2018	5429	2830	8259	83102

Table 1: Statistics of Our Dataset. To evaluate the performance on shifting distribution, the data is splitted by year.

5.2 Experiment Settings

Evaluation Metrics. We evaluated the prediction performance of all methods in terms of Recall, Precision, F1-Score, and AUC@ROC (AUC), respectively.

Evaluation Algorithms. Two important inspiration of our work are to introduce image as an additional modality and to limit the resources to pre-posting profile. Neither perspective has been addressed in previous work. Thus the focus

[5] https://webrobots.io/kickstarter-datasets/

[6] www.kickstarter.com/blog/happy-birthday-kickstarter

[7] www.kickstarter.com/blog/happy-8th-birthday-kickstarter

Level	Quantile	No. of Words	No. of Images
1	- 0.25	150	1
2	- 0.5	260	4
3	- 0.75	450	10
4	- 1(0.9)	4572(725)	113(18)

Table 2: Statistics about Number of Words and Images in Profiles We used 0.9 quantile as cut-off for feature dimension in the MDL model, so the values inside the parentheses are the actual value for text length and image numbers in our experiments.

of our evaluation lies in the investigation of different modalities and feature representations for project success prediction without any post-posting information. We compared our proposed framework with the linear SVM on all investigated cases. Our study explored the following questions and corresponding methods:

1. *Which performance levels can be achieved by textual profile information only?* We studied classifications based on textual information only: **SVM-BoW** and **MDL-Text**, and evaluated the boundaries of textuality only. The best text-based approach serves as the baseline.
2. *Which performance level is achievable by visual information?* We investigated the suitability of the visual modality only and evaluated different aggregation strategies for **SVM-BoVW** and **MDL-IMG**. The best image-based approaches are reported.
3. *Which multimodal representation performs best? Does the multimodal combination of different feature representations facilitate classification?* We compared the methods with all modalities with Text, Image, and Metadata (TIM), and reported the results of **SVM-TIM** and **MDL-TIM**. Moreover, further study on varied combination of different modalities (**MDL-Text/IMG, MDL-Text/Meta, MDL-IMG/Meta**) are conducted to learn the most contributing element. Meanwhile, the parameters are tuned for each model, like the number of CNN layers, the value of kernel size for CNN layer, the value of pooling size for max pooling layer, the number of fully connected layer and the value of neuron units, the dropout rate for each dropout layer.
4. *How sensitive of the performance to the modalities?* More specifically, is the performance consistent on different project profiles with varied textual length and image numbers? We performed the ablation analysis to different granularity to test this.

Hyper-Parameter Tuning. For each evaluated case, we varied the most influential parameters to train the models. The models are tuned based on validation part and the optimal parameters are reported accordingly based on F1 measure. Our MDL model is implemented in Python using Keras with TensorFlow backend. We used the RMSprop [Tieleman and Hinton, 2012] optimizer and the learning rate is set to 1e-5 for 100 epochs. We set the batch size to 128 campaign projects and employed early stopping with 20 epochs and dropout [Srivastava *et al.*, 2014] to prevent overfitting. To reduce the computational cost of training CNN, we truncated the length of text description and the number of images in the profile. The cut-off value we used is 0.9 quantile for both. Specifically, it's 725 for words in text and is 18 for images as shown in Table 2. Nevertheless, we are still dealing with heavier sparsity of image fusion.

6 Results and Discussion

Baseline. As shown in Table 3, compared with MDL-Text, SVM-BoW achieved decent performance in terms of all metrics, especially the precision and recall. Thus SVM with textual modality only is the baseline.

Methods	Recall	Precision	F1 Score	AUC
SVM-BoW	**0.7356**	**0.7424**	**0.7387**	**0.7356**
MDL-Text	0.6831	0.7153	0.6920	0.7788

Table 3: Results of single modality on textuality only

Visual Modality. We trained a linear SVM classifier with BoVW feature obtained by mini-batch k-means with clusters size varied from 30 to 300. With respect to the visual representation for MDL-IMG, we tried different approaches to aggregate the feature map $\ell_i, i \in [1, n]$ where $n = 18$ for given I_k, e.g., average pooling, flattening, and stacking. The stacking approach performed best on the MDL-IMG model. As we can see from Table 4, the MDL-IMG model yield better visual representations than BoVW. This demonstrated the deficiency of BoVW while dealing with large dataset due to the complexity and diversity of image pattern, and confirmed the superiority of the state-of-art feature transfer learning from computer vision. However, its performance is worse than MDL-Text.

