Visual Descriptor Extraction from Patent Figure Captions: A Case Study of Data Efficiency Between BiLSTM and Transformer

Anonymous Author(s)

ABSTRACT

Technical drawings used for illustrating designs are ubiquitous in patent documents, especially design patents. Different from natural images, these drawings are usually made using black strokes with little color information, making it challenging for models trained on natural images to recognize objects. To facilitate indexing and searching, we propose an effective and efficient visual descriptor model that extracts object names and aspects from patent captions to annotate benchmark patent figure datasets. We compared two state-of-the-art named entity recognition (NER) models and found that with a limited number of annotated samples, the BiLSTM-CRF model outperforms the Transformer model by a significant margin, achieving an overall F1=96.60%. We further conducted a data efficiency study by varying the number of training samples and found that BiLSTM consistently beats the transformer model on our task. The proposed model is used to annotate a benchmark patent figure dataset.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**.

KEYWORDS

NLP, NER, big data, entity recognition, deep learning

ACM Reference Format:

Anonymous Author(s). 2018. Visual Descriptor Extraction from Patent Figure Captions: A Case Study of Data Efficiency Between BiLSTM and Transformer . In *JCDL '22: ACM/IEEE JOINT CONFERENCE ON DIGITAL LIBRARIES, June 20 – 24, 2022, Hybrid Conference, Cologne, Germany and Online.* ACM, New York, NY, USA, 5 pages.

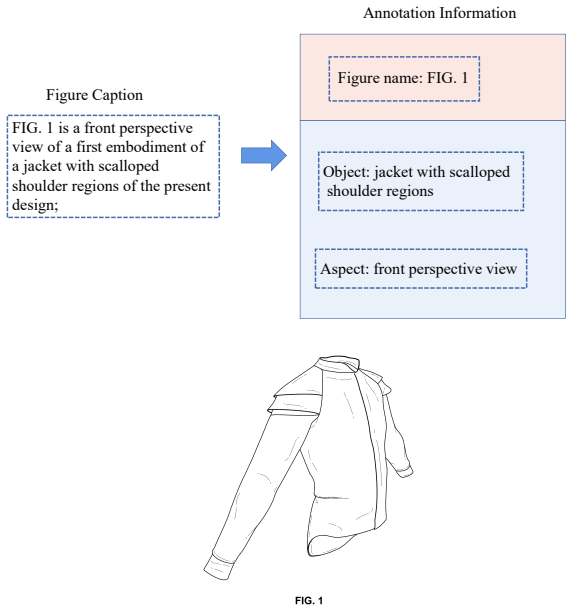

Figure 1: A technical drawing in US patent USD0836887-20190101 is annotated. Patent name can be parsed directly due to XML structure. Object and aspect can be obtained by our model.

1 INTRODUCTION

The number of patents in the United States has been steadily increasing since 2004[1]. Nowadays, there are about 7000 patents approved every week in the United States. This poses great labor and infrastructure challenges to searching and comparing figures in new and existing patents. Patent figures contain many different types, such as technical drawings, block diagrams, flow charts, plots, and grey scale natural images. In this paper, we focus on extracting textual descriptors for technical drawings in design patents. These drawings are of special interest because of two reasons. First, a survey paper [1] indicates drawings are an important component of patents and our study on a small set of figures randomly selected from US patents indicates that they represent approximately 95% in the design patents. Second, these descriptors can be especially useful to build search systems [2–4] that aid patent examiners to search for similar designs, which will speed up the patent examination and approval processes.

Automatic extraction of textual descriptors can also help us to build a large-scale dataset for training image captioning models. [4]. Most content-based approaches based on computer vision methods rely on models trained on a large number of annotated natural images such as ImageNet [5]. However, such methods may not perform well on technical drawings [4] because unlike natural images, technical drawings are usually formed by straight lines, curves, and dots. Most technical drawings are black and white. The lack of rich color information makes it more challenging to identify objects and their aspects solely based on visual information. To develop learning-based computer vision methods for such drawings, a large set of annotated images of technical drawings is needed. As shown in Figure 2, the benchmark dataset ImageNet [5] is annotated manually, while we use a NLP (Natural Language Processing)-based method for the annotation of patent drawings, which can automate this process. In addition, extracted patent descriptors can also be used for patent image retrieval, patent figure classification, and building patent knowledge graphs.

