International Journal of Advanced Culture Technology Vol.7 No.4 295-301 (2019)
DOI 10.17703/IJACT.2019.7.4.295

IJACT 19-12-36

Deep Learning Research Trend Analysis using Text Mining

Jee Young Lee

Department of Software, SeoKyeong University, Korea
J.Ann.LEE@skuniv.ac.kr

Abstract

Since the third artificial intelligence boom was triggered by deep learning, it has been 10 years. It is time to analyze and discuss the research trends of deep learning for the stable development of AI. In this regard, this study systematically analyzes the trends of research on deep learning over the past 10 years. We collected research literature on deep learning and performed LDA based topic modeling analysis. We analyzed trends by topic over 10 years. We have also identified differences among the major research countries, China, the United States, South Korea, and United Kingdom. The results of this study will provide insights into research direction on deep learning in the future, and provide implications for the stable development strategy of deep learning.

Keywords: *Artificial intelligent, Deep learning, Text mining, Topic modeling, Research trend analysis*

1. INTRODUCTION

Artificial intelligence (AI) has repeated expectations and disappointments from research institutions and industry since professor John McCarthy first mentioned it in 1956 [1]. As the 2010s began, the advances in processor computing power, network advancement, and the explosion of big data caused the third AI boom. The driving force behind the third AI boom is the deep learning which evolution of artificial neural network. Now, ten years later, there are concerns that deep learning's performance has not met expectations. To develop the third AI boom triggered by deep learning, it is necessary to analyze the trends of deep learning research over the last decade and discuss future research directions and strategies. In this regard, this study systematically analyzes the trends of deep learning research during the past 10 years, from 2010 to 2019. To this end, we analyzed research topics and analyzed the differences between China, the United States, Korea, and the United Kingdom. The results of this study will provide insight into the direction of future deep learning research.

2. RELATED WORKS

2.1 Deep learning

Deep learning is an evolution of artificial neural network. As shown in Figure 2, there is a hidden layer between the input layer and the output layer, and it is a predictive analysis method that increases the accuracy by adjusting the weight of the edge connecting the nodes of the layer by learning [1]. Deep learning allows computational models composed of multiple processing layers to learn data representations with multiple

Manuscript received: October 11, 2019 / revised: October 28, 2019 / Accepted: November 10, 2019
Corresponding Author: J.Ann.LEE@skuniv.ac.kr
Tel:+82-02-940-7520, Fax: +82-02-940-7521
Author's affiliation
Adjunct Professor, Dept. of Software, SeoKyeong University, Korea

levels.

Deep learning has been receiving continuous attention with outstanding achievements in the fields of image, voice, and natural language processing by partially eliminating the existing constraints caused by neural network learning.

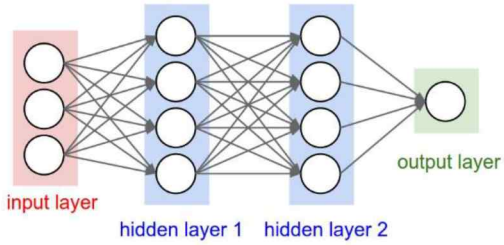


Figure 1. Deep learning architecture

2.2 Text mining and LDA based topic modeling

Text mining is accomplished through a series of text analysis and processing that extracts meaningful information using natural language processing (NLP) in unstructured text. Topic modeling is a text mining method that identifies and classifies topics that are latent in a document [2,3]. Topic modeling inferences topics by clustering words with similar meanings to find topics in a large unstructured document set [4,5].

The most popular topic m6deling technique is the potential probability estimation technique developed by Blei et al. [5]. The LDA assumes that a single document contains multiple topics or that multiple documents can share a common theme. The LDA calculate the probability that individual documents and words will be included in a particular topic and the probability that individual words derived from the entire document will be included in a particular topic.

