

Korean Tourist Spot Multi-Modal Dataset for Deep Learning Applications

Changhoon Jeong ¹, Sung-Eun Jang ², Sanghyuck Na ¹ and Juntae Kim ^{1,*}

¹ Department of Computer Science and Engineering, Dongguk University, Seoul 04620, Korea; chjeong@dongguk.edu (C.J.); shna@dongguk.edu (S.N.)

² Department of Intelligence, Dongguk University, Seoul 04620, Korea; jse9512@dongguk.edu

* Correspondence: jkim@dongguk.edu

Received: 6 September 2019; Accepted: 8 October 2019; Published: 12 October 2019



Abstract: Recently, deep learning-based methods for solving multi-modal tasks such as image captioning, multi-modal classification, and cross-modal retrieval have attracted much attention. To apply deep learning for such tasks, large amounts of data are needed for training. However, although there are several Korean single-modal datasets, there are not enough Korean multi-modal datasets. In this paper, we introduce a KTS (Korean tourist spot) dataset for Korean multi-modal deep-learning research. The KTS dataset has four modalities (image, text, hashtags, and likes) and consists of 10 classes related to Korean tourist spots. All data were extracted from Instagram and preprocessed. We performed two experiments, image classification and image captioning with the dataset, and they showed appropriate results. We hope that many researchers will use this dataset for multi-modal deep-learning research.

Dataset: <https://doi.org/10.5281/zenodo.3381859>

Dataset License: MIT License

Keywords: social network service; Korean tourist spot; deep learning; multi-modal learning; Korean text

1. Summary

Recently, as deep learning has emerged as a big topic in various fields, the importance of datasets is increasing [1]. In the field of computer vision, there are many single modal datasets, such as CIFAR-10, CIFAR-100, ImageNet, and MNIST [2–4]. Also, in the field of natural language processing, various deep-learning models could be studied due to datasets like the IMDB (Internet Movie Database) review dataset and Stanford Sentiment Treebank dataset [5,6].

One of the new areas in deep learning is multi-modal deep learning [7]. A multi-modal dataset is a pair of single modal data, such as image, text, audio, and video, and the methods of integrating them provide us important insights for solving real-world problems. By building multi-modal datasets and using them for various applications, such as image captioning, multi-modal classification, and cross-modal retrieval, we can solve a wider variety of real-world problems. The various datasets, including MSCOCO, NUS-wide, Yelp, XMedia, and Flickr30k are provided as open-source materials for multi-modal deep learning research [8–12]. However, there are not enough multi-modal datasets available in Korean, even though the Korean language is structurally more complex and challenging to preprocess than English.

Therefore, we built a Korean tourist spot (KTS) dataset by collecting multi-modal data related to Korea's tourist spot domain and integrating them into a single dataset. All data are collected from a

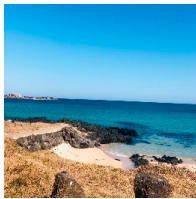
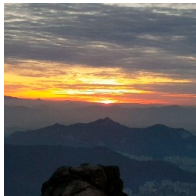
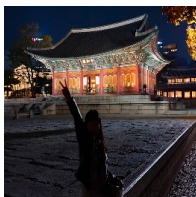

social network service (Instagram), and each data instance consists of an image, text, hashtags, and likes of a post. The dataset has 10 classes of 1000 instances each (total of 10,000 instances).

The KTS dataset can be used not only to perform simple image classification or sentiment analysis for text data, but also to perform various multi-modal tasks, such as image captioning, multi-modal classification, and recommendation-system simulation. In the experimental part of this study (see Section 4), we conducted two simple experiments, image classification, and image captioning. The experiment results show that meaningful performances can be achieved with this dataset in general deep learning.

