


Article

Empowering Multi-Span Question Answering with Expansive Information Injection using Large Language Models

Zhiyi Luo ¹ , Yingying Zhang¹ and Shuyun Luo ^{1,*}

¹ School of Computer Science and Technology and the Key Laboratory of Intelligent Textile and Flexible Interconnection of Zhejiang Province, Zhejiang Sci-Tech University, Hangzhou, China; luzhiyi@zstu.edu.cn

* Correspondence: shuyunluo@zstu.edu.cn;

Abstract: Retrieval-based question answering in the automotive domain requires a model to comprehend and articulate relevant domain knowledge, accurately understand user intent, and effectively match the required information. Typically, these systems employ a encoder-retriever architecture. However, existing encoders, which rely on pretrained language models, suffer from limited specialization, insufficient awareness of domain knowledge, and biases in user intent understanding. To overcome these limitations, this paper constructs a Chinese corpus specifically tailored for the automotive domain, comprising question-answer pairs, document collections, and multitask annotated data. Subsequently, a pretraining-multitask fine-tuning framework based on masked language models is introduced to integrate domain knowledge as well as enhance semantic representations, thereby yielding benefits for downstream applications. To evaluate system performance, an evaluation dataset is created using ChatGPT, and a novel retrieval task evaluation metric called Mean Linear Window Rank (MLWR) is proposed. Experimental results demonstrate that the proposed system (based on BERT_{base}), achieves accuracies of 77.5% and 84.75% for Hit@1 and Hit@3 respectively, in the automotive domain retrieval-based question answering task. Additionally, the MLWR reaches 87.71%. Compared to a system utilizing a general encoder, the proposed multitask fine-tuning strategy shows improvements of 12.5%, 12.5%, and 28.16% for Hit@1, Hit@3, and MLWR, respectively. Furthermore, when compared to the best single-task fine-tuning strategy, the enhancements amount to 0.5%, 1.25%, and 0.95% for Hit@1, Hit@3, and MLWR, respectively.

Keywords: deep learning; pretrained language model; retrieval-based question answering; multi-task learning; fine-tuning

Citation: Luo, Z.; Zhang, Y.; Luo, S. Title. *Mathematics* **2023**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

Copyright: ©2023 by the authors. Submitted to *Mathematics* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In summary, the main contributions in this paper are as follows:

- We construct Chinese question and answer corpora, document corpora, and multi-task annotated corpora specifically tailored to the automotive domain.
- We propose a joint learning framework with a pretraining-multitask fine-tuning architecture to incorporate domain knowledge and conduct a comparative analysis of the contributions of various auxiliary task objectives to model performance.
- We create an evaluation dataset based on ChatGPT using a semi-automated approach, along with the introduction of the MLWR metric for evaluation.

