PSYCHOLOGY AND EDUCATION (2022) 59(2): 267-276
ISSN: 1533-6939

BREAKING NEWS ARTICLE ANNOTATION USING IMAGE AND TEXT PROCESSING

T.kathyayani ,Assistant professor: Mr.K.Narsimhulu
DEPARTMENT OF COMPUTERSCIENCE AND EGINEERING
SREYAS INSTITUTE OF ENGINEERING AND TECHNOLOGY, NAGOLE,
BANDLAGUDA, HYDERABAD , TELANGANA, INDIA
E-mail: kathyayanikatty@gmail.com

ABSTRACT:

In areas like automated captioning and image retrieval, modern techniques at the confluence of Computer Vision and Natural Language Processing have made ground-breaking advances built on recent Deep Neural Network designs. A substantial training set of photos with human annotations that characterise their visual content is required for most of these learning approaches. More complicated scenarios where written descriptions are only weakly linked to visuals are the focus of this research.

Connotative and ambiguous relationships are typically expressed in news items by the textual content, which can only be deduced via the use of connotative and ambiguous visuals. Source identification, article illustration, and article geolocation are some of the applications for an adaptable CNN architecture that shares most of its structure. A novel loss function based on Great Circle Distance is suggested for geolocation in place of Deep Canonical Correlation Analysis for article visualisation. And we've just unveiled BreakingNews, a new dataset that includes about 100,000 news pieces, each of which is enhanced with a variety of meta-data (such as GPS coordinates and user comments).

Using a variety of Deep Learning architectures, we demonstrate that this dataset is suitable for investigating all of the aforementioned challenges, and we offer a baseline performance using multiple representations of the textual and visual characteristics. As a consequence of these encouraging findings, we intend to drive advancement in the field by highlighting the limits of present state-of-the-art.

INTRODUCTION

Recently, there has been a rise in the study of how visuals and words interact. Computer vision (CV) and natural language processing (NLP) researchers have made significant strides together, resulting in breakthroughs in understanding connections between images and text. It's been claimed that tasks like automated picture labelling (Figure 8) have yielded results comparable to those of

a three-year-old (Figure 16), image retrieval (Figure 21), (Figure 31), (Figure 45), (Figure 48), and image production (Figure 6), (Figure 88). The comeback of deep learning for modelling data, made possible by the development of new parallel processors and GPU architectures, as well as

the release of new massive datasets required to train deep models with many parameters, is one of the primary reasons for the success of these techniques.

With the growing popularity of crowd sourcing methods, these datasets that include visual and verbal information have proliferated. The UIUC Pascal Sentence Dataset [66], the SBU captioned picture dataset [61], Flickr8K [31], Flickr30K [84], and MS-COCO [51] are the most well-known of these data sets. In all of these datasets, each picture is annotated with a human-written text (ranging from 1K to 1M photos) (between 1 and 5 per image). Short and accurate sentences (less than 20 words) describe the visual content of the picture and the activity going place in a concise way. Other, more complicated materials, such as illustrated news pieces, on the other hand, have gotten less attention. Current NLP and Computer Vision success stories imply that the approaches are ready for more difficult tasks than those presented by current datasets, we feel.

Computer Vision, on the other hand, has been focusing on tasks such as sentiment analysis, popularity prediction, summarization, source identification, and geolocation, among others. In this work, we provide a variety of educational approaches. A CNN architecture that can be used for source identification, article illustration, and geolocation prediction can be used for all of these tasks, but the final layers must be replaced and retrained for each specific challenge. Due to

A Long-Short-Term Network (LSN) integrates the tasks of choice generation, picture representation, and text representation (LSTM). BreakingNews, a large-scale collection of news stories with extensive meta-data, has been used to test these algorithms. Our database contains around 100K news stories, each of which is accompanied by one to three photographs and their accompanying subtitles. The articles are also enhanced with data such as photographs from Google Images, tags, superficial and deep linguistic characteristics (e.g. parts of speech, semantic subjects or the consequence of a sentiment) and more.

GPS latitude/longitude coordinates, reader comments, and an analyzer are included. Each article dates back to the beginning of 2014, and they've been sourced from a variety of newspapers and media outlets.

This dataset serves as a great baseline for advancing the development of integrated vision and language. There is no direct relationship between text and pictures inBreakingNews. This is in contrast to previous datasets, which have a more direct correlation between photos and text (see examples in Fig. 1). Since the visual-language links are more delicate, new inference tools that can reason at a higher and more abstract level are required to acquire them. Additionally, the suggested dataset is designed to handle additional difficulties, such as source/media agency recognition or the estimate of GPS locations, in addition to article illustration and picture captioning.