2. Data Description

Instagram is a photo- or video-sharing social networking service owned by Facebook [13]. Instagram-user posts include images, texts, hashtags, likes, user ids, and other users' comments. We extracted the images, texts, hashtags, and likes from the above elements, using a web scraping technique, and sensitive information (e.g., user ID and URLs of the post) was removed. The KTS dataset has 10,000 instances collected from the posts related to Korean tourist spots uploaded to Instagram. Table 1 shows a schematic of the dataset. For example, the first row shows an instance of “beach” sub-class that contains an image, texts “is this real life?? Real-time Udo. Jeju-do is awesome. Sea color is also beautiful”, hashtags meaning “#travel”, “#Udo (island)”, and “#Hagosudong beach”, and the likes count.

Table 1. The overall schematic of Korean tourist spot (KTS) dataset.

Subclass	Image	Text	Hashtag	Likes
beach (nature-scene)		"이 풍경 실화냐?? 실시간 우도. 제주도 너무 예쁘다. 바다 빛깔도 예술이야."	"#여행" "#우도" "#하고수동해수욕장"	27
mountain (nature-scene)		"죽음의 등산. 오랜만에 등산에 실례서 시작한 등산. 김밥도 싸서 올라갔는데 이게 뭐야.. 너무 힘들어 계단이며 바위며.. 6시간동안 등산하고 몸살났음."	"#등산" "#관악산" "#김밥" "#몸살" "#알배김"	27
palace (person-made)		"2시간 해설 들으면서 덕수궁 탐방, 너무 내 취향이야"	"#덕수궁" "#덕수궁궁궐야행" "#한양길라잡이" "#주말나들이" "#궁궐"	32
tower (person-made)		"대부도에 있는 높은 빌딩 이름하여 시화나래휴게소 달전망대 이름도 참 길다...바닥이 유리로 만들어져서 다 보임! 고소공포증 있는 사람에게는 비추천"	"#대부도" "#대부도여행지" "#대부도전망대" "#시화나래전망대" "#시화호" "#시화방조제"	54

2.1. Class Structure

Table 2 shows the class structure. The super-class is divided into “person-made” and “nature-scene” for the tourist spot domain, and each has five sub-classes: amusement park, palace, park, restaurant,

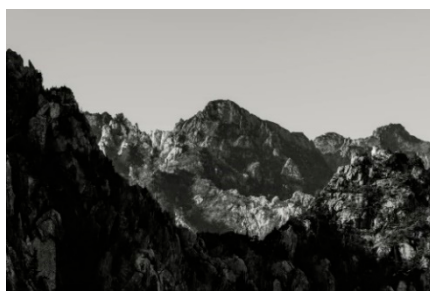
and tower for person-made and beach, cave, island, lake, and mountain for nature-scene. There are 1000 instances (image, text, hashtags, and likes) for each sub-class.

Table 2. A class structure.

Super-Class	Person-Made					Nature-Scene				
sub-class	amusement park	palace	park	restaurant	tower	beach	cave	island	lake	mountain

2.2. Data Structure

There is a total version and split version of the dataset. The split version is provided in a 7:1:2 ratio, divided by train, valid, and test. The total version contains all the data to allow users to divide the dataset to the desired ratio. We also provide the code to split the dataset. Each of the four folders (total, train, valid, and test) consists of two super-classes and ten sub-classes, like the class structure presented in Table 2. Each class has an image folder that contains image data and a json file that includes text, hashtags, and likes in json format. Figure 1 shows an example of the data structure of the KTS dataset. For instance, the first picture shows the 2nd image data for the mountain class in the total folder. The json file contains data such as text, likes, etc., which form a pair for this image. The “text” refers to the texts that are extracted from the posts and the “label” refers to the name of the sub-class of instance. The “hashtag” refers the hashtags of post, and the “img_name” refers to image file names that are stored in the image folder. “likes” refers to the numbers of likes of the post at the time of data collection. The data structure allows users to load a json file and an image file together.



