Image Recommendation for Wikipedia Articles

Oleh Onyshchak[1] and Miriam Redi[2]

[1] Ukrainian Catholic University, Lviv, Ukraine
o.onyshchak@ucu.edu.ua
[2] Wikimedia Foundation. London, UK

Abstract. Multimodal learning, which is simultaneous learning from different data sources such as audio, text, images, is a rapidly emerging field of Machine Learning. It is also considered as machine learning at the next upper level of abstraction. This allows tackle more complicated problems such as creating cartoons from a plot or speech recognition based on lips movement. In this paper, we propose to research whether state-of-the-art techniques of multimodal learning, will solve the problem of recommending the most relevant images for a Wikipedia article. In other words, we need to create a shared text-image representation of an abstract notion which paper describes, so that having only a text description machine would ”understand” which images would visualize the same notion accurately.

Keywords: multimodal learning · text-image similarity · image recommendation

1 Introduction

Every day we perceive the world around us through multiple cognitive feelings such as sight, smell, hearing, touch, taste. Moreover, our ability to consolidate all the information from different sources into one complete picture helps us comprehensively understand the world.

With a trend to digitizing in the last few decades, more and more information is recorded in different kinds of media such as audio, image, video, text, and 3D modeling. That also created new challenges of efficiently processing significant amounts of recorded information, where we already have substantial achievements. However, every type of digital storage only captures some subset of available information. For example, imagery only captures visual appearance, while audio – the sound, just as our eyes and ears do. Thus, all the scientific progress in processing some data carrier is bounded by the limitation of what that medium can capture.

In other words, to represent a dog digitally, we have to have more than just a visual representation. Similar to humans, we need to combine all the information streams, which describe the entity from different perspectives, into one comprehensive representation.

In: Proceedings of the 1st Masters Symposium on Advances in Data Mining, Machine Learning, and Computer Vision (MS-AMLV 2019), Lviv, Ukraine, November 15-16, 2019, pp. 29–39

That is the motivation for multimodal representation learning, which aims to combine different types of data into a complete representation of a real-world entity. In that context, the word ”modality” refers to a particular way of encoding information. Thus a problem in the domain of for example image processing is called unimodal, while a problem in the domain of multiple information encodings, for example image to caption generation, is called multimodal since it works with both image and text modalities [1].

By having a complete representation of an entity, which was created via multimodal data that captures complementary / supplementary information subsets of an object, we have more comprehensive computational ”understanding” of that entity. That helps us increase the precision of existing data science applications, and extend the limits to more abstract problems such as not only identify the objects in an image, but understand the value. For example [1], early research on speech recognition showed that, by involving the visual modality of lips movement on top of sound modality, we get extra information that allows us to increase the quality of voice recognition task, just as it works for humans [2].

In this project, we are going to research possible approaches toward image recommendation for Wikipedia articles problem, which is also the part of multimodal representation learning domain. That is, based on the article text, we need to recommend images describing the entity described in the article. In other words, we need to create a high-level representation of some entity, described by both text and images. Furthermore, we are interested to find out which image representation of the notion is the best suited for a given text description.

In scope of this project, we are going to explore the State-of-the-Art techniques of multimodal representation learning and analyze whether these could be applied to solving this problem. We believe this project will be valuable from both research and application perspectives.

This paper presents a project proposal of Master Thesis, which will formally define the problem, provide a rigorous overview of the State-of-the-Art approaches in Multimodal Representation Learning domain, specify the goals of the project, suggest an approach to envisioned solution, and provide a time plan of the thesis.

2 Motivation

Wikipedia is the biggest collection of human knowledge containing more than 35 million pages and having nearly 9 billion views per month[1] And it continually growing, having more than 500 new pages per day[2], and all of that only in its English version.

As a part of 2030 strategy, one of the key goals is to break down any barriers for accessing free information[3]. By researching the possibilities to recommend images for Wikipedia editors in an automated way, it will help get better media enrichment of

[1] https://stats.wikimedia.org/v2/#/en.wikipedia.org

[2] https://en.wikipedia.org/wiki/Wikipedia:Statistics

[3] https://meta.wikimedia.org/wiki/Strategy/Wikimedia movement/2017/Direction

articles, which in turn will make information easier and faster to comprehend [11]. Automation would also help reduce time and effort to be spent in searching for and adding appropriate article visualizations.

In addition to the motivation of making Wikipedia better, this work might present some useful insights to the development of multimodal learning field. Since this is: (1) a real-world problem, which might give us interesting insights of how to apply and adjust current academia progress; and (2) we have a more complicated problem setting of one large article corresponding to a multiple images, instead of more simplified one-to-one correspondence of images and respective tags or descriptive sentences.