With the advance of natural language processing models, it is possible to mine the text to extract visual descriptors of objects in patent figures. In this work, we focus on extracting *object* names and *aspects* from figure captions. Figure 1 shows one example of technical drawings and captions, with objects and aspects.

[1] https://www.statista.com/statistics/256738/number-of-patents-in-force-in-the-us/

JCDL '22, June 20 – 24, 2022, Hybrid Conference Cologne, Germany and Online

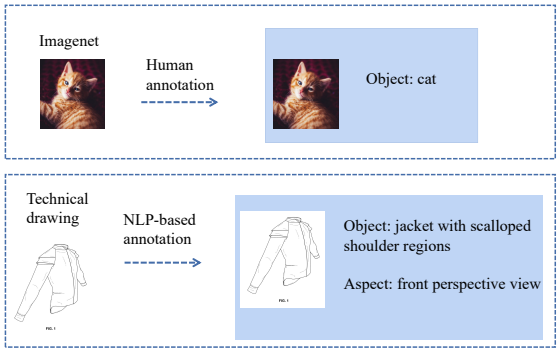

Figure 2: Existing benchmark dataset such as Imagenet uses human annotation while we use NLP-based annotation for patent figures.

Named Entity Recognition (NER) aims at recognizing mentions of rigid designators from text belonging to predefined semantic types such as person, location, and organization [6]. In general, the entities appearing in natural language can be beyond the scope of these named entities, such as domain knowledge entities [7], biomedical entities, and materials compositions [8]. A simple rule-based extractor such as a grammar-based noun phrase chunker does not generalize well because the text span of an object name or an aspect can be a subphrase or a superphrase of another phrase. An accurate extraction model should incorporate the grammatical and semantic information in the context. Recently, large-scale pre-trained models have shown advantages on representing text in NER tasks. However, we cannot directly apply pre-trained NER models because the tag types of pre-trained NER models do not match the tag types of our task. Therefore, we build a ground truth corpus by manually annotating a set of figure captions which are selected from US patents and then use it to train deep learning-based NER models.

The contributions of the paper are as follows:

(1) We propose a BiLSTM-CRF (Bidirectional Long Short-Term Memory Conditional Random Field) model to extract visual descriptions of technical drawings in US design patents. Our model outperforms the transformer model by a significant margin, using a training set consisting of 2700 annotated captions.
(2) We compiled a dataset containing 3300 captions with human-annotated visual descriptors, which is publicly available for future training and evaluation of NLP models. This will be provided in Github.
(3) We performed a data efficiency study and found that transformer exhibits a much steeper data efficiency curve than BiLSTM-CRF, while the performance is worse than BiLSTM-CRF in all the scenarios we examined.
(4) We automatically extracted object and aspect descriptors of 68094 patent figures, using the trained BiLSTM-CRF model. This dataset is released in Github and can be used for technical drawing object recognition and patent image captioning tasks.

2 RELATED WORK

Many previously published papers applied the BiLSTM-CRF architecture for NER tasks. One of the early works [9] introduced this architecture in NER. Then a series of papers used the BiLSTM architecture and achieved outstanding performance (with F1>0.91)

FIG. 1 shows a front perspective view of a collapsible pallet consolidator having full height walls;

FIG. 13 is a right side elevation view of the lock ring as shown in FIG. 9 showing my new design;

Figure 3: Examples of annotated captions for patent figures.

[10][11]. In addition to classic named entities explored in the papers above, this model was also used for extracting domain knowledge entities for scientific paper recommendation [12]. More works based on this architecture can be found in the review paper [6].