KTS/total/nature-scene/mountain/images/2.jpg

```
{
  "text": "대호가 살고있다...",
  "label": "mountain",
  "hashtag": [
    "#설악산",
    "#북백사진",
    "#풍경사진",
    "#강원도",
    "#picture",
    "#photography",
    "#sel2470gm",
    "#photo",
    "#sonya9"
  ],
  "img_name": "2",
  "likes": "39"
},
```

KTS/total/nature-scene/mountain/mountain.json



KTS/total/person-made/palace/images/27.jpg

(a)

```
{
  "img_name": "27",
  "label": "palace",
  "likes": "11",
  "text": "남자친구랑 덕수궁 돌담길 걸으면 헤어진다며?",
  "hashtag": [
    "#덕수궁",
    "#중화전",
    "#덕수궁돌담길",
    "#반되니까",
    "#또다른느낌",
    "#멋지다",
    "#조명별",
    "#내가왕이다",
    "#음직하네",
    "#장군인줄"
  ]
},
```

KTS/total/person-made/palace/palace.json

(b)

Figure 1. An example of the data structure.

2.2.1. Images

Every image in sub-classes is saved in jpg format and numbered from 1 to 1000 for each data instance. It is composed of the images that can represent each class well, and image classification tasks can be performed using only this image data. We conducted some experiments, and the results are described in Section 4.1. Since the images posted on Instagram are stored without any modification, the image sizes are various. Because the images are related to the tourist spots, there are many components

in images, including people, but there is no image that clearly shows a face to be recognized as a specific person. Figure 2 shows the examples of images.

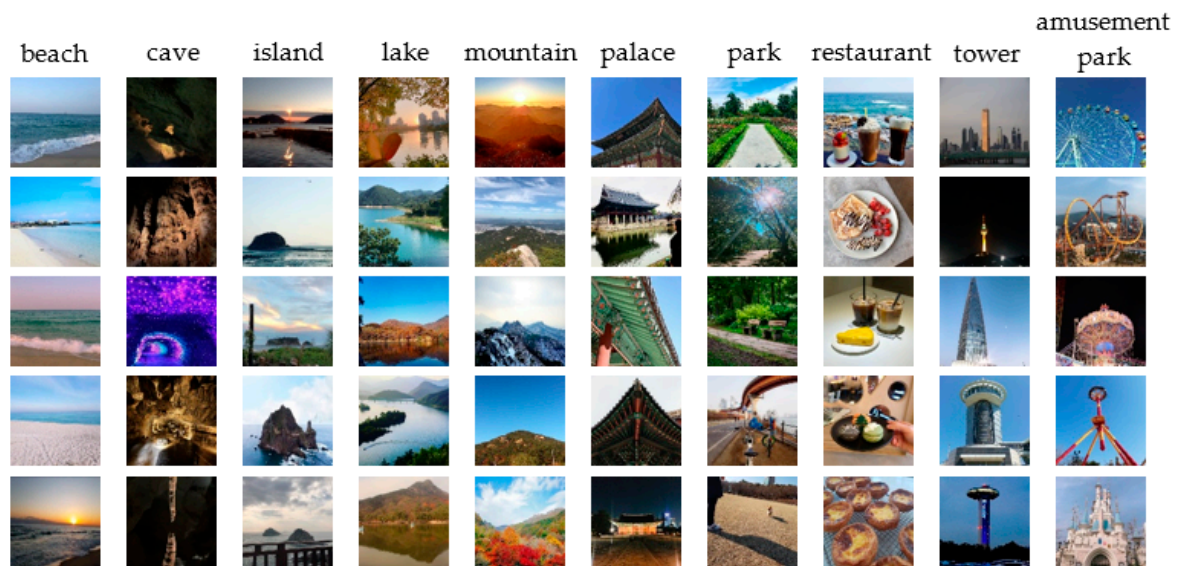


Figure 2. The examples of images.