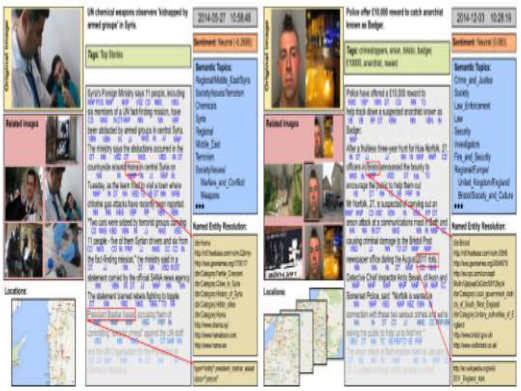


Fig. 1: Breaking News

1. LITERATURE SURVEY

Tasks Description And Related Work:

This paper's objectives and datasets are then established. This is the last step. Identifying the root cause This task involves analysing the substance of news articles in order to discover where they originally appeared in a news outlet. Additionally, it may be used in sociological research. To test our hypothesis that different news organisations have distinct visual styles, we used this image as an example.

There may be a way to deal with this problem by taking into account the political leanings of each news organisation and modelling its language, images, and theme choices. As a result, despite the fact that we don't know of any alternative solutions, there are several articles on the topic [44, [58], [72].

Text Illustration:

The goal of this challenge is to find a picture (or a limited collection of images) based on a query in a news story's text. Classifiers may be trained to represent pictures using intermediate semantic concepts, which can subsequently be given to specific keywords or multi-attribute textual descriptions [3, [4], [45], [48],[67],[17]] to solve this issue. Direct comparisons between images and texts are made possible by mapping both images and sentences to intermediate spaces where comparisons may be made. Text input for all of these systems is limited to single words or brief phrases. Many initiatives are aimed at making news stories easier to read.

With the BBC News dataset, [22] uses joint topic models for text-to-image tagging[23] and analyses its performance.

[9] assumes that the photos are accompanied by brief text descriptions and tags, which may therefore be readily matched to the text documents represented by word frequencies. There are alternative methods that use the Google search engine for article illustration (which also assumes that each picture in the database has text connected with it) and combine numerous searches derived from the article's title [50] or the narrative keywords [34]. The narrative visualising engine proposed by [38] works in a similar fashion, parsing a tale to identify important words and then using those words to choose annotated photos from a database. Text illustration from a generative viewpoint is an alternative to picture retrieval algorithms.

Static [11] and dynamic [64] scene representations may be generated by extracting object groupings from phrases and generating computer graphics. Human avatars may be animated by inferring emotions from text and then applying that

information to an avatar. Advanced language parsers have recently been utilised to autonomously create 2D [88] and 3D [6] scenes by extracting objects and their relationships from words. It has also been studied how to illustrate brief sentences, such as chat messages or tweets.

Geolocation :

In this assignment, we'll look at how to locate news stories based on text and visual clues. An strategy for geolocating online pages by identifying and disambiguating place names in the text was presented in one of the pioneering publications on a similar subject. Textual information was used to identify the most probable location of Flickr photographs that had been tagged. Many image-based algorithms, on the other hand, rely on enormous datasets of geotagged photos to geolocate generic sceneries on the global Earth scale [28], [40], [79], or to recognise places [7].

2. DESCRIPTION

Breakingnews Dataset :

When a news piece was only a few paragraphs long, we were a long way from where we are now. The papers are supplemented with images and videos, and readers are able to add their own thoughts to the original material. Understanding the impacts and interactions of various media has obvious practical uses, such as making journalists' jobs easier by proposing images from a repository automatically, or finding the most effective approach to publicise a certain piece of content. However, there are no criteria for scientific study that reflect this multimodality.

Our new dataset of news articles2 with photos, captions, geographical information, and comments will be used to assess CNN and LSTM architectures on a wide range of tasks, as a result of these considerations: This dataset is designed to serve as a reference point for exploring the present boundaries of the most recent state-of-the-art algorithms in deep learning, which have been proved to be quite successful when dealing with pictures.

Using Image Processing to Detect Text:

A classic example of a contemporary conundrum, text detection uses a combination of machine learning, computer vision, and image processing. Existing programmes, such as Google Lens and Cam Scanner, are excellent at this task. OCR (optical character recognition) algorithms are used in both of these applications to convert photos into text.

Using OCR was something I wanted to do as part of a larger effort. In spite of the many libraries currently in existence, I wanted a particular level of control and customization for my specific use case. This was also the perfect time to make the shift to OpenCV for computer vision and image processing.