Figure 2 shows the LDA as a graphical model [4,5]. N is the number of words per document and indicates the length of the document. D represents a set of documents. $W_{d,n}$ represents the n-th word of d document, which is determined by Z and β. Z represents a topic by word with a per-word topic assignment and is determined by the value θ_d, which is the subject distribution of the document. θ is the weight of each subject obtained using the parameter α and the Dirichlet distribution.

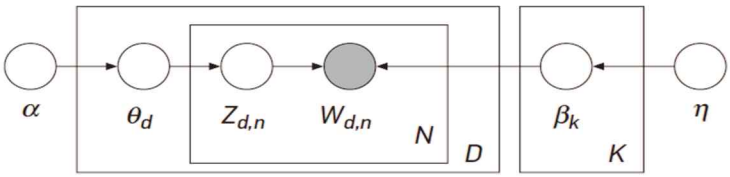


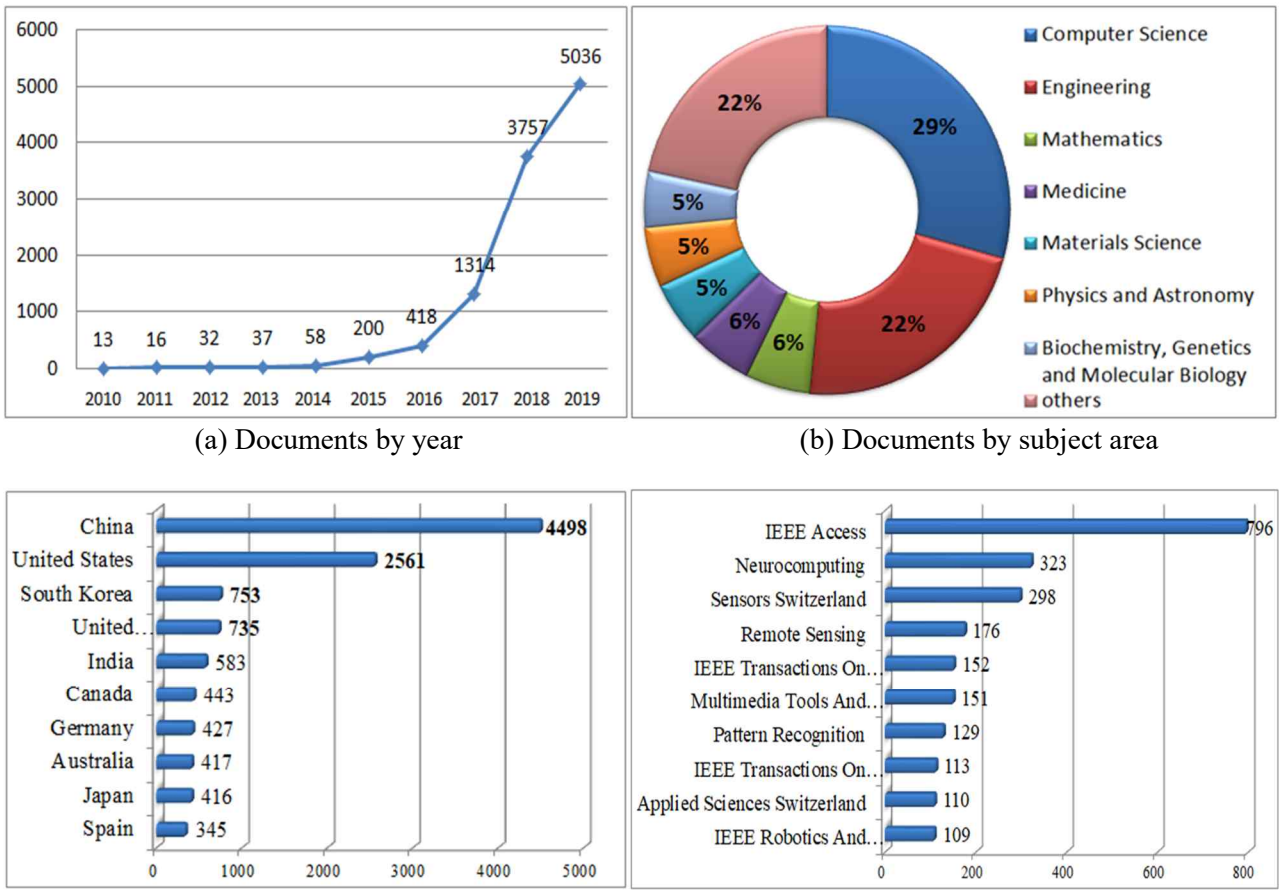
Figure 2. LDA graphical model for Topic modeling [4]