2.2.2. Texts

Since the texts are extracted from Instagram posts written by Koreans, they are mostly in Korean, and some special symbols (., !, ?, @, +, *, etc.) and numbers are also included. The personal information of the user and others in the post is not included. The names of persons which must appear in the context are replaced as “칠수” for the male and “영희” for the female, which are very common, names like “Jack” and “Jill” in English. In addition, grammar and spelling errors in Korean commonly found in Instagram are reflected without correction. Table 3 shows the examples of texts. As an example, the text for restaurant sub-class is “Crispy egg tart is delicious.”

Table 3. The examples of texts.

Coarse Label	Fine Label	Example
person-made	amusement park	진짜 재밌지만 떨어지는 구간 외에는 그저 추웠다고 한다
	palace	새해 첫날부터 열심히 역사공부, 아픈 역사를 다시 느끼게 해준 설명
	park	시월의 어느 멋진 날에 용산가족공원. 붙잡고 놓아주기 싫은 가을풍경
	restaurant	바삭한 에그타르트 맛있어
nature-scene	tower	야간 드라이브로 간 빛가람전망대! 맛있는 커피도 마셨다
	beach	배 한 척 떠 있는 바다. 이마저도 멋있네
	cave	감성적인 느낌이 물씬 나는 울진 성류굴에서 여행 기분을 만끽해본다
	island	범섬의 아침과 한라산입니다 제주는 오늘도 따뜻합니다
	lake	경치는 좋았지만 아직은 좀 추웠던... 오가는 것도 고생이다
	mountain	아쉽게도 정상에선 설경을 볼 순 없었지만 소문대로 칼바람

2.2.3. Hashtags and Likes

The hashtags are in Korean and English, and they are closely related to the content of the posts. “likes” refers to the number of likes of posts at the time of data collection. Due to the nature of the Instagram, the number of hashtags in a post varies from 0 to tens, and the number of likes also varies from 0 to thousands. Figure 3 shows the distribution of hashtags and likes in the dataset.

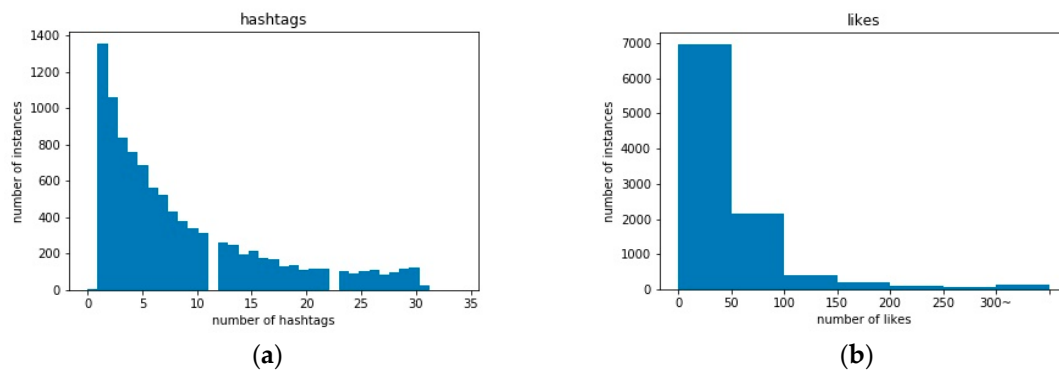


Figure 3. The distribution of hashtags (a) and likes (b). The x-axis represents the number of hashtags, and likes and the y-axis represents the number of instances, respectively.

We trained and visualized the distributed representation of hashtag words using Word2Vec model [14]. Figure 4a shows the two-dimensional visualization of the hashtag word vectors learned for each of the nature-scene and person-made classes. The Word2Vec visualization shows that the hashtags have somewhat separate distributions for each class. We also visualized the frequency of hashtags using the WordClouds application [15], to verify if hashtags consist of words that are appropriate for each super-class. Figures 4b and 4c show the results of WordClouds, visualize hashtags that are frequent in person-made and nature-scene classes, respectively. The larger the word size, the more words are included in the dataset. For example, in Figure 4b, there are many words related to person-made classes, such as “서울 (Seoul)”, “궁궐 (palace)”, and “식당 (restaurant)”, and in Figure 4c, there are many words related to nature-scene classes, such as “바다 (sea)”, “제주 (Jeju)”, and “풍경 (scene)”.