2. Related Work

3. Our Approach

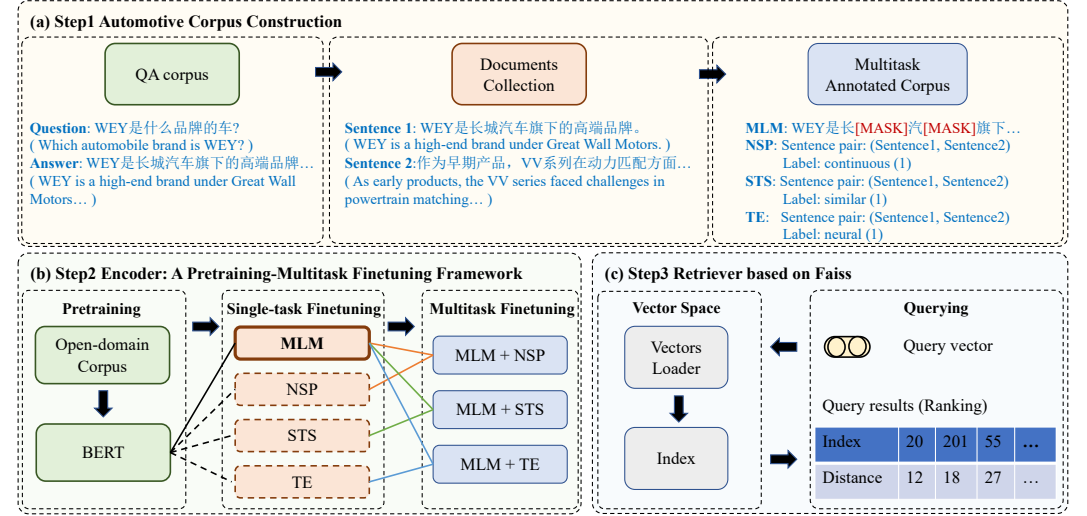


Figure 1. An overview of our retrieval-based QA system. **(a) Step 1:** Automotive Corpus Construction. **(b) Step 2:** The encoder module under a pretraining-multitask fine-tuning framework. **(c) Step 3:** The retriever module based on Faiss library.

3.1. Corpus Construction

We construct a QA dataset specific to the automotive domain by gathering question-answer pairs from authoritative resources, including professional databases, websites and other relevant resources. This dataset comprises a total of 7,157 question-answer pairs, with each pair consisting of a question and its corresponding answer. Table 1 provides an illustrative example of the dataset.

Table 1. The constructed Chinese QA corpus in automotive domain.

An Example of QA pairs	
{	"id": 0,
	"question": "WEY 是什么品牌的车?",
	(Which automobile brand is WEY?)
	"answer": "WEY 是长城汽车旗下的高端品牌...",
	(WEY is a high-end brand under Great Wall Motors...)
}	

Table 2. The constructed Chinese QA corpus in automotive domain.

Name	Data Format
MLM Corpus	[MASK][MASK] 车队首次在 1994 年的 JGTC 第四站比 [MASK] 中亮相, 获得了资格赛第二名的位置。(The [MASK] team made its debut in the fourth round of the JGTC in 1994, securing the second position in the qualifying [MASK].)
NSP Corpus	Sentence1: 上世纪 90 年代, TRD 为丰田 TOM'S 车队打造了 Supra 赛车。(In the 1990s, TRD built Supra race cars for the Toyota TOM'S team.) Sentence2: 丰田车队首次在 1994 年的 JGTC 第四站比赛中亮相, 获得了资格赛第二名的位置。(The Toyota team made its debut in the fourth round of the JGTC in 1994, securing the second position in the qualifying race.) Label: continuous (1)
STS Corpus	Sentence1: 丰田车队首次在 1994 年的 JGTC 第四站比赛中亮相, 获得了资格赛第二名的位置。(The Toyota team made its debut in the fourth round of the JGTC in 1994, securing the second position in the qualifying race.) Sentence2: 在 1994 年的 JGTC 第四站比赛中, 丰田车队首次参赛, 并在资格赛中获得了第二名的成绩。(In 1994, during the fourth round of the JGTC, the Toyota team made its debut and secured a second-place finish in the qualifying race.) Label: similar (1)
TE Corpus	Sentence1: 一个年轻的黑人正试图向另外两个人解释一些事情。(A young black person is trying to explain something to two other individuals.) Sentence2: 一位年轻的黑人男子正在和另外两位说话。(A young black man is talking to two other people.) Label: entailment (0)

4. Experiments

In this section, we first create an evaluation dataset for the automotive domain. Next, we introduce evaluation metrics utilized in our experiments. Finally, we conduct a comparative analysis against competing models to substantiate the effectiveness of our system.

4.1. Evaluation Dataset

Firstly, we create a retrieval corpus (shown in Table 3) utilizing the question set outlined in Section 3.1. Then, we randomly sample 100 questions from the retrieval corpus. These selected questions are then subjected to synonymous rephrasing using ChatGPT, which ultimately leads to the creation of the evaluation dataset (shown in Table 4).

Table 3. The retrieval corpus consisted of all questions in QA dataset.

Questions
{ 未来车门会是什么样? } (What will the car doors of the future look like?)
{ WEY 是什么品牌的车? } (Which automobile brand is WEY?)
{ 长安沃尔沃和吉利沃尔沃有什么区别? } (What is the difference between Changan Volvo and Geely Volvo?)