3. RESEARCH METHODS

3.1 Data collection

In this study, we analyze research topics and trend of deep learning using abstract texts of research articles. We collected 10,881 data for 10 years from 2010 to September 2019 on journal articles and conference papers that include "deep learning" as keywords in SCOPUS.

The collected data is shown in Figure 3. (a) shows documents by year, which has surged in recent three years. (b) shows the percentage by subject area, with computer science 29%, engineering 22%, mathematics 6%, and medicine 6%. In (c) documents by country, China is the most followed by the United States, South Korea and the United Kingdom. (d) shows the journals in which the articles were published.



(a) Documents by year (b) Documents by subject area

(c) Documents by country (d) Documents by source

Figure 3. Data collection results

3.2 Data preprocessing

We have preprocessed on abstracts of collected text data for analysis. We converted the words in the document into lowercase letters, and then performed tokenization to separate them into words. We also removed stop words, which are not necessary for analysis. Next, we performed lemmatization to extract the lemma for words used in various forms in the sentence. We used python 3.7, NLTK, and machine learning library scikit-learn for preprocessing and text mining.

3.3 Topic modeling

To perform LDA based topic modeling, you must determine the number of topics that are hyper parameters. In this study, topic coherence, which is a technique to evaluate the performance of topic modeling, was used to derive the optimal number of topics.

Topic coherence is a performance evaluation method proposed by Newman et al. [7]. The better the topic modeling, the more semantically similar words are gathered within the topic, which increases the coherence of the topic. The similarity calculation between words uses pointwise mutual information (PMI) index. The higher the PMI value, the higher the relevance between words [7].

In Equation (1), PMI (w_i, w_j) is calculated by using the probability of word w_i, probability of word w_j and the probability of a word pair (w_i, w_j) appearing at the same time.

$$\mathrm{PMI}(w_i, w_j) = \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \quad (1)$$

In this study, the topic coherence was calculated while the number of topics was varied from 2 to 40, and the number of topics with the highest coherence score was found. As a result, the number of topics in the LDA model was set to 8.

3.4 Normalized topic frequency

We calculated the topic frequency according to the year to analyze the trend of deep learning research topics over time. The LDA algorithm assumed that one document can contain multiple topics or that multiple documents can share a common topic. Therefore, all of the topic proportions assigned to one document should be considered and analyzed. We performed the normalization according to the topic proportions to identify the topic frequency by year. The equation for the normalized topic frequency is shown below [8].

Let $D_t = \{d_t^1, d_t^2, \ldots, d_t^{n_t}\}$ be the set of documents at time index (i.e., year) t, d_t^j be the jth document in this set ($\forall j$ = 1, …, n_t), where n_t represents the total number of documents in D_t. Correspondingly, the topic frequency for topic ID i at time index t is defined according to Equation (2).

$$\mathrm{Tf}_{i,t} = \frac{\sum_{j=1}^{n_t} \theta_{i,d_t^j}}{n_t} \quad (\forall i = 1, \ldots, K;\ \forall t = 1, \ldots, T) \qquad (2)$$

In this study, K is 8, so the topic ID i is a value from 0 to 7. Given that the data period is from 2010 to 2019, the time index t ranges from 1 = 2010 to T = 2019. The topic frequency graph normalized over time is shown in Figure 5.