Transformer is a powerful model for many tasks such as machine translation and language model pre-training, but is less used for NER tasks. As mentioned in [6], transformer models will fail if they are not pre-trained on a huge corpus and when the training data is limited. [13] and [14] are two of the few papers that achieved SOTA performance with transformer models in NER tasks. They modified the original transformer model in different ways in order to improve performance.

A recent work [15] compared LSTM (long short-term memory networks) and BERT (a transformer-based architecture) for a small corpus for intent classification and found LSTM models could achieve significantly better results than a BERT model. Another work [16] compared the transformer models with LSTM models in the task of speech recognition in terms of training time and found transformer takes less time.

3 DATA

3.1 USPTO Patent Database

The dataset used in our experiments is from the United States Patent and Trademark Office (USPTO) patent database, which consists of the full-text (in XML format) and figures (in TIFF format) of patents ranging from 1976 to the present with new patent files released on a weekly basis. We parse the XML documents to obtain figure captions associated with figures. The figure captions are enclosed in special XML tags, which enables them to be accurately extracted.

3.2 Annotation of Ground-truth Data

To build the ground truth corpus, we randomly selected 3300 figure captions from 3300 patent figures in the 2020 dataset. Each caption is manually annotated by researchers in our lab using *brat*, a web-based annotation tool[2]. Two examples of annotated captions are shown in Figure 3. The annotation follows the BIO schema with five tags `['B-ASPECT', 'I-ASPECT', 'B-OBJECT', 'I-OBJECT', 'O']`, in which "I" indicates a tag inside an entity, "O" indicates a token belonging to no entity, and "B" indicates the tag being the beginning of an entity. During the annotation, we tend to annotate superphrase that provide more specific descriptions of an OBJECT. For example, "cooker with lid" and "men's shirt" will be annotated as OBJECT instead of "cooker" and "shirt".

[2]https://brat.nlplab.org/

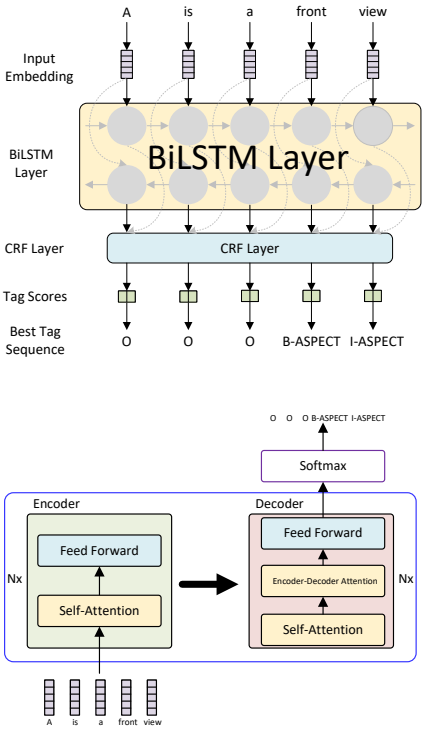

Figure 4: Architectures of BiLSTM-CRF and Transformer models.

The annotation results include a total of 2464 OBJECTS and 2958 ASPECTS. Then the ground truth corpus is split into training, validation, and testing datasets, each consisting of 2700, 300, and 300 captions respectively.

4 MODELS

As summarized by [6], deep-learning-based NER models are usually composed of 3 parts: distributed representations for input, a context encoder, and a tag decoder. The context encoder processes a sequence of tokens and outputs a vector of scores for all possible tags for each token. The tag decoder converts the scores into probabilities and then chooses the tags with the highest probability. We focus on two architectures: BiLSTM-CRF and the transformer.

