Image Recommendation System for Wikipedia Articles

Oleh Onyshchak[1]

Ukrainian Catholic University, Lviv, Ukraine
o.onyshchak@ucu.edu.ua
https://apps.ucu.edu.ua/en/

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.
TODO: finish abstract

Keywords: Multimodal Learning · Text-Image Similarity · Image Recommendation.

Table of Contents

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 kids of medial such as audio, image, video, text, and 3D modeling. That also created new challenges of efficiently processing a significant amount of recorded information, where we already have very significant achievements. However, every medium of digital information only captures some subset of available information. For example, image only captures visual appearance, while audio - the sound, just as our sight and ears do. Thus all scientific progress in processing some data medium is bounded by limitation of what that medium can capture. In other words, to digitally create a notion of dog, we cannot only have a visual representation. Just as humans, we need to combine all the information streams, which describes the same entity from different perspectives, and combine them into one comprehensive representation.

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 e.g., processing images is called unimodal, while a problem in the domain of multiple information encodings e.g., 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 will help us to increase the precision of existing data science applications and also extend its limits to more abstract problems such as not only identifying the objects on an image but understanding its value. For example[1], early researches on speech recognition showed that by involving visual modality of lips movement on top of sound modality, we get extra information which allows us to increase the quality of voice recognition task, just as it does for humans[2]

In this project, we are going to research a topic which is also a part of multimodal representation learning, specifically the "Image Recommendation System for Wikipedia Articles". That is, based on the article's text information, we need to recommend images describing the same notion. In scope of the project, we are going to combine state-of-the-art techniques of multimodal representation learning and processing of massive databases. We believe this project will be valuable from both a research and an application perspective. The former benefits from 1) evaluating the best approach to apply the state-of-the-art techniques on massive Wikimedia Commons database[TODO: link for Commons] 2) possibility to employ a set of handcrafted metadata, associated with that database.

This report is an official Project Proposal of Master's Thesis, which will define a problem, provide rigor overview of state-of-the-art approaches in problem's domain, specify goals of the project, suggest a solution approach and provide a time plan of the thesis. To fully comprehend material, the reader should be familiar with Convolutional Neural Network(CNN)[?], TODO: finish the list.

2 Problem Formulation

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

However, the quality of a significant amount of those articles is not satisfied. There are a variety of problematic points, but one of them is the absence of supporting images, which make information far easier to comprehend. Also, because finding and adding relevant images is quite a tedious job, we have plenty of pages which does not have any visualizations.

In order to facilitate the qualitative growth of Wikipedia, we are going to implement an Image Recommendation System for Wikipedia articles. So that when the user edits or creates an article, relevant images from Wikimedia Commons[?] would be suggested.

That is, having a text with wiki formatting, we need to identify the most relevant images from Wikimedia Commons dataset.

3 Data

All data is publicly available on Wikipedia. Specifically, we have more than 35 million Wikipedia pages with a fair amount of them containing some images. We also have Commons image dataset, containing more than 55 million images. That's the real-world data, where ideally our solution should be used.

But for training purposes we would only use a reliable subset of above specified data. In particular, Wikipedia has a notion of featured articles[?], which are the best articles with qualitative text and a lot of supporting visualization. In other words, it's a high quality manually created dataset of more than 5000 articles[?], each of which has multiple associated images. Although, it still required proper preprocessing and cleaning before using.

4 Related Work

TODO: include a paragraph about content based image retrieval and explain why it's not optimal?

While during the last decades there was much progress in a field of unimodal representation, research in multimodal learning was limited by simple concatenation of unimodal features[4]. However, during recent years, the scientific landscape in this domain has been 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. Thus they were also incorporated in multimodal learning. As Guo et al. suggested[1], we can categorize 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 modal to some intermediate space, from where we later generate features of another modal.

In this chapter, we will cover available techniques to extract features from text and image modalities, overview available solutions in each type of multimodal learning, and then discuss in details the most significant techniques from the Image Recommendation perspective.