3 Problem

We are going to research how the State-of-the-Art multimodal learning techniques perform on a task of recommending images for Wikipedia articles. In other words, having a text with wiki formatting, we need to rank images from Wikimedia Commons database [19] by relevance.

4 Data

All data is publicly available on Wikipedia. Specifically, we have more than 35 million Wikipedia pages with a fair amount of them enriched with images. We also have Commons image dataset [19], containing more than 55 million images[4]. That is the real-world data, where ultimately the solution should be applied.

In research, we will use a reliable subset of the above-specified data for training. In particular, Wikipedia has a notion of featured articles[5], which are the best articles having quality text and a lot of supporting visualization. In other words, it is a high quality dataset of more than 5000 articles, each having multiple associated images, that was manually created. Nevertheless, it still requires proper preprocessing and cleaning before using.

Particularly, by text we mean the entire article textual content cleared from Wikipedia formatting along with some extra metadata such as categories or title. Images are also collected with additional metadata such as filenames or descriptions. More details could be found on Kaggle[6]

5 Related Work

In the last decades, there was much progress in the field of unimodal representation; research in multimodal learning was limited by simple concatenation of unimodal features [4]. However, in recent years, the scientific landscape in this domain has being

[4] https://en.wikipedia.org/wiki/Wikimedia Commons

[5] https://en.wikipedia.org/wiki/Wikipedia:Featured articles

[6] https://www.kaggle.com/jacksoncrow/extended-wikipedia-multimodal-dataset

rapidly evolving [3]. One of the triggers for it was the success of deep learning models, which have a powerful representation ability with multiple levels of abstractions. Therefore, these models were incorporated in multimodal learning. As Guo et al. suggested [1], we can divide all the multimodal learning approaches into three categories:

1. Joint representation, which aims to integrate modality-specific features into some common space
2. Coordinated representation, which aims to preserve modality-specific features, while introducing a space to measure multimodal similarities
3. Intermediate representation, which aims to encode features of one modality to some intermediate space, from where we later generate the features of another modality

In this section, we will cover available techniques to extract features from text and image modalities, overview available solutions in each type of multimodal learning, and then summarize their applicability for our problem.

5.1 Unimodal Representation

Image. The most popular model used in feature extraction from images are different types of Convolutional Neural Networks (CNN), such as AlexNet [8], VGGNet [9], and ResNet [10]. When working with big datasets, it is preferable to use a pre-trained version of the chosen CNN. This field has tremendous development in recent years, and currently we already have well-defined solutions for most problems.

Text. A popular way to extract features from text is to encode it to vector, as is done in word2vec [13] or Glove [14] algorithms. Although, the common problem with those approaches is when some words are not present in vocabulary or out-of-vocabulary error. However, there is a variety of alternative solutions to this problem, such as character embeddings [15]. Specifically, an alternative and more powerful tool for dealing with text is Recurrent Neural Networks (RNN) [16], which is more context-aware and can make better encoding of the *n*-th word, knowing what was already in a sentence. One of the most successful realizations of RNN is Long Short-Term Memory (LSTM) [17].

5.2 Joint Representation

The main idea of joint representation is to integrate multimodal features into a single input, which we then process as some artificial unimodal input with well-known machine learning techniques. More formally, it aims to project unimodal representations into a shared semantic subspace, where the multimodal features can be fused [3], as shown in Fig. 1(a). Until recently, it was the primary technique in multimodal learning, where shared features were fused by concatenating these together. However, now, the most popular choice is to use a distinct hidden layer, where modality-specific features will be combined into a single output vector.

Concatenation approach was historically the first and is still commonly applicable in video classification [18], event detection [20], and visual question answering [21].

However, its main disadvantage is neglecting the fact that different modalities have not only supplementary information that is representing the same notion from different perspectives, but also complementary information, where one modality captures the information, which another cannot capture. For example, lips movement and audio of a speech are mostly supplementary sources, while images of some bird and audio of it singing are mostly complementary sources. Because of that, much information gets lost in that shared space.