4.1 Distributed Representations

We use pre-trained distributed language models for the input word embeddings:

- **GloVe:** GloVe [17] is a context-free embedding. We use the version pre-trained on 2B tweets with 27B tokens[3].
- **ELMo:** ELMo [18] is a context-dependent word embedding model. We use the implementation of AllenNLP[4] trained on WMT 2011 News Crawl data with 13.6M parameters.
- **BERT:** BERT [19] is a language mode trained under the transformer architecture. We use the basic version "bert-base-uncased" with 110M parameters.
- **RoBERTa:** RoBERTa [20] modifies the BERT pretraining procedure with parameters slightly increased. We use "roberta-base" with 125M parameters.

[3]https://nlp.stanford.edu/projects/glove/
[4]https://allennlp.org/elmo

NER Models	Embedding Models	Precision	Recall	F1
BiLSTM-CRF	*RoBERTa&OpenAI*	95.87	**96.50**	96.18
	RoBERTa	95.63	95.84	95.74
	BERT	96.27	96.06	96.17
	ALBERT	94.42	96.28	95.34
	DistilBERT	**96.92**	96.28	**96.60**
Transformer	*RoBERTa&OpenAI*	86.68	92.56	89.52
	RoBERTa	90.33	94.09	92.17
	BERT	93.98	95.62	94.79
	ALBERT	93.42	96.28	94.83
	DistilBERT	90.02	94.75	92.32

Table 1: Compare the performance of BiLSTM-CRF and transformer architectures. The biggest value in each column is marked in bold.

- **ALBERT:** ALBERT [19] has significantly fewer parameters than a traditional BERT architecture. We use "albert-base-v1" with 11M parameters.
- **RoBERTa&OpenAI:** This is RoBERTa fine-tuned by OpenAI on the outputs of the 1.5B-parameter GPT-2 model[5]. We use "roberta-base-openai-detector" with 125M parameters.
- **DistilBERT:** DistilBERT [21] leverages knowledge distillation during the pre-training phase while retaining the understanding capability of BERT. We use "'distilbert-base-uncased"' with 66M parameters.

4.2 BiLSTM-CRF Model

The BiLSTM-CRF architecture is based on Recurrent Networks. It's organized as shown in Figure 4 with a pre-trained embedding input, a bi-directional LSTM model as context encoder, and a CRF layer as tag decoder.

The BiLSTM layer captures contextual dependency for each token in both forward direction and backward direction, and then outputs the scores of possible tags for each token. The CRF layer learns the joint relationship between tags, excludes unreliable tag combinations, and further refines the probability scores for each token.

4.3 Transformer Model

The transformer model is based on the multi-head attention mechanism with a encoder-decoder structure. Similar to BiLSTM, the transformer model encodes the contextual information. However, the attention mechanism is able to select important relevant words in the sentence based on semantic information when processing each input token, which brings more powerful understanding capability. Transformer uses softmax as the last layer by default.

5 EVALUATION

The models are evaluated using standard metrics: precisions, recalls, and F1 scores. For a certain entity (such as OBJECT), the *precision* is defined as *the number of correctly extracted entities* divided by *the total number of entities extracted as* OBJECT. The *recall* is defined as *the number of correctly extracted entities* divided by *the total number of entities labeled as* OBJECT *in ground truth*. We use the strict string matching as the criteria. Either the extracted entity is a subphrase or a superphrase of the labeled entity was counted as a false sample.

[5]https://github.com/openai/gpt-2-output-dataset/tree/master/detector

NER Model	Embedding Models	All Entities			Aspect			Object		
		P	R	F1	P	R	F1	P	R	F1
BiLSTM-CRF	*RoBERTa&OpenAI*	95.87	96.50	96.18	99.20	99.20	99.20	91.87	93.20	92.53
	RoBERTa	95.63	95.84	95.74	98.81	99.20	99.01	91.75	91.75	91.75
	BERT	96.27	96.06	96.17	98.41	98.80	98.61	93.63	92.72	93.17
	AlBERT	94.42	96.28	95.34	97.64	98.80	98.22	90.57	93.20	91.87
	DistilBERT	96.92	96.28	**96.60**	99.20	99.20	**99.20**	94.09	92.72	**93.40**
	ELMo	95.06	95.06	95.06	98.34	98.34	98.34	91.41	91.41	91.41
	Glove	96.33	91.57	93.89	99.45	99.45	99.45	92.47	82.82	87.38

Table 2: BiLSTM-CRF-based NER models with different embeddings. P: precision. R: recall. Highest F1 scores are marked in bold.