Methods	Recall	Precision	F1 Score	AUC
SVM-BoVW	0.6570	0.6623	0.6592	0.6569
MDL-IMG	**0.6738**	**0.6809**	**0.6768**	**0.7340**

Table 4: Results of single modality on visuality only

Multimodal Combination. We utilized all collected textual, visual and meta information to evaluate their effects on performance and investigated different feature fusing approaches. The fusion techniques could be complicated as [Tan *et al.*, 2018], we found the FC layer achieved most effective results in our work. As demonstrated in Table 5, both SVM-TIM and our MDL-TIM outperformed baseline method SVM-Text, which confirms the motivation outlined in the introduction: images as an important component in profile could help to improve the predictive performance. Meanwhile, our proposed model MDL outperforms SVM which demonstrated the superiority of deep transfer learning in computer vision over the BoVW model.

Methods	Recall	Precision	F1 Score	AUC
SVM-TIM	0.7411	**0.7595**	0.7483	0.7411
MDL-TIM	**0.7505**	0.7568	**0.7534**	**0.8326**

Table 5: Comparison with all modalities

Ablation Analysis. Additionally, we investigated the most contributing modality by removing one factor at a time. The performances of different modality combinations (Text/Image, Text/Meta, Image/Meta) is reported in Table 6. As we can see that IMG/Meta performs worst after removing text from model, thus text still carries more predictive message in general. Meanwhile, we observed that meta data could introduce more distinctive signals. It becomes clear after checking the images on Kickstarter because the patterns in different categories vary from each other indeed, e.g., styles of product design do differ with comic books tremendously in all respects.

Combinations	Recall	Precision	F1 Score	AUC
Text/IMG	0.7335	0.7241	0.7278	0.7995
IMG/Meta	0.7108	0.7191	0.7143	0.7807
Text/Meta	0.7385	0.7320	0.7348	0.8162

Table 6: Ablation Analysis of MDL

Correlation	Text+	Image+	All
Text Level	0.618	—	—
Image Level	—	0.776	0.570

Table 7: Correlation between accuracy and images/text levels. The entries with p.value $>$ 0.05 are removed.

Sensitivity analysis of the modalities. Furthermore, we investigated the effects of textual length and image numbers to test the sensitivity of performance on different modalities. Firstly, we split profiles into 4 levels as statistics shown in Table 2. Then we analyzed the accuracy of models MDL-Text/Meta (Text+), MDL-Image/Meta (Image+) and MDL-TIM (All) from Table 5 and reported the fine-sorted results in Fig. 3. In general, MDL-TIM outperforms significantly in most cases. Moreover, we surprisingly found that the MDL-Image/Meta outperforms MDL-Text/Meta in the last column (image level 4 across all text levels). This indicates better prediction can be achieved if given more images than text. Interestingly, we found that better success prediction won't be necessarily achieved if given more text by checking the trend from bottom row (text level 1 across all image levels) to the top row (text level 4 across all image levels). The correlation between accuracy and text/image level reported in Table 7 also supported our observation. Giving this important revelation, it's suggested that the performance of success prediction will be enhanced by integrating images despite of limited text description. Yet to have the best outcome, all modalities should be considered. To this end, we demonstrated that the prevailing role of images in crowdfunding success prediction.