Figure 4. A 2D visualization of hashtag vectors trained by Word2Vec (a), and a WordClouds of hashtags in person-made class (b) and nature-scene class (c).

3. Methods

As described in Table 2, we divided the class structure into super-classes (person-made, nature-scene) and five sub-classes for each super-class. In order to deal with tourist spot domain, we designed a collecting and preprocessing procedure so that sub-class data can cover tourist spot domain, as well. Data collection and data preprocessing took two months, from January to February in 2019, and were implemented through Python [16]. In addition, we used the BeautifulSoup package [17] for data collection.

3.1. Data Collection

The posts were obtained from Instagram, using queries related to the sub-classes. For example, to collect data for the sub-class “beach”, we used specific tourist-spot beach names, such as “경포해수욕장 (Gyeongpo beach)” and “해운대 (Hawoondae)”, as queries. Then, we extracted the post’s information from the HTML code of the post. After extracting the information from the HTML code, the images were saved as jpg files, and the texts, hashtags, likes, user ids, comments, and post URLs

were saved as json files. If multiple images were registered in the post, only the first image shown at the front was collected.

3.2. Data Preprocessing

3.2.1. Images

We selected only the instances of images related to Korean tourist spots and sub-classes from all collected data. We also excluded instances of images that contain sensitive information, such as a face that is recognizable as a specific person or personal information such as a phone number. In consideration of the difficulty of utilization, instances of images with framed decorations or white or black solid margins on the edges of the images were removed. Instances with sensitive parts, such as copyrighted images with logos, were also excluded. All these processes were done manually, and the images were not resized or cropped.

3.2.2. Texts and Hashtags

In the case of texts and hashtags, they were refined for 10,000 instances acquired through the images preprocess. We removed emojis, except for some special symbols, and characters such as Chinese or Japanese were removed or translated. In addition, like the images, information such as names, user ids, and phone numbers, that can identify an individual was also removed or modified. Also, comments and post URLs were removed from the instances.

4. Experiments

We conducted two simple experiments to verify that all the data were collected appropriately. The first experiment was the image classification using some of the recent Convolutional Neural Networks [18], which is described in Section 4.1. In the second experiment, a simple image-captioning task [19] is preformed using images and texts, which is described in Section 4.2.

4.1. Image Classification Using DCNN (Deep Convolutional Neural Networks)

To verify that the images were adequately collected, we fine-tuned several deep CNN models using the images. The selected CNN models were VGG16 [20], ResNet18 [21], and DenseNet121 [22], and the hyperparameters were set to the optimizer as momentum SGD [23], the learning rate as 0.001 (scheduling), and batch size as 4. All experiments were performed in the same setting. Since there were also 10 classes, similarly, we fine-tuned CIFAR-10 dataset [2] and compared their performances. Table 4 shows the Top-1 accuracies for each model. As described in Table 4, the deep CNN models show good performances for the KTS dataset, like the CIFAR-10. In fact, we can achieve higher performance if we fine-tune VGG or ResNet using CIFAR-10 properly, but in this experiment, we trained all the models with the same hyperparameters setting.

Table 4. Top-1 classification accuracies of deep CNN models.

Model	CIFAR-10	KTS (Ours)
VGG16	0.8668	0.9155
ResNet18	0.8724	0.9025
DenseNet121	0.8823	0.9160

4.2. Image Captioning Using CNN and LSTM (Long Shot Term Memory)

The images of KTS dataset can be used in a single modal experiment like an image classification, but, basically, the dataset is a multi-modal in which multiple single-modal data make up one instance, so it can be used on many complex tasks. We conducted a simple image captioning using images and texts in the dataset. The goal of image captioning is to convert a given image into a text description.