Specifically, we transform the rephrasing task description and the question text into prompts, which are fed into ChatGPT to generate the rephrased results. These results are then subject to manual screening to create the final evaluation dataset. The prompts are designed following specific principles: (1) Role assignment: ChatGPT assumes the role of an automotive engineer, for instance, with a prompt like “Imagine you are an automotive engineer”. (2) Rephrasing task description: a detailed explanation of the rephrasing task requirements is provided, such as “You will receive a text related to the automotive domain

and your task is to rewrite it, ensuring that the length remains similar to the original while preserving its meaning". (3) Result requirements: the desired outcomes of ChatGPT's generation are described, for example, with a prompt like "Please provide 6 rephrased results for each data point and rank them in descending order based on their quality". Subsequently, human annotators manually review the rephrased results generated by ChatGPT, carefully examining each question's rephrased versions, and selecting the top 4 synonyms with the highest quality. This process yields a dataset comprising 400 user queries, as depicted in Table 4. Each query corresponds to a question in the *queries* field and has a corresponding reference question in the retrieval corpus, indicated by the *reference* field. This user query dataset is utilized to evaluate the performance of the retrieval-based question answering model.

Table 4. The evaluation dataset in our experiments.

An illustration Example
<pre> { "reference": "长安 cs75plus 车型热销背后的几点思考", (Some thoughts behind the hot sales of the Changan CS75 Plus model.) "queries": ["长安 cs75plus 车型热销背后原因解析", (Analysis of the reasons behind the high sales of the Changan CS75 Plus model.) "购买长安 cs75plus 车型的几点原因", (Several reasons for purchasing the Changan CS75 Plus model.) "长安 cs75plus 车型为什么能够热销，列出几点原因", (Please list a few reasons to explain why the Changan CS75 Plus model sells well.) "长安 cs75plus 车型，热销背后的原因思考", (Reflection on the reasons behind the popularity of the Changan CS75 Plus model.)] }</pre>

4.2. Evaluation Metrics

The hit rate at K (Hit@K) is a widely used evaluation metric for assessing the performance of retrieval systems. It measures the capability to rank the correct target of user queries among the top K retrieval results. When conducting Hit@K for evaluation, a query is considered a hit if the true target is included in the top K retrieval results; otherwise, it is categorized as a miss. In our experiments, we utilize Hit@1, Hit@3, and Hit@5 as evaluation metrics for the retrieval model. Formally, the Hit@K is computed as:

$$Hit@K = \frac{N_K}{|Q|}, \quad (1)$$

where N_K represents the total number of hits for N retrieval tasks, and $|Q|$ is the total number of true targets across all queries.

While Hit@K is a straightforward metric, it solely focuses on the top-ranked results, disregarding other potentially valuable outcomes. To account for the ranking of all results, we utilize the Mean Reciprocal Rank (MRR) as an evaluation metric. MRR is computed as follows:

$$MRR = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{1}{r_i}, \quad (2)$$

where $|S|$ represents the total number of queries, and r_i denotes the ranking of the true target corresponding to the i -th query in the retrieved results.

In addition, we present a novel evaluation metric designed based on real user behavior. Considering that the result page of the QA system has the capability to display multiple results, the difference between the top-ranked and fifth-ranked feedback results holds minimal significance in terms of user experience. However, the MRR metric assigns a score difference of 0.8 between the first and fifth ranks (with a maximum score of 1). Moreover, considering that users typically limit their exploration to the initial pages of retrieval re-

sults, if a rank exceeds a certain threshold, it indicates a failure of the query sample to produce the correct result within the system. Despite a low ranking, MRR still provides a positive evaluation score when used for assessment. To address these concerns, we propose a new evaluation metric called Mean Linear Window Rank (MLWR). MLWR linearly adjusts the impact of ranking on the evaluation metric and incorporates a window size to exclude query samples with rankings beyond the predefined window. This approach better aligns with the user experience requirements of the system and provides a comprehensive assessment of retrieval system performance. Formally, MLWR is computed as follows:

$$MLWR = \sum_{i=1}^{|S|} \frac{\max\{0, N - r_i + 1\}}{N \times |S|}, \quad (3)$$

where $|S|$ represents the total number of samples, r_i denotes the ranking of the true target corresponding to the i -th query in the retrieved results, and N is the window size.

