

Deep Learning Methods for Solving the NLI Problem

Subjects: Computer Science, Artificial Intelligence

Contributor: Petros Eleftheriadis, Isidoros Perikos, Ioannis Hatzilygeroudis

Natural language inference (NLI) is one of the most important natural language understanding (NLU) tasks. NLI expresses the ability to infer information during spoken or written communication. The NLI task concerns the determination of the entailment relation of a pair of sentences, called the premise and hypothesis. If the premise entails the hypothesis, the pair is labeled as an “entailment”. If the hypothesis contradicts the premise, the pair is labeled a “contradiction”, and if there is not enough information to infer a relationship, the pair is labeled as “neutral”.

Keywords: deep learning ; natural language processing ; natural language inference

1. Introduction

During human communication, a lot of information is conveyed. Usually, the receiver gains a lot more information than what is uttered by the speaker. This is due to the natural inference capabilities of humans or general world knowledge. For example, the sentence “John grew up in Spain” gives a lot more information than simply the birthplace of John. Researchers can assume that John speaks Spanish, and he went to school in Spain. Additionally, researchers can assume that John adopts the Spanish culture and therefore likes or plays football; furthermore, he likes Spanish food. All this information is taken for granted when researchers talk to each other, but to a computer, much more work needs to be done to feed this extra information. This is what natural language inference (NLI) deals with, and this is why, from its beginning, NLI was thought to be necessary for full natural language understanding ^[1].

Natural language inference is the task of determining the entailment relation between a premise and its hypothesis. This relation is usually described with one of three labels: entailment, contradiction, or neutral. An NLI sentence pair is classified as entailment if, given its premise, a human would be happy to infer the hypothesis. If the hypothesis directly contradicts the premise, the pair is labeled as a contradiction. In the case where not enough information is present to label a pair as entailment or contradiction, the ‘neutral’ label is annotated. In some datasets, the problem is set as a 2-label classification task, with one of the labels being ‘entailment’ and the other ‘not entailment’. The definition of “a human being happy to infer the hypothesis” may seem vague, but it has to be, given that the language itself is complicated, and many times, the same sentence can have different meanings to different people. In this line, it is well stated that the sentence pairs should be annotated by people that are awake, careful, moderately intelligent and informed ^[2].

In **Table 1**, three examples of sentence pairs for NLI and the resulting labels are presented. The *premise* is the first sentence, which provides us with the context to be used. The *hypothesis* is the second sentence, in which researchers will be asked whether it can be inferred from the premise. The label describes the entailment relationship of the two sentences. Given the first example in **Table 1**, researchers see that the correct label is *entailment* because, by reading the premise, researchers can safely infer that the hypothesis is also true. The next example, however, is not an entailment, as the hypothesis adds information that is not present in the premise. Researchers cannot safely say that the costume the girl is wearing is indeed a fairy costume, as it could be any type of costume. On the other hand, researchers do not have information that says that the costume is not a fairy costume; therefore, the correct label for this pair is *neutral*. In the last case, the first thought is that the pair should be labeled as a contradiction since researchers are talking about two different oceans. However, one could say that a boat sinking in the Pacific Ocean does not negate a boat sinking in the Atlantic Ocean, meaning that both sentences could be true. To avoid such conflictions, researchers always consider that researchers are talking about one single event. Therefore, the correct label is *contradiction*.

Table 1. Examples of NLI sentence pairs and their label.

Premise	Hypothesis	Label
A soccer game with multiple males playing	Some men are playing a sport	Entailment
The girl went out wearing a costume	The girl went out wearing a fairy costume	Neutral

Premise	Hypothesis	Label
A boat sank in the Pacific Ocean	A boat sank in the Atlantic Ocean	Contradiction

The NLI task needs special handling and has posed great challenges to the NLP research community since its formulation in 2005 with the recognizing textual entailment (RTE) challenges ^[3]. Since then, a lot of progress has been made both in terms of available data, as well as in the development of models that try to face the NLI problem. In 2015, the first large-scale NLI dataset was collected and provided neural models with 550,000 examples of crowdsource-labeled data ^[4]. These assisted in gaining more attention to the field of NLI, which brought many advances but also many criticisms. In 2017, the GLUE benchmark for natural language understanding ^[5] was published, which includes four NLI task datasets (MNLI, QNLI, RTE, WNLI). The collection method of each dataset is different, and so each dataset evaluates a different type of inference ability. This benchmark later evolved into SuperGLUE ^[6], which includes even harder and more challenging tasks.

As already stated, the main type of inference models today are deep neural network models. Up to 2017, the dominant type of inference model consisted of LSTMs (a deep learning model) encoding the premise and hypothesis, then applying an attention mechanism before passing the final vector to a SoftMax function for classification. A model that fits in this description is ESIM ^[7], which at that time demonstrated a state-of-the-art 88.0% accuracy on the SNLI dataset. Later adaptations of such models included additional learning from a knowledge graph to incorporate external world knowledge into the model ^[8]. Recently, the invention of transformers ^[9] and, more specifically, BERT ^[10] has switched research to using pre-trained models on large corpora of text that are then fine-tuned on specific data. Such models (BERT, XLNet, RoBERTa, ALBERT) quickly became the state-of-the-art models for NLI, performing with over 90% accuracy on some of the most challenging datasets.

However, many researchers questioned the ability of these models' inference abilities as they argued that the models take advantage of various annotation artifacts within the datasets to achieve their results. One of the first papers that addressed this issue was Poliak et al.'s hypothesis-only approach ^[11]. In their paper, they showed that only using the hypothesis sentence for some