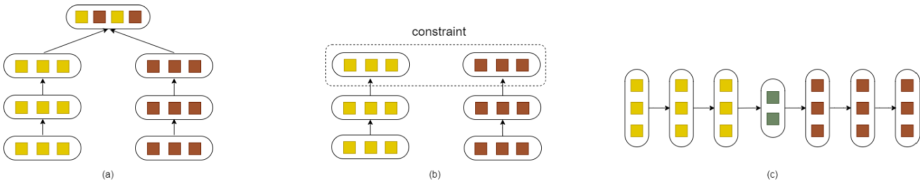

Fig. 1. Three types of frameworks for deep multimodal representation[7]: (a) joint representation aims to learn a shared semantic subspace; (b) coordinated representation framework learns separated but coordinated representations for each modality under some constraints; (c) intermediate representation framework translates one modality into another and keep their semantics consistent [1].

Although it has advantages of being a simple method and producing modality-invariant common space of features, it cannot be used to infer the separated representations for each modality [1]. Thus, the methods from this category are not applicable to our problem.

5.3 Intermediate Representation

Intermediate representation models aim to encode features of one modality to some intermediate space, from which later features of another modality can be generated (or decoded), as shown in Fig. 1(c). To prevent the intermediate space from being related only to a source modality, during encoder-decoder training we maximize, e.g., the likelihood of the target sentence given source image, so that the error function employs the error of decoding. Subsequently, the generated intermediate representation tends to capture the shared semantics from both modalities [1].

Some interesting application of that model was proposed by Mor et al. [22], where the algorithm encodes a musical track into intermediate space, which is then decoded by multiple decoders into a space of some specific instrument. In other words, the encoder extracts instrument-invariant generic musical features, which then each decoder transforms into features of its target instrument.

The general advantage of this approach is that it is one of the best ways to generate new features in a target domain. Thus, this technique is used in image caption [23], video description [24], and text to image [25] generations. The disadvantages of that

[7] This figure has been drawn for this paper for illustrative purposes based on the inspiration from [1].

model are: (1) it can only encode one modality; (2) the complexity of designing a feature generator should be taken into account [1]; and (3) intermediate space extracts only shared subspace from two modalities. Moreover, because we need to query existing information rather than generate one, these methods are also not suitable for our problem solution.

5.4 Coordinated Representation

The last type of multimodal learning is coordinated representation. Instead of learning from a joint representation, it learns from modal-specific representations separately but with a shared constraint, which is some loss function identifying cross-modal similarity / correlation. Since different modalities hold unique information about an object, this approach operates with all available knowledge. A visual explanation can be seen in Fig. 1(b). Regarding a constraint function, a commonly used option is a cross-modal similarity function, where learning objective is to preserve both inter-modality and intra-modality similarity structure. In other words, it would force cross-modal distance for elements with the same semantics to be as small as possible, while with dissimilar – as big as possible.
The cross-modal ranking is a widely used constraint, where the loss function is defined in the following way:

$$\sum_i \sum_{t^-} max(0, \alpha - S(i,t) + S(i,t^-)) + \sum_t \sum_{i^-} max(0, \alpha - S(t,i) + S(t,i^-)) \,, \quad (1)$$

where (i,t) is a matching image-text pair, α is margin, S is a similarity function, i^- is mismatching pair to t and vise-versa. Frome et al. [27] used a combination of dot-product similarity and margin rank loss to learn a visual-semantic embedding model (DeViSE) for visual recognition [1]. DeViSE trains deep networks for both image and text features, and then adjusts features based on the above-mentioned ranked loss, though in a more simplified form.

Alternative to cross-modal ranking, another widely used constraint is Euclid distance, which is also used for ensuring that similarity structure for both intra-modality and inter-modality is preserved. That is, for inter-modality, we map text and image features into low-dimensional space, where we can calculate the distance between feature vectors. The idea here is to ensure that inter-modality features of the same semantics are as close as possible [30]. While for intra-modality, we want to preserve the similarity between neighborhood items, that is:

$$d(m_i, m_j) + m < d(m_i, m_k), \forall m_j \in N(m_i), \forall m_k \notin N(m_i) \,, \quad (2)$$

where m is a data point of any modality, m_i is a point of interest, $N(m)$ denotes the neighborhood of m [31].

Hence, coordinated representation preserves all modality-specific information. It also explicitly compares features from different modalities, thus having data from one modality, we can identify the closest data point from another modality. Because of those properties, it is used for cross-modal retrieval [31], retrieval-based visual description [29], and transfer knowledge across modalities [30]. Thus, it can be applied to our problem of image recommendation for articles, and we will proceed with those methods.