4. RESULTS AND DISCUSSION

4.1 Results of topic analysis on deep learning

The topic related to the outcomes of the LDA algorithm is the distribution of words. We can derive proportions of words which are distributed according to the topic. The results of the topic modeling based on LDA are shown in Table 1. The first column shows the topic IDs from 0 to 7. In the second column, we list 5 words distributed in the topic, in a descending order starting from the highest probability of allocation. The third column shows the top two representative documents based on topic proportions. We marked the topic proportions in parentheses. The last column is a label which symbolizes the topic. The LDA algorithm does not automatically generate labels for topics. Therefore, we discussed and decided the label of the topic with three deep learning experts, referring to the documents with the largest topic proportions and word distributions on the specific topic.

Table 1. Topic related to deep learning extracted using LDA

Topic ID	Top 5 terms	Top 2 representative documents	Topic label
0	layer, signal, architecture, structure, parameter	d201 (0.98) d1024 (0.98)	Architecture
1	segmentation, dataset, label, sample, domain	d7948 (0.97) d2276 (0.95)	Domain dataset
2	recognition, human, task, visual, challenge	d10349 (0.86) d7159 (0.82)	Human recognition
3	image, detection, object, quality, map	d3051 (0.97) d907 (0.96)	Image detection
4	model, prediction, predict, perfor	d7345 (0.99)	Prediction model

	mance, term	d112 (0.98)	
5	technique, application, algorithm, design, problem	d8110 (0.98) d1963(0.97)	Application algorithm
6	feature, information, face, video, extract	d5353 (0.98) d6028 (0.98)	Feature extract
7	classification, convolutional, patient, class, diagnosis	d2259 (0.98) d3040 (0.98)	Patient diagnosis

Figure 4 shows the number of documents by topic for four countries. We analyzed four major countries, China, USA, Korea, and UK, identified in data collection. As shown in Figure 3 (c), most research has been conducted in China. In particular, Topic 6 shows that China's research is relatively more than other countries. In Topic 5, more research was being done in the United States.