As shown in Table 1, the BiLSTM-CRF architecture outperforms the transformer architecture in all embedding scenarios. The highest difference is generated when using roBERTaOpenAI as the input word embedding, with 96.18% for BiLSTM-CRF and 89.52% for Transformer.

As shown in Table 2, the BiLSTM-CRF architecture has precision, recall, and F1 scores higher than 90% in almost all scenarios. The transformer-based embeddings have better performance than other embeddings such as ELMo and GloVe. The model with DistilBERT embedding achieved F1 scores of 93.40% and 99.20% for object and aspect names, respectively, and a micro-average F1 score of 96.60% on the overall level. This is consistent with [22] which used DistilBERT to generate better sentence embeddings compared with BERT and RoBERTa.

6 ERROR ANALYSIS

One limitation of the BiLSTM-CRF model is that it may omit the object names randomly. The second limitation is that the extracted text span may be shorter than the annotated text span. For example, in "`FIG. 11 is a top view of the winged sofa with curved legs in first position;`" the object extracted by our best model is "winged sofa" while the annotated text is "winged sofa with curved legs". However, in a different example, "`FIG. 1 is a top front perspective view of a table with bench seating depicting the new design;`" the model extracts the complete and correct object name "table with bench seating", although the structure of this example is similar to the first one.

7 DATA EFFICIENCY ANALYSIS

In this paper [23], data efficiency is characterized as the performance of models given various amounts of training data. To compare the data efficiency between these two models, we varied the training data size from 500 to 2700.

As shown in Figure 5, the data efficiency of the transformer model depends on the input word embedding. In addition, the f1-score of the transformer model increases as the training data size increases from 500 to 2700, but it still consistently underperforms the BiLSTM-CRF model. A ground truth dataset with about 500 samples can be used to train a decent biLSTM-CRF model with F1= 95.15%, but is far from enough to train a transformer model. This is likely to be attributed to a much higher number of free parameters in the transformer than a BiLSTM-CRF. The data efficiency problem makes the transformer model a suboptimal option for NER tasks with a relatively small amount of samples.

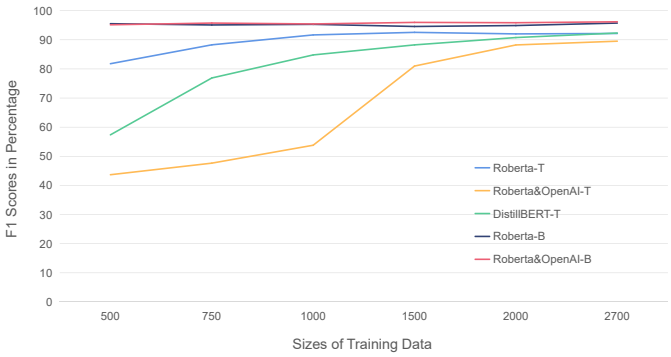

Figure 5: How performance grows with the increase in the amount of training data. T: Transformer, B: BiLSTM-CRF.

8 APPLICATION

In this section, we apply the DistilBERT-BiLSTM-CRF model on a corpus of unlabeled data, consisting of 68094 figures from 8032 design patents in 2019. we extracted 4832 distinct object names and 1811 distinct aspect names (67679 object names and 67105 aspect names in total, taking duplicates into account). Some most frequent types such as "display screen", "container" and "shoe" appear 577, 377 and 311 times, respectively. In the training data, there are 557 distinct object names and 346 distinct aspect names. Out of the extraction results, only 120 objects and 127 aspects appear in the training data, which are 0.24% and 7% of the total number. The majority of objects (>99%) and aspects (93%) is new, indicating the generalization power of the BiLSTM-CRF model.