4.3. Comparison Results

In this section, we compare the performance of pretrained models, single-task fine-tuned models, and multitask fine-tuned models on the automotive domain QA retrieval task. By thoroughly examining the experimental results, we identify the most effective multitask fine-tuning strategy.

Table 5. Comparison of Hit@K, MRR, MLWR using single-task fine-tuning models based on $BERT_{base}$.

Model	Hit@1(%)		Hit@3(%)		Hit@5(%)		MRR(%)		MLWR(%)	
	CLS	MEAN	CLS	MEAN	CLS	MEAN	CLS	MEAN	CLS	MEAN
$BERT_{base}$	30.00	64.50	39.50	72.25	43.25	77.25	36.57	70.21	47.37	79.97
FT-MLM	47.25	77.00	55.50	83.50	57.00	85.75	52.19	80.94	58.96	86.76
FT-STs	13.00	60.25	18.50	71.25	22.25	74.75	17.73	66.58	26.21	76.19
FT-TE	5.00	24.75	8.50	30.25	9.00	32.75	7.27	28.89	10.73	35.41
FT-NSP	14.75	59.00	14.75	67.75	20.75	71.00	18.53	64.53	24.56	73.29

Table 6. Comparison of Hit@K, MRR, MLWR using multitask fine-tuning models based on $BERT_{base}$.

Model	Hit@1(%)	Hit@3(%)	Hit@5(%)	MRR(%)	MLWR(%)
STS+MLM [CLS]	74.25	84.00	85.25	79.54	87.12
NSP+MLM [CLS]	8.25	15.00	18.75	13.08	21.65
STS+MLM [MEAN]	77.50	84.75	87.00	81.60	87.71
NSP+MLM [MEAN]	73.25	78.75	80.50	76.77	82.17

Table 7. Comparison of Hit@K, MRR, MLWR using the best single-task and multitask fine-tuning models based on $BERT_{large}$ and RoBERTa.

Model	Hit@1(%)		Hit@3(%)		Hit@5(%)		MRR(%)		MLWR(%)	
	CLS	MEAN	CLS	MEAN	CLS	MEAN	CLS	MEAN	CLS	MEAN
$BERT_{large}$	26.50	64.50	39.50	72.25	43.25	77.25	36.57	70.21	47.37	79.97
FT-MLM	29.25	76.75	36.50	82.00	39.50	85.25	33.88	80.42	41.56	86.19
MLM+STS	18.50	78.00	22.20	83.50	23.50	85.75	21.13	81.67	25.16	88.42
RoBERTa	69.00	73.00	77.75	79.50	79.25	82.75	74.16	77.41	81.59	84.61
FT-MLM	73.25	75.75	79.25	81.75	83.25	84.00	77.63	79.79	85.31	86.81
MLM+STS	37.50	78.75	48.25	84.50	51.75	87.75	44.46	82.50	55.16	88.79

To illustrate the generalization capability of our framework in mitigating bias within pretrained models, we extend our evaluation beyond the BERT-base model to also include the BERT-large and RoBERTa models. As shown in Table 7, our framework exhibits effective generalization across a variety of pretrained models. It is worth noting that RoBERTa, which is solely pretrained using the MLM task, demonstrates good performance with the CLS encoding, achieving a Hit@1 score of 69%. However, in pretraining models such as BERT-base and BERT-large, the CLS representation primarily learns from the NSP task, resulting in sub-optimal performance when directly utilizing the CLS encoding. These findings highlight the negative impact of the NSP task as a training objective on QA tasks, while emphasizing the benefits of the MLM task as a training objective for downstream QA tasks.

4.4. Case Study

In this section, we conduct a case study to qualitatively analyze competing models. Table 8 presents the results obtained when inputting an example user query, namely “如何对高尔夫 7 系列车型进行自我检修?” (How to perform self-maintenance on Volkswagen Golf7 series?), and showcases the responses retrieved by the competing models utilizing BERT_{base} in proposed framework. The baseline BERT model provides a historical overview of the 7th Generation Volkswagen Golf, while the single-task FT-MLM model compares the 308S and Golf 7 models. However, neither of these models is capable of providing an accurate answer to the query. In contrast, the STS+MLM joint fine-tuning model effectively responds with the precise and specific steps to perform maintenance on the FAW-Volkswagen Golf7, thereby correctly addressing the query. The STS+MLM joint model demonstrates its effectiveness in capturing the query intentions, surpassing other competing models in terms of qualitative performance.