Matlab's built-in image processing toolbox has been my go-to solution up until now. It's

a wonderful set of tools that makes image processing tasks a breeze. To learn more about this python library, I decided to use Open CV as it's widely used in business (and my student Matlablicence is soon ending).

Locating text is a rather straightforward process conceptually. Letters should be separated from their contrasting backdrop using a threshold filter. Each letter becomes a blob that may be segregated using pixel areas after this is accomplished. Not merely the letters themselves were of importance to me in my application (bold, italics, underline, etc.).

Thresholding is the initial stage in this process, as it is in the majority of image processing operations. First, I experimented with the global threshold approach developed by Otsu. At first, I believed this would be easier to programme, but as I learned more about OpenCV, I found that its functions enable programmers to simply swap out threshold methods.

Differences in outcomes may be seen between the two strategies. Both may be appropriate, depending on the information sought. The Otsu technique is more beneficial if the whole word is required. Adaptive thresholds are preferable if individual letters are required. Even if a whole word is needed, as will be shown in a future episode, the word may be inferred from individual letters. Adaptive filters are also easier to use if a picture isn't clean or has bad lighting.

In theory, the letters may be deciphered with the image threshold applied. Using this method directly results in a very obvious flaw. The letters themselves will be deemed to be cavities if they appear inside other letters. It's because of OpenCV's pixel area definitions that this issue arises. To discover pixel areas in Matlab, the programme "regionprops" is used. In OpenCV, however, the command "contours" is used. The command used is 'findContours()', and the arguments it is supplied may alter the outcome.

Make sure you know which curves are cavities so you don't make this mistake! It is necessary to draw the contours in the same direction as their underlying hierarchies. We're looking for a way to tell whether a contour is internal or exterior. FindContours() is called, and the second parameter is set to RETR CCOMP. Arrays of lists that explain the connection between each contour and the others will now be returned by the function. The number of the internal contour's parent, or -1 if there isn't one, may be found in one of the indices.

We don't need to worry about any shapes that have a parent with this information. Keep track of the most extreme pixel positions for each contour by iterating over them one by one until the requirement is met. In order to create the bounding boxes for each letter, we'll utilise these coordinates.

We get astonishingly clean results by following the steps outlined in this article. Letters are properly separated and delimited. The method is actually "too good" in certain areas, such as when the dots above the letter I are regarded as a single object in some instances. However, the outcomes are not fault-free either. To my eye, the letters "t" and "y" look to be too near to each other in this particular font and text editor. "rty" is a single letter in the word "liberty," which we can see in the word itself.

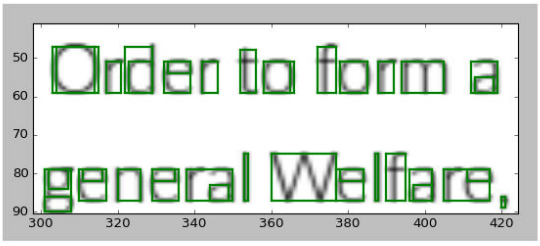
This script will eventually be applied to handwritten text. Two distinct people's handwriting was used to assess the script's accuracy and legibility. While I'll let the reader make the distinction between the two sets of handwriting, one was intended to be as neat and orderly as possible. There are certain areas of concern that will need to be addressed after this test.

A non-contiguous letter, for example, will be recognised as a collection of distinct items. The letter 'T' in the picture below serves as a good example of this. Top of the "T" is believed to be an independent item from bottom of the "T." The contrary is true, too. Script-style letters that merge numerous letters into a single sign will be read as such. 'Fo' in the word 'For' below is an example of this. Both of these problems will need creative solutions.

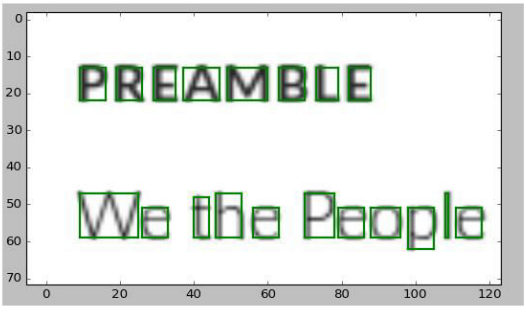
I can think of a number of things right away. We can look for letters that have a larger bounding box than other letters in its cluster, paragraph, or sentence. An overly big bounding box may signal that the letter was mistakenly mixed. The question of a proper separation emerges from here.

Breaking letters may need the determination that a bounding area does not match an existing letter in our alphabet before we may join them. Most likely, a complex neural network will be used to figure this out. Once an item is ruled out as a letter, the neighbouring bounding boxes around it are examined to see whether a more pronounced letter can be formed.