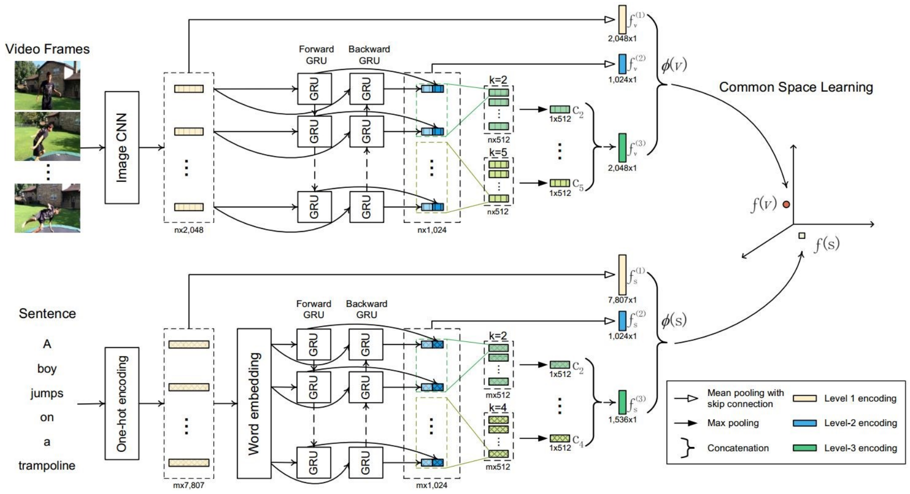

Fig. 2. Example of Coordinated Representation learning pipeline[8]

6 Solution Approach

Based on the analysis of the related work, coordinated representation techniques were identified as the most relevant approach to solve our problem. Coordinated representation approach aims to exploit modality-specific features fully, thus we train each feature modality separately.

To make the system learn right features in each modality, we map all of them into space where inter-modality similarity can be evaluated but also preserving the intra-modality similarity structure [12, 27, 28]. Then we identify loss function, by enforcing ranking function (1) in that space to return high values for mismatches modality pairs and small otherwise. That would be a loss function, which each modality-specific model will be minimizing, thus empowering modality-specific feature learning. You can see visualization of this idea on Figure 2.

We will focus on integrating recent Word2VisualVec [5] and dual encoding [6] models to our more broader and more realistic problem settings. They showed impressive results but were evaluated on a narrower problem. More specifically, they were working with the Flickr dataset [7] where one image corresponds to 5 descriptive sentences. In our settings, we have one article corresponding to multiple images, where all of them having additional metadata such as category, name, description.

This paper is supported by Github repository[9] with all experiments.

[8] The figure is adopted from the GitHub resource (https://github.com/danieljf24/dual_encoding/blob/master/dual_encoding.jpg) by the author(s) of [6]. This is done for illustrative purpose. The resource is freely available for use under the conditions of the Apache License 2.0.

[9] https://github.com/OlehOnyshchak/WikiImageRecommendation

7 Methodology

7.1 Methodological Approach

The hypothesis under test is ”it is possible to implement a model to recommend relevant Commons [19] images for a specific Wikipedia article using multimodal learning techniques” and implies quantitative research approach. It is aiming to discover whether state-of-the-art techniques of multimodal representation learning can solve this specific problem for Wikipedia with not worse precision.

7.2 Methods of Data Collection

Existing Wikipedia data will be used to conduct the research. More specifically, we will use a collection of featured articles[10] where each page went through thorough manual review procedure by the Wikipedia community and represent the best Wikipedia can offer. Thus, it is theoretically the best possible quality for machine learning algorithms.

7.3 Methods of Analysis

We will select candidate algorithms by analyzing recent literature surveys of a corresponding domain, and choosing the most prominent state-of-the-art approaches described there. We will also check the most cited approaches to solve a similar problem. In that way, we can ensure that all state-of-the-art methods existing in that field would be reviewed and then the most applicable would be adequately tested.

7.4 Evaluation

Since we have a labeled dataset, classical evaluation metrics would be applied here. Currently, the most appropriate approach is rank-based performance metric [26] P@K (K=1,5,10), where P is the percentage of articles for which corresponding images are found within the top $K * N_{images}$ images, where N_{images} is the number of images of this article.

When scaling up on real-world image dataset size, evaluation metrics will require additional improvements such as merging visually similar images from top-ranked matches, although it is out of the scope of testing our hypothesis.