9 CONCLUSION

In this paper, we proposed an effective and data-efficient visual description extractor using the BiLSTM-CRF sequence tagging model. The model achieves F1=99.20% for aspect and F1=93.40% for object extraction, evaluated using 300 figure captions from US patents in 2020. The object names and aspects extracted can be used for automatically labeling a large number of technical drawings, which can further be used for training computer vision based models and patent figure retrieval. We found that the BiLSTM-CRF model is more data-efficient than the transformer model for this task. A more systematic study about this will be our future work.

REFERENCES

[1] Allan Hanbury, Naeem Bhatti, Mihai Lupu, and Roland Mörzinger. Patent image retrieval: a survey. In *Proceedings of the 4th workshop on Patent information retrieval,*

pages 3–8, 2011.

[2] Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (notebook papers/labs/workshop)*. Citeseer, 2011.

[3] Stefanos Vrochidis, Anastasia Moumtzidou, and Ioannis Kompatsiaris. Concept-based patent image retrieval. *World Patent Information*, 34(4):292–303, 2012.

[4] Liping Yang, Ming Gong, and Vijayan K Asari. Diagram image retrieval and analysis: Challenges and opportunities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 180–181, 2020.

[5] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[6] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[7] Jian Wu, Md Reshad Ul Hoque, Gunnar W. Reiske, Michele C. Weigle, Brenda T. Bradshaw, Holly D. Gaff, Jiang Li, and Chiman Kwan. A comparative study of sequence tagging methods for domain knowledge entity recognition in biomedical papers. In *JCDL '20: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, Virtual Event, China, August 1-5, 2020*, pages 397–400. ACM, 2020. doi: 10.1145/3383583.3398602. URL https://doi.org/10.1145/3383583.3398602.

[8] Leigh Weston, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. *J. Chem. Inf. Model.*, 59(9):3692–3702, 2019. doi: 10.1021/acs.jcim.9b00470. URL https://doi.org/10.1021/acs.jcim.9b00470.

[9] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[10] Liyuan Liu, Jingbo Shang, Xiang Ren, Frank Xu, Huan Gui, Jian Peng, and Jiawei Han. Empower sequence labeling with task-aware neural language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[11] Abbas Ghaddar and Philippe Langlais. Robust lexical features for improved neural network named-entity recognition. *arXiv preprint arXiv:1806.03489*, 2018.

[12] Md Reshad Ul Hoque, Dash Bradley, Chiman Kwan, Agnese Chiatti, Jiang Li, and Jian Wu. Searching for evidence of scientific news in scholarly big data. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 251–254, 2019.

[13] Qipeng Guo, Xipeng Qiu, Pengfei Liu, Yunfan Shao, Xiangyang Xue, and Zheng Zhang. Star-transformer. *arXiv preprint arXiv:1902.09113*, 2019.

[14] Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*, 2019.

[15] Aysu Ezen-Can. A comparison of lstm and bert for small corpus. *arXiv preprint arXiv:2009.05451*, 2020.

[16] Albert Zeyer, Parnia Bahar, Kazuki Irie, Ralf Schlüter, and Hermann Ney. A comparison of transformer and lstm encoder decoder models for asr. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 8–15. IEEE, 2019.

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[18] Suzana Ilić, Edison Marrese-Taylor, Jorge A Balazs, and Yutaka Matsuo. Deep contextualized word representations for detecting sarcasm and irony. *arXiv preprint arXiv:1809.09795*, 2018.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[21] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[22] Lamia Salsabil, Jian Wu, Muntabir Hasan Choudhury, William A Ingram, Edward A Fox, Sarah J Rajtmajer, and C Lee Giles. *A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software*. Association for Computing Machinery, 2022. doi: 10.1145/3487553.3524658.

[23] Amina Adadi. A survey on data-efficient algorithms in big data era. *Journal of Big Data*, 8(1):1–54, 2021.