In subsequent posts, I'll discuss the next phases of the project now that the script is operating well on regular text. Letter recognition, paragraph estimate, and font and style identification are some of the other goals I've set for myself. determination.



Cavities in the letters 'g', 'o', 'e', and 'a' are selected as separate objects



Successful letter detection

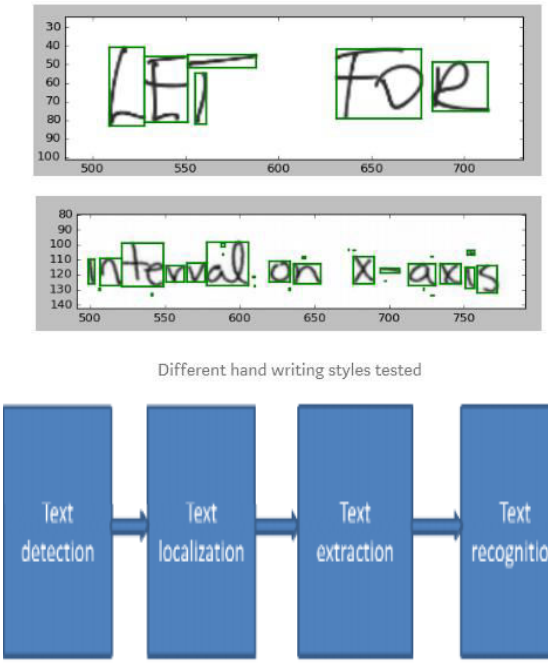


Figure 1 basic block of texture ana lysis

There were a number of research institutions in the 1960s that developed digital image processing techniques, or digital picture processing as it was more commonly known, for use in satellite imagery, wire-photo standards conversion, medical imaging, videophones, character recognition, and photograph enhancement. Many of these techniques are still in use today. [3] Early image processing aimed to enhance the picture's quality. It was designed to enhance the appearance of human beings. When a picture is processed, it goes from a low-quality input to a higher-quality output. Image enhancement, restoration, encoding, and compression are some of the most used image processing techniques. The American Jet Propulsion Laboratory was the first to successfully implement the technology (JPL).

Images from the Space Detector Ranger 7 transmitted back in 1964 were processed using methods such as geometric correction, gradation transformation, noise reduction, and so on, taking into consideration the sun's position and the lunar environment. An enormous achievement has been achieved thanks to the successful mapping of the moon's surface by a computer. With more advanced image processing applied to the spacecraft's roughly 100,000 photographs, a topographic map, a colour map and a panoramic mosaic of the moon were created. These incredible discoveries established the groundwork for a future manned mission to the moon.

As a result of the limited computer power available in those days, processing was quite expensive. In the 1970s, when cheaper computers and specialised hardware became accessible, digital image processing spread. As a result of this, dedicated problems like television standards conversion could be solved in real-time. In all but the most specialised and computer-intensive tasks, general-purpose computers have increasingly replaced specialty hardware. Digital image processing has become the most frequent way of image processing due to the availability of fast computers and signal processors in the 2000s. It is also the most cost-effective approach.

3. RESULTS AND CONCLUSION

Here, we'll go through our findings from the four experiments that we looked at. We first discuss CNN results for source detection, geolocation prediction and article illustration, followed by a discussion on LSTM and a mixed LSTM/CNNs model for caption generation. Step-by-Step Instructions Technical details, such as how the dataset was prepared and how hyper-parameters were set up for representation and learning, are discussed before we analyse the experimental findings. Considerations for the experimental set-up in relation to the dataset: Training, validation, and test sets are each comprised of 60 percent of the articles in the BreakingNews dataset. To make sure the tests were fair, we used the VGG19 characteristics and a cosine distance to make sure there was nearly no picture overlap across the sets.

Approximately 106 pairs of near-identical images were found in this sample set. Images that can be represented in text In the training set, there are 47,677 distinct tokens with a frequency of 5 or more (out of 115,427), hence the BoW representation is Db = 47,677-dimensional. In terms of captions, the size of the BoW is 13,507 (outof89,262uniqueterms). On figure out the embedding space for Word2Vec, we used the skipgram approach to the BreakingNews training set. The embedding space was Dw = 500; the window size was 30; the sampling rate was 1e-5; and the number of negative samples was 5. These hyperparameters were selected for a limited fraction depending on the illustrative goal.

We also tried out a publicly accessible Word2Vec model trained on the Google 100 billion word dataset, but the results were less impressive. As an alternative to Word2Vec, we looked into Glove [63], Doc2Vec [49], and other tools. However, our dataset shows that these embeddings perform poorly.