Table 8. Retrieved answers from baseline BERT_{base} model, single-task FT-MLM model, and multitask STS+MLM model.

Input Query: “如何对高尔夫 7 系列车型进行自我检修?” (How to perform self-maintenance on Volkswagen Golf7 series?)	
Model	Retrieved answers
BERT _{base}	1. 高尔夫 Mk1: 1974 年 5 月, 第一代高尔夫... 2. 高尔夫 Mk2:1983 年 8 月, 大众推出第二代... (1. Golf Mk1: In May 1974, the first generation Golf was introduced.. 2. Golf Mk2: In August 1983, Volkswagen unveiled the second generation...)
FT-MTM	1. 动力组合对比: 高尔夫一共有三种排量... 2. 质保政策对比: 高尔夫官方保修周期为 3 年或者... (1. Powertrain Comparison: The Golf is available in three different engine displacements... 2. Warranty Policy Comparison: The official warranty period for the Golf is 3 years or...)
MLM+STS	1. 更换 4.5 升 5w40 美孚全合成机油、博世机滤、博世空气滤芯、博世空调滤芯、博世汽油滤芯、3m 燃油添加剂、免拆三元催化清洗剂 2. 动平衡、更换原厂轮毂盖 (1. Replace 4.5 liters of 5w40 Mobil fully synthetic engine oil, Bosch oil filter, Bosch air filter, Bosch cabin air filter, Bosch fuel filter, 3M fuel additive, and non-dismantle catalytic converter cleaner. 2. Dynamic balancing and replace original wheel hub covers.)

5. Conclusion

In this paper, we first focus on constructing QA corpora, document sets, and multitask fine-tuning datasets specifically tailored to the automotive domain. Then, we propose a pretraining-multitask fine-tuning learning framework to develop an encoder model that incorporates domain knowledge, enhancing the semantic representation of texts and enabling effective retrieval QA applications. The experimental results reveal that within the multitask fine-tuning framework (based on BERT_{base}), the joint fine-tuned model MLM+STS

achieves the highest performance. It attains Hit@1 and Hit@3 accuracies of 77.5% and 84.75%, respectively, along with an MLWR of 87.71%. These outcomes signify substantial improvements of 12.5, 12.5, and 28.16 percentage points, respectively, compared to the baseline BERT model. Additionally, when compared to the best-performing single-task fine-tuned model, FT-MLM, the observed enhancements amount to 0.5, 1.25, and 0.95 percentage points, respectively.

Funding: This research was supported by the Natural Science Foundation of Zhejiang Province, China (Grant No. LQ22F020027) and the Key Research and Development Program of Zhejiang Province, China (2023C01041).