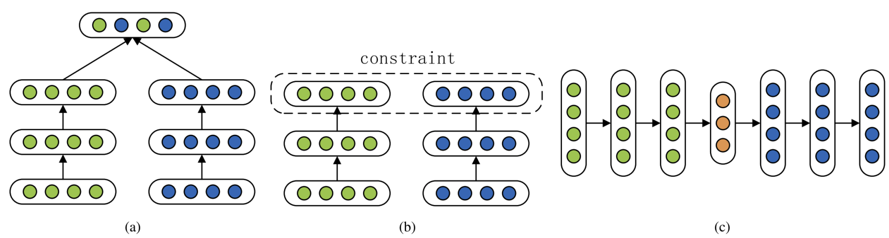

Fig. 1. Three types of frameworks about deep multimodal representation. (a) Joint representation aims to learn a shared semantic subspace.(b) Coordinated representation framework learns separated but coordinated representations for each modality under some constraints. (c) Encoder-decoder framework translates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

4.1 Unimodal Representation

Image The most popular model used in feature extraction from images are different types of Convolutional Neural Network(CNN), such as LeNet[5], AlexNet[6], GoogleNet[7], VGGNet[8] and ResNet[9]. When working with big datasets, it is preferable to use pre-trained version of chosen CNN.

Text A popular way to extract features from the text is to encode it to vector, as is done in word2vec[?] or Glove[?] algorithms. They map words into one-hot encoded vector space of language vocabulary. Although, the common problem with those approaches is when some words are not present in vocabulary or out-of-vocabulary error. However, there are also a variety of solutions to this problem, such as character embeddings[?] or character n-grams[?].

An alternative and more powerful tool for dealing with text is recurrent neural network(RNN)[?], which are 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)[?].

4.2 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 Figure[?], features of every modality are mapped into shared subspace, where the conceptions shared by modalities will be extracted and fused into a single vector[1] TODO: add figure. Up until recently, that was the primary technique in multimodal learning, where shared features were fused by concatenating them 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.

This approach was historically the first one and is still commonly applicable in video classification[?][?], event detection[?][?], sentiment analysis[?][?], and visual question answering[?]. However, its main disadvantage is neglecting the fact that different modalities have not only supplementary information, that is which show the same notion from different perspectives, but also complementary information, where one modality captures the information which another cannot. 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 supplementary sources. Because of that, much information gets lost in that shared space.

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. Thus methods from this category are not applicable to our problem

4.3 Intermediate Representation

Intermediate Representation models aim to encode features of one modal to some space, from which later features of another modal can be decoded FIGURE[?]. In that encoder-decoder framework, encoder maps source modality into a latent vector, and then, based on the vector, the decoder will generate new features of target modality[1]. To prevent the intermediate space from being related only to a source modality, during encoder-decoder training we maximize, e.g., the likelihood of target sentence given source image, so that error function employs the error of decoding. Subsequently, the generated intermediate representation tends to capture the shared semantics from both modalities.

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

The general advantage of such approach is that it is the only way to generate new features in a target domain. Thus this technique is used in Image Caption[?], Video Description[?], and Text to Image[?] generations. The disadvantages of that model are that 1) it can only encode one modality and 2) complexity of designing a feature generator should be taken into account[1]. However, since in our problem, we need to query existing information, and because intermediate space also extracts only shared subspace from two modalities, those models are not applicable for our case.

4.4 Coordinated Representation

The last type of multimodal learning is a coordinated representation. Instead of learning from a joint representation, it learns from modal-specific representations separately but with a shared constraint, which is some error function identifying cross-modal similarity/correlation. Since different modalities hold unique information about an object, that approach operates with all available knowledge. A visual explanation can be seen in Figure[?]. Regarding constraint function, a commonly used option is cross-modal similarity functions, 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 be as small as possible, while with dissimilar - as big as possible.

The cross-modal ranking is a widely used constrain, where the loss function is defined in the following way

$$loss = \sum_{i} \sum_{t}^{-} max(0, \alpha - S(i,t) + S(i,t^{-})) + \sum_{t} \sum_{i}^{-} max(0, \alpha - S(t,i) + S(t,i^{-})),$$

where (v,t) is a matching pair from the image-text match, α is margin, S is a similarity function, v^{-} is mismatching pair to t and vise versa. Frome et al. [?] 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 adjust features based on above mentioned ranked loss, though in more simplified form.

After the success of DeViSE, Kiros et al. [?] extended this model in order to create image captions. Specifically, they used the full version of cross-modal ranking as a loss function and also employed LSTM to learn text features. Socher et al. [?] also used DeViSE model to perform cross-modal retrieval of text and images. They introduced dependency trees based recursive neural network (DTRNN) to encode language modality and argued that the proposed DTRNN is robust to surface changes such as word order.