8 Time Plan

You can find timetable of milestones in the Table 1

[10] https://en.wikipedia.org/wiki/Wikipedia:Featured articles

Table 1. Time Plan

Date	Milestone
10 Sep 2019	Kick Start
16 Sep 2019	Project Proposal’s Abstract Submission
30 Sep 2019	Project Proposal Submission
1 Nov 2019	Start of Implementation
15 Nov 2019	Finalize Approach and Solution
1 Dec 2019	Start of Evaluation
10 Dec 2019	Finalize Evaluation Planning
23 Dec 2019	Finalize Implementation
27 Dec 2019	Finalize Evaluation
31 Dec 2019	Finalize Review of Related Work
8 Jan 2020	Thesis Final Submission

9 Conclusive Remarks

The goal of the project is to research whether it is possible to implement a system, which would recommend relevant Commons [19] images for a specific Wikipedia article. Thus. we are planning to investigate the scientific landscape in that area and provide report whether it can solve our specific problem of image recommendation with Wikipedia dataset. We do not expect to create a complete end-to-end solution but rather investigate a path towards it is feasible.

References

1. Guo, W., Wang, J., Wang, J.: Deep multimodal representation learning: a survey. IEEE Access **7**, 63373–63394 (2019). doi: 10.1109/ACCESS.2019.2916887
2. McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264,** 746–748 (1976). doi:10.1038/264746a0
3. Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: a survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2018). doi: 10.1109/TPAMI.2018.2798607
4. D’mello, S.K., Kory, J.: A review and meta-analysis of multimodal affect detection systems. ACM Computing Surveys **47**(3), Article No 43 (2015). doi: 10.1145/2682899
5. Dong, J., Li, X., Snoek, C.G.M.: Predicting visual features from text for image and video caption retrieval. IEEE Transactions on Multimedia 20(12), 3377-3388 (2018). doi: 10.1109/TMM.2018.2832602

6. Dong, J., Li, X., Xu, C., Ji, S., He, Y., Yang, G., Wang, X.: Dual encoding for zero-example video retrieval. In: 2019 IEEE Conference on Computer Vision and Pattern Recognition, pp. 9346–9355. IEEE Press, New York (2019)
7. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo J.C., Hockenmaier, J., Lazebnik, S.: ”Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models. In: 2015 IEEE International Conference on Computer Vision. IEEE Press, New York (2015)
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: 25th International Conference on Neural Information Processing Systems. Volume 1, pp. 1097–1105. Curran Associates Inc., New York (2012)
9. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778. IEEE Press, New York (2016)
11. Vogel, D.R., Dickson, G.W., Lehman, J.A.: Persuasion and the role of visual presentation support: the UM/3M study. Minneapolis: Management Information Systems Research Center, School of Management, University of Minnesota (1986)
12. Jiang, Q.-Y., Li, W.-J.: Deep cross-modal hashing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3232–3240. IEEE Press, New York (2017)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
14. Pennington, Jeffrey, Richard Socher, and Christopher Manning. ”Glove: global vectors for word representation. In: 2014 ACL Conference on Empirical Methods in Natural Language Processing, pp. 1532–1543. Association for Computational Linguistics (2014)
15. Kim, Y., Jernite, Y., Sontag, D., Rush, A.M.: Character-aware neural language models. In 13th AAAI Conference on Artificial Intelligence, pp. 2741–2749 (2016)
16. Elman, J.L.: Finding structure in time. Cognitive science **14**(2), 179–221 (1990)
17. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
18. Jiang, Y.-G., Wu, Z., Wang, J., Xue, X., Chan, S.-F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(2), 352–364 (2017)
19. Wikimedia Commons, Wikimedia. https://commons.wikimedia.org/wiki/Main Page
20. Habibian, A., Mensink, T., Snoek, C.G.M.: Video2vec embeddings recognize events when examples are scarce. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(10), 2089–2103 (2016)
21. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
22. Mor, N., Wolf, L., Polyak, A., Taigman, Y.: A universal music translation network. arXiv preprintarXiv:1805.07848 (2018)
23. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164. IEEE Press, New York (2015)
24. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., Saenko, K. Translating videos to natural language using deep recurrent neural networks. arXiv preprint arXiv:1412.4729 (2014)

25. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprintarXiv:1605.05396 (2016)
26. Ma, L., Lu, Z., Shang, L., Li, H.: Multimodal convolutional neural networks for matching image andsentence. In: 2015 IEEE International Conference on Computer Vision, pp. 2623–2631. IEEE Press, New York (2015)
27. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.'A., Mikolov, T.: DeViSE: a deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems. Volume 26, pp. 2121—2129. Curran Associates, Inc. (2013)
28. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
29. Socher, Richard, Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics **2**, 207–218 (2014)
30. Pan, Y., Mei, N., Yao, N., Li, H., Rui, Y.: Jointly modeling embedding and translation to bridge video and language. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 4594–4602. IEEE Press, New York (2016)
31. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013. IEEE Press, New York (2016)