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

References

- Gupta, N.; Tur, G.; Hakkani-Tur, D.; et al. The AT&T spoken language understanding system. *IEEE Transactions on Audio, Speech, and Language Processing*, **2005**, *14*(1), 213–222.
- Ferrucci, D.G.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.A.; Lally, A.; Murdock, J.W.; Nyberg, E.; Prager, J.; Schlaef, N.; Welty, C. Building Watson: An Overview of the DeepQA Project. *AI Magazine* **2010**, *31*(3), 59–79.
- Qu, C.; Yang, L.; Qiu, M.H. Open domain question answering using early fusion of knowledge bases and text. Available online: <https://doi.org/10.48550/arXiv.1809.00782>. (accessed on 19-05-2023).
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention Is All You Need. In Proceedings of Advances in Neural Information Processing Systems (ANIPS), California, USA, 2017; 5998–6008.
- Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Minneapolis, 2019; 4171–4186.
- Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. Available online: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. (accessed on 19-05-2023).
- Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language Models are Unsupervised Multitask Learners. Available online: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. (accessed on 19-05-2023).
- Brown, T.; Mann, B.; Ryder, N.; et al. Language models are few-shot learners. In proceedings of Advances in Neural Information Processing Systems, Virtual-only Conference, 2020, 33: 1877–1901.
- Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Austin, Texas, USA, 2016, 2383–2392.
- Rajpurkar, P.; Jia, R.; Liang, P. Know what you don't know: Unanswerable questions for squad. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 2018, 784–789.
- Li, H.; Tomko, M.; Vasardani, M.; Baldwin, T. MultiSpanQA: A dataset for multi-span question answering. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Seattle, Washington, 2022, 1250–1260.
- Yang, Wei; et al. End-to-end open-domain question answering with bertserini. Available online: <https://doi.org/10.48550/arXiv.1902.01718>. (accessed on 19-05-2023).
- Zhang, Q.; Chen, S.S.; Xu, D.K.; Cao, Q.Q.; Chen, X.J.; Cohn, T.; Fang, M. A Survey for Efficient Open Domain Question Answering. Available Online: <https://arxiv.org/abs/2211.07886>. (accessed on 19-05-2023).
- Jones, S.K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **1972**, *28*(1), 11–21.
- Baldwin, T.; Marneffe M.C.; Han, B.; et al. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In Proceedings of the workshop on noisy user-generated text, Beijing, China, 2015, 126–135.
- Derczynski, L.; Maynard, D.; Rizzo, G.; et al. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* **2015**, *51*(2), 32–49.
- Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. Available online: <https://arxiv.org/abs/1301.3781>. (accessed on 19-05-2023).
- Harris, Z.S. Distributional structure. *Journal of Documentation* **1954**, *10*(2–3), 146–162.
- Pennington, J.; Socher, R.; Manning, C.D. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Maryland, USA, 2014; 1532–1543.
- McCann, B.; Bradbury, J.; Xiong, C.; Socher, R. Learned in translation: Contextualized word vectors. In Proceedings of Advances in Neural Information Processing Systems (NIPS), California, USA, 2017; 6297–6308.

21. Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, USA, 2018; 2227–2237.
22. Devika, R.; Vairavasundaram, S.; Mahenthara, C.S.J.; et al. A Deep Learning Model Based on BERT and Sentence Transformer for Semantic Keyphrase Extraction on Big Social Data. In *IEEE Access* **2021**, *9*, 165252–165261.
23. Natarajan, B.; Rajalakshmi, E.; Elakkiya, R.; et al. Development of an End-to-End Deep Learning Framework for Sign Language Recognition, Translation, and Video Generation. In *IEEE Access* **2022**, *10*, 104358–104374.
24. Bentley, J.L. Multidimensional binary search trees used for associative searching. *Communications of the ACM* **1975**, *18*(9), 509–517.
25. Liu, T.; Moore, A.W.; Gray, A.; New algorithms for efficient high-dimensional nonparametric classification. *JMLR* **2006**, *7*, 1135–1158.
26. Omohundro SM. Five balltree construction algorithms. Berkeley, California, USA: International Computer Science Institute, 1989.
27. Friedman, J.H.; Bentley, J.L.; Finkel, R.A. An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software* **1977**, *3*(3), 209–226.
28. Jégou, H.; Douze, M.; Schmid, C. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2011**, *33*(1), 117–128.
29. Indyk, P.; Motwani, R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, New York, USA, 1998; 604–613.
30. Malkov, Y.A.; Yashunin, D.A. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. Available online: <https://arxiv.org/abs/1603.09320>. (accessed on 19-05-2023).
31. Jégou, H.; Perronnin, F.; Douze, M.; Sánchez, J.; Pérez, P.; Schmid, C. Aggregating local descriptors into a compact image representation. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010; 3304–3311.
32. Wang, K.; Liu, Z.; Lin, Y.; Lin, W.; Han, S. HAQ: Hardware-Aware Automated Quantization with Mixed Precision. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Angeles, USA, 2019; 1669–1678.
33. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* **2019**, *7*(3), 535–547.
34. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D. RoBERTa: A Robustly Optimized BERT Pretraining Approach. Available online: <https://arxiv.org/abs/1603.09320>. (accessed on 19-05-2023).
35. Clark, C.; Lee, K.; Chang, M.W.; Kwiatkowski, T.; Collins, M.; Toutanova, K. SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics* **2020**, *8*, 64–77.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.