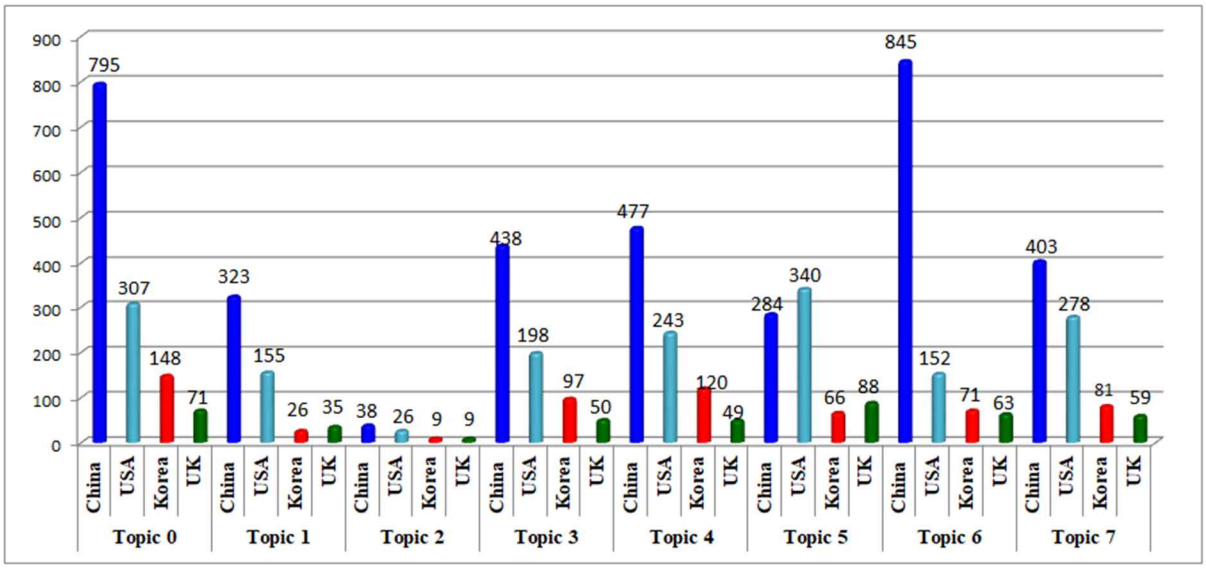
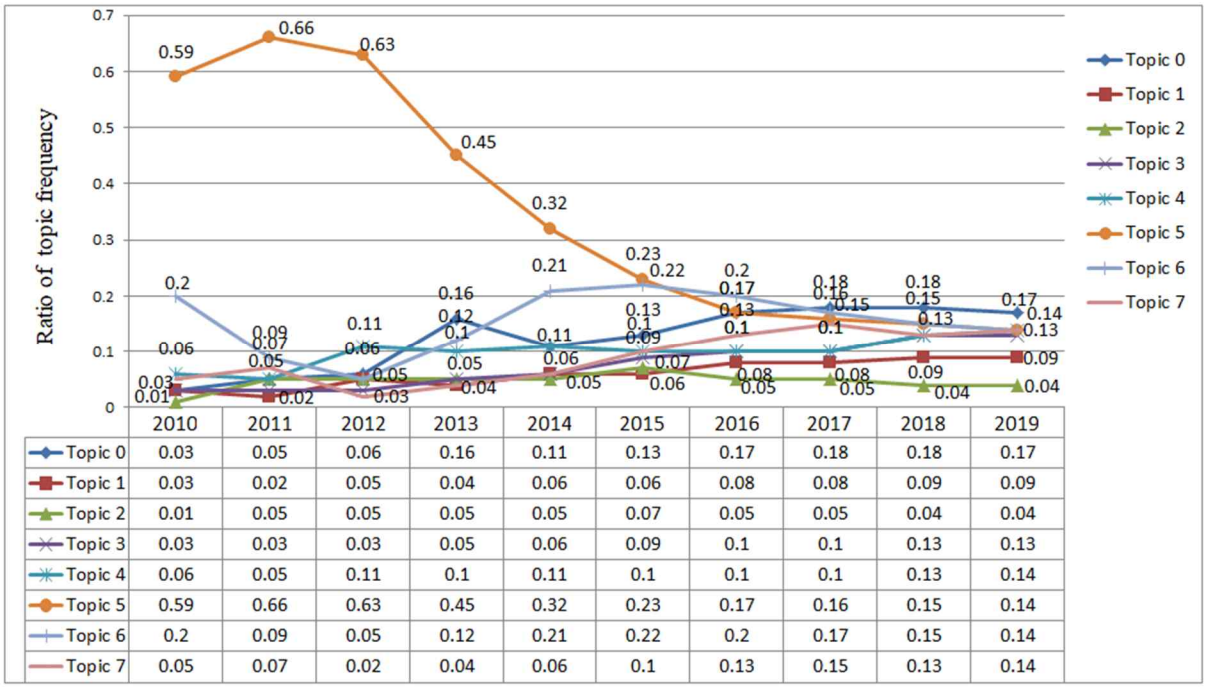


Figure 4. The number of documents by topic for four countries

4.2 Topic frequency trend

Figure 5 shows the normalized topic frequency over time from 2010 to September 2019 based on Equation (2). In Figure 5, we can see that topic 5 (application algorithm) is being studied much more than other topics. It can be seen that there was a high interest in deep Learning's application in early 2010. However, the decrease in the frequency of topic 5 since 2013 is attributed to the fact that the performance of deep learning in the application sector did not meet expectations. In early 2010, topic 0, topic 1, topic 2, and topic 3 which are technology areas of deep learning were studied relatively little. It is presumed that the focus was on deep learning applications, and did not pay much attention to the research underlying the development. As a result, the performance of the applied field was sluggish because the technical research was not supported, and the frequency of topic 5 dropped sharply. Since 2014, topics related to technology have been increasing evenly, and stable technology development is expected. In particular, the rapid increase in topic 7 (patient diagnose)

explains the growing interest in deep learning in the field of medical diagnostics [9].