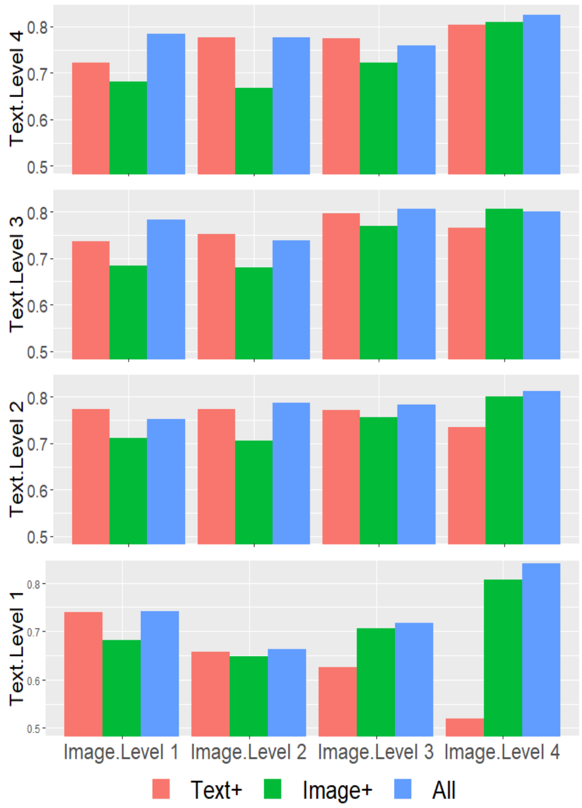


Figure 3: Investigation on the ablation of modalities grouped by the length of text profile and numbers of images. Accuracy is reported here by split Levels of Word and Image Number in Profile for models MDL-Text/Meta (Text+), MDL-Image/Meta (Image+) and MDL-Text/Image/Meta (All). The profiles are split into groups by definition in Table 2. The higher the level is, more words or images in the profiles.

7 Conclusion

In this work, we explore the project success prediction problem and developed a multimodal deep learning framework to bridge the gap of missing visual modality in previous prediction algorithms. Particularly, the exploration of multi-modalities demonstrates the outperformance of our proposed approach in terms of capturing success signal from pre-posting profile of textual, visual, and meta data. The extensive experiments are conducted on real world data set collected from Kickstarter. The empirical studies illustrate that visual images are non-negligible as text and superior performance could be achieved by collaborating with them. The corresponding results show that our model can deliver the best performance over alternative methods. The ablation analysis of the modalities also provides useful insights for project creators.

References

[Csurka *et al.*, 2004] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[Etter *et al.*, 2013] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. Launch hard or go home!: predicting the success of kickstarter campaigns. In *Proceedings of the first ACM conference on Online social networks*, pages 177–182. ACM, 2013.

[Greenberg *et al.*, 2013] Michael D Greenberg, Bryan Pardo, Karthic Hariharan, and Elizabeth Gerber. Crowdfunding support tools: predicting success & failure. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*, pages 1815–1820. ACM, 2013.

[Kiela and Bottou, 2014] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, 2014.

[Kornblith *et al.*, 2018] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? *arXiv preprint arXiv:1805.08974*, 2018.

[Li *et al.*, 2016] Yan Li, Vineeth Rakesh, and Chandan K Reddy. Project success prediction in crowdfunding environments. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 247–256. ACM, 2016.

[Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Ma *et al.*, 2016] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *AAAI*, volume 3, page 16, 2016.

[Mitra and Gilbert, 2014] Tanushree Mitra and Eric Gilbert. The language that gets people to give: Phrases that predict success on kickstarter. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 49–61. ACM, 2014.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[Perronnin and Larlus, 2015] Florent Perronnin and Diane Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3743–3752, 2015.

[Ramisa Ayats, 2017] Arnau Ramisa Ayats. Multimodal news article analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 5136–5140, 2017.

[Roller and Im Walde, 2013] Stephen Roller and Sabine Schulte Im Walde. A multimodal lda model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, 2013.

[Sharif Razavian *et al.*, 2014] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Srivastava *et al.*, 2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[Tan *et al.*, 2018] Fei Tan, Zhi Wei, Jun He, Xiang Wu, Bo Peng, Haoran Liu, and Zhenyu Yan. A blended deep learning approach for predicting user intended actions. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 487–496. IEEE, 2018.

[Tieleman and Hinton, 2012] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[Wang *et al.*, 2016] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[Yuan *et al.*, 2016] Hui Yuan, Raymond YK Lau, and Wei Xu. The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, 91:67–76, 2016.

[Zhang *et al.*, 2015] Hao Zhang, Gunhee Kim, and Eric P Xing. Dynamic topic modeling for monitoring market competition from online text and image data. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1425–1434. ACM, 2015.

[Zhao *et al.*, 2017] Hongke Zhao, Hefu Zhang, Yong Ge, Qi Liu, Enhong Chen, Huayu Li, and Le Wu. Tracking the dynamics in crowdfunding. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 625–634. ACM, 2017.

[Zheng *et al.*, 2014] Haichao Zheng, Dahui Li, Jing Wu, and Yun Xu. The role of multidimensional social capital in crowdfunding: A comparative study in china and us. *Information & Management*, 51(4):488–496, 2014.