Word2Vec features have a median rank (MR) of 60 for the illustration task, whereas Glove features have an MR of 100 and Doc2Vec features have an MR of 200 when trained and tested on a subset of 1,000 pairings.

As a result, we concentrated on using Word2Vec capabilities in other contexts. The following is a breakdown of the article: Caffe was used to create the suggested CNN architecture, including our new GCD and CCA losses [35]. A basic learning rate of 0 was employed in all studies unless otherwise stated.

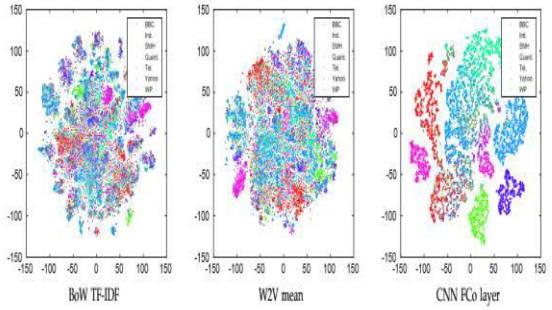


Fig. 4: Source detection: t-SNE embedding of shallow and deep features for the articles in the test set.

For every 1,000 runs, the error rate fell by an average of 0.01 percent. The momentum of the stochastic gradient descent solver was adjusted at 0.9. The weight decay regularisation value was set at 0.0005 for this analysis. Batch size m = 64 was used in

this study, which was cross-validated on the validation set to produce the findings published here. Additionally, the projection space's dimensionality was cross-validated, and the cosine distance was utilised to assess distance. When exhaustive search was not feasible, we tried as many network designs and hyper-parameter values as we could to find the best solution.

So, for example, we tested with various combinations of kernel width and stride in the convolutional layer and discovered that a combination of these values generated the greatest results. Using a kernel size of 5007 instead of 5005 decreased accuracy by 5 points, while using a kernel size of 5003 instead of 5005 decreased accuracy by 2 points. To put it another way, a stride of 2 and 5 instead of 1 reduced accuracy by a total of 1 and 5 points, respectively, in this experiment. NVIDIA Tesla K40C GPU with 12G RAM was used for all of our investigations. Consensus training is about 10 hours. The overall amount of parameters learned by the CNN varies depending on the job. Neuraltalk24 was used to implement the two LSTM topologies seen in Fig. 3-b1 and b2.

REFERENCES

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In ICML, 2013.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In International Conference on Computer Vision (ICCV), 2015.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan. Matching words and pictures. The Journal of Machine Learning Research, 3:1107–1135, 2003.

[4] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In ICCV, volume 2, pages 408–415. IEEE, 2001.

[5] L. Cao, J. Yu, J. Luo, and T. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In ACM International Conference on Multimedia, pages 125–134. ACM, 2009.

[6] A. Chang, M. Savva, and C. Manning. Interactive learning of spatial knowledge for text to 3d scene generation. Sponsor: Idibon, page 14, 2014.

[7] D. Chen, G. Baatz, K. Köser, S. Tsai, R. Vedantham, T. Pylvä, K. Roimela, X. Chen, J. Bach, M. Pollefeys, et al. City-scale landmark identification on mobile devices. In CVPR, pages 737– 744. IEEE, 2011.

[8] X. Chen and C. Zitnick. Mind's eye: A recurrent visual representation for image caption generation. In CVPR, 2015.

[9] F. Coelho and C. Ribeiro. Image abstraction in crossmedia retrieval for text illustration. Lecture Notes in Computer Science, 7224 LNCS:329–339, 2012.

[10] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. JMLR, 12(08):2493–2537, 2011.

[11] B. Coyne and R. Sproat. Wordseye: an automatic text-to-scene conversion system.

In Conference on Computer Graphics and Interactive Techniques, pages 487–496. ACM, 2001. [12] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In International Conference on World Wide Web, pages 761–770. ACM, 2009.

[13] J. Deng, W. Dong, R. Socher, K. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR, pages 248– 255, 2009.

[18] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010. [19] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. Platt, C. Zitnick, and G. Zweig. From captions to visual concepts and back. In CVPR, 2015. [20] H. Fang, S. Gupta, F. Iandola, R. Srivastava, L. Deng, and P. Dollár others. From captions to visual concepts and back. In CVPR, pages 1473–1482, 2015.

[21] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, pages 15–29. Springer, 2010.

[22] Y. Feng and M. Lapata. Topic Models for Image Annotation and Text Illustration. Conference of the North American Chapter of the ACL: Human Language Technologies, (June):831–839, 2010.