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Topic 0	0.03	0.05	0.06	0.16	0.11	0.13	0.17	0.18	0.18	0.17
Topic 1	0.03	0.02	0.05	0.04	0.06	0.06	0.08	0.08	0.09	0.09
Topic 2	0.01	0.05	0.05	0.05	0.05	0.07	0.05	0.05	0.04	0.04
Topic 3	0.03	0.03	0.03	0.05	0.06	0.09	0.1	0.1	0.13	0.13
Topic 4	0.06	0.05	0.11	0.1	0.11	0.1	0.1	0.1	0.13	0.14
Topic 5	0.59	0.66	0.63	0.45	0.32	0.23	0.17	0.16	0.15	0.14
Topic 6	0.2	0.09	0.05	0.12	0.21	0.22	0.2	0.17	0.15	0.14
Topic 7	0.05	0.07	0.02	0.04	0.06	0.1	0.13	0.15	0.13	0.14

Figure 5. Normalized ratio of topic from 2010 to 2019

5. CONCLUSION

We conducted research to provide insight into the research trends of deep learning and future research directions. For this purpose, topic analysis was performed on deep learning related literature over the past 10 years to derive 8 topics. We analyzed the normalized topic frequency and identified the frequency of literature in China, USA, Korea, and UK. Based on the results of this study, we have concluded that: Topic 2, human recognition was found to be less researched than other topics. Considering that the Fourth Industrial Revolution is the era of human-friendly intelligent information services, it is necessary to expand research on Topic 2. In addition, Topic 6, Feature extract, compared to other topics, especially more, has been done in China. Considering that it is an important factor in determining the performance of deep learning, we should pay more attention to research on this topic.

The 1st and 2nd AI booms did not develop properly and faced the ice age. In order not to repeat past mistakes, a strategy is needed to continue to develop and expand the current AI boom. We need to study the basic technologies such as architecture, algorithm, feature detection, parameter, etc. before researching deep learning based application services. The results of this study are expected to provide insights into research design of deep learning, investment strategies, and the long-term development strategy of the deep learning industry.

REFERENCE

[1] LeCun, Y., Bengio, Y., & Hinton, G. “Deep learning,” *nature,* Vol. 521, No.7553, pp. 436-444, 2015. doi:http://doi.org/10.1038/nature14539

[2] Amado, A., Cortez, P., Rita, P., & Moro, S. “Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis”, European Research on Management and Business Economics, Vol. 24, No.1, 1-7, 2018.

doi:http://doi.org/10.1016/j.iedeen.2017.06.002
[3] Lee, J. (2019). "A Study on Research Trend Analysis and Topic Class Prediction of Digital Transformation using Text Mining". International journal of advanced smart convergence, 8(2), 183-190. doi:http://doi.org/10.7236/IJASC.2019.8.2.183
[4] Blei, D.M., "Probabilistic topic models," *Commun. ACM*, Vol. 55, No. 4, pp. 77-84, 2012. doi:http://doi.org/10.1145/2133806.2133826
[5] Steyvers, M. and T. Griffiths, "*Probabilistic topic models"*, Handbook of latent semantic analysis, Vol. 427, No. 7, pp. 424-440, 2007.
[6] Blei, D.M., A.Y. Ng, and M.I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research"*, pp. 993-1022, 2003.
[7] Newman, D., S. Karimi, and L. Cavedon, "External evaluation of topic models," *Australasian Doc. Comp. Symp*., 2009.
[8] Bastani, K., Namavari, H., & Shaffer, J. "Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints," *Expert Systems with Applications,* Vol. 127, pp. 256-271, 2019. doi:http://doi.org/10.1016/j.eswa.2019.03.001
[9] Shen, D., Wu, G., & Suk, H.-I. "Deep learning in medical image analysis," *Annual review of biomedical engineering*, Vol. 19, pp. 221-248, 2017. doi:https://doi.org/10.1146/annurev-bioeng-071516-044442