In addition to cross-modal ranking, another widely used constraint is Euclid distance. It 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[Pan et al. ??]. While for intra-modality, we want to preserve the similarity between neighborhood items, that is:

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

where m is any data point of any modality, m_i point of interest, N(m) - denotes neighborhood of m[Wang et al.?]

So, Coordinated Representation preserves all modality-specific information. It also explicitly compares features from different modalities, thus having data from one, we can identify the closest data point from another modality. Because of that properties, it is used for cross-modal retrieval[?][?], retrieval-based visual description[?], and transfer knowledge across modalities[??]. Thus it can be applied for our problem of Image Recommendation for articles.

4.5 A New Benchmark and Approach for Fine-grained Cross-media Retrieval???

Probably not

4.6 Deep Cross-Modal Hashing

Should we specify it here?

5 Solution Approach

We are planning to implement a multimodal learning system with Coordinated Representation. After dataset preprocessing, we will extract original features as 1) for text it will be some variation of word2vec algorithm, which handles out-of-vocabulary error 2) for the image we will convert all the images to some fixed size with three color channels. Then for each of them, we will create a separate deep network model and use the best performing option out of 1) LSTM[?] or DTRNN[Socher et al?] for text and 2) ResNet[?] or AlexNet[?] for images.

For loss function, we will employ cross-modal ranking function defined in DeViSE[?] and also Euclidian metrics for preserving intra-modality and inter-modality similarity structure[above?][?]

6 Methodology

6.1 Methodological Approach

The research question is defined as "What approach is the most efficient to recommend Commons image for a specific Wikipedia article" and implies Quantitative research. It is aiming to discover the best approach of solving a specific task, which will be valuable both as a knowledge and as a working system.

6.2 Methods of Data Collection

Existing Wikipedia data will be used to conduct the research. More specifically, we will use a collection of featured articles, which got the title after a thorough manual review procedure of the Wikipedia community. Those articles are carefully reviewed and present the best articles Wikipedia can offer. Thus it is theoretically the best possible quality for machine learning algorithms.

6.3 Methods of Analysis

We will select candidate algorithms by analyzing all recent literature surveys of a corresponding domain, and choosing the state-of-the-art approaches described there. We will also check the most cited approaches to solve similar topic. For all the above methods, we will try to find and adjust described algorithms, if they are publicly available. Otherwise, we will carefully implement them by following all the guidance from original papers. In that way, we can ensure that all state-of-the-art methods existing in that field would be identified and adequately tested.

7 Goals

TODO: formulate goals

8 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 Implementation
15 Nov 2019	Finalise Approach and Solution
1 Dec 2019	Start Evaluation
10 Dec 2019	Finalise Evaluation Planning
23 Dec 2019	Finalise Implementation
27 Dec 2019	Finalise Evaluation
31 Dec 2019	Finalise Review of Related Work
8 Jan 2020	Thesis Final Submission

TODO: put somewhere definition of heterogeneity gap

References

1. Guo, Wenzhong, Jianwen Wang, and Shiping Wang. "Deep Multimodal Representation Learning: A Survey." IEEE Access 7 (2019): 63373-63394.
2. McGurk, Harry, and John MacDonald. "Hearing lips and seeing voices." Nature 264.5588 (1976): 746.
3. Baltrušaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal machine learning: A survey and taxonomy." IEEE Transactions on Pattern Analysis and Machine Intelligence 41.2 (2018): 423-443.
4. D'mello, Sidney K., and Jacqueline Kory. "A review and meta-analysis of multimodal affect detection systems." ACM Computing Surveys (CSUR) 47.3 (2015): 43.
5. LeCun, Yann, et al. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86.11 (1998): 2278-2324.
6. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.
7. Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
8. Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).
9. He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
10. D'mello, Sidney K., and Jacqueline Kory. "A review and meta-analysis of multimodal affect detection systems." ACM Computing Surveys (CSUR) 47.3 (2015): 43.
11. D'mello, Sidney K., and Jacqueline Kory. "A review and meta-analysis of multimodal affect detection systems." ACM Computing Surveys (CSUR) 47.3 (2015): 43.