Usually, an encoder–decoder framework is used for this task. The encoder uses a CNN model, and the decoder uses a recurrent model, such as LSTM or GRU [24,25]. In this experiment, we used pre-trained DenseNet152 encoder and LSTM decoder [22,24], and the hyperparameters were set to the Adam [26], the learning rate as 0.001, and batch size as 128. Table 5 shows several samples of the test results for image captioning.

Table 5. Sample test results for image captioning.

Sub-Class	Beach	Mountain	Palace	Amusement Park
image				
ground truth	날씨 좋다. 경포대	야호! 이 풍경 보려고 등산하지	주말나들이 너무 좋다. 경회루도 너무 이쁘다	너랑 놀이공원 가고 싶어
caption (prediction)	겨울바다 너무 좋은 것. 여행은 혼자다.	어제 도봉산을 올랐어요! 날씨가 많이 춥지 않은 덕분에 즐거운 산행을 했지요!	경회루 좋아 멋있어	밤되니 이쁘다 회전목마 우리 애기랑 사진한 장 찍기가 하늘의 별따기.

In Table 5, the image of the first column is from the sub-class “beach”. The ground truth text for this is described in the second row, and it translates to “The weather is nice, Gyeongpodae”. The caption generated by the neural network for this test data is described in the third row. It translates into English as “The winter sea is so good. I like traveling alone”.

The vocabulary we built for image captioning has 23,147 words. In this experiment, we trained the model using cross-entropy loss to reduce the differences between target sentences and prediction sentences. Also, we measured a perplexity [27], which is a simple way of evaluating language models. Figure 5 shows that training loss and perplexity are reduced for each epoch for the image-captioning experiment.

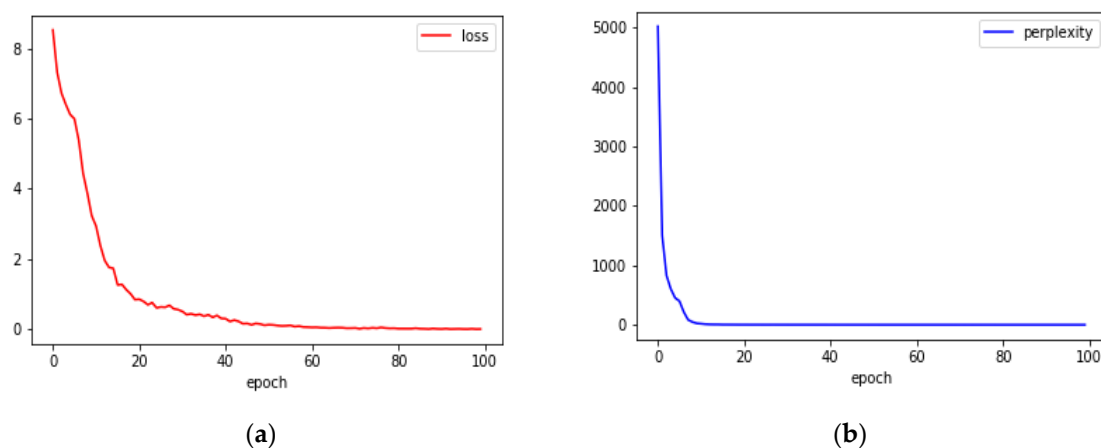


Figure 5. The loss (a) and perplexity (b) graphs for image captioning.

5. Conclusions and Future Work

We created the KTS dataset for multi-modal tasks in the field of machine learning. The KTS dataset was designed for research with Korean texts, and it consists of images, texts, hashtags, and likes of Instagram posts on Korean tourist spots. The dataset can be used to perform a variety of multi-modal tasks, such as image captioning, multi-modal classification, and recommendation-system simulation, as well as single-modal tasks, such as image classification and sentiment analysis. We provide not only the dataset but also the code for loading and preprocessing the data (see <https://github.com/DGU-AI-LAB/Korean-Tourist-Spot-Dataset>).

In the future, we plan to increase the size of the dataset by adding more classes and other modality data, such as audio and video. We also plan to expand the dataset by providing English texts, as well as Korean texts, so that many researchers can use it.

Author Contributions: All authors contributed equally to this work and have read and approved the final manuscript.

Funding: This research was funded by the Ministry of Science, ICT, Republic of Korea, grant number (NRF-2017M3C4A7083279).

Acknowledgments: This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT (NRF-2017M3C4A7083279).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pouyanfar, S.; Sadiq, S.; Yan, Y.; Tian, H.; Tao, Y.; Reyes, M.P.; Shyu, M.-L.; Chen, S.-C.; Iyengar, S.S. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.* **2018**, *51*, 92. [CrossRef]
2. Krizhevsky, A.; Nair, V.; Hinton, G. The CIFAR-10 Dataset. Available online: <https://www.cs.toronto.edu/~kriz/cifar.html> (accessed on 11 October 2019).
3. Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; Li, F.-F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.
4. LeCun, Y.; Cortes, C.; Burges, C.J.C. MNIST Handwritten Digit Database. Available online: <http://yann.lecun.com/exdb/mnist/> (accessed on 11 October 2019).
5. Maas, A.L.; Daly, R.E.; Pham, P.T.; Huang, D.; Ng, A.Y.; Potts, C. Learning word vectors for sentiment analysis. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Portland, OR, USA, 19–24 June 2011; pp. 142–150.
6. Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C.D.; Ng, A.; Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Seattle, WA, USA, 18–21 October 2013.
7. Zhang, Q.C.; Yang, L.T.; Chen, Z.K.; Li, P. A survey on deep learning for big data. *Inf. Fusion* **2018**, *42*, 146–157. [CrossRef]
8. Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014.
9. Chua, T.-S.; Tang, J.H.; Hong, R.C.; Li, H.J.; Luo, Z.P.; Zheng, Y.T. NUS-WIDE: A real-world web image database from National University of Singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval, Fira, Greece, 8–10 July 2009.
10. Yelp Dataset Challenge. 2014. Available online: <https://www.yelp.com/dataset/challenge> (accessed on 11 October 2019).
11. Peng, Y.X.; Huang, X.; Zhao, Y.Z. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Trans. Circuits Syst. Video. Technol.* **2017**, *28*, 2372–2385. [CrossRef]

12. Plummer, B.A.; Wang, L.W.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
13. Instagram. 2010. Available online: <https://www.instagram.com> (accessed on 11 October 2019).
14. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, CA, USA, 5–10 December 2013.
15. WordClouds. Available online: <https://www.wordclouds.com/> (accessed on 11 October 2019).
16. Python. 1991. Available online: <https://www.python.org> (accessed on 11 October 2019).
17. BeautifulSoup Documentation. Available online: <https://www.crummy.com/software/BeautifulSoup/bs4/doc/> (accessed on 11 October 2019).
18. Rawat, W.; Wang, Z.H. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef] [PubMed]
19. Hossain, M.D.; Sohel, F.; Shiratuddin, M.F.; Laga, H. A comprehensive survey of deep learning for image captioning. *ACM Comput. Surv.* **2019**, *51*, 118. [CrossRef]
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556. Available online: <https://arxiv.org/abs/1409.1556> (accessed on 10 April 2015).
21. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016.
22. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
23. Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the importance of initialization and momentum in deep learning. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013.
24. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
25. Chung, J.Y.; Gulcehre, C.; Cho, K.Y.; Bengio, Y.S. Gated feedback recurrent neural networks. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015.
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2017**, arXiv:1412.6980. Available online: <https://arxiv.org/abs/1412.6980> (accessed on 30 January 2017).
27. Brown, P.F.; Pietra, V.J.D.; Mercer, R.L.; Pietra, S.A.D.; Lai, J.C. An estimate of an upper bound for the entropy of English. *Comput. Linguist.* **1992**, *18*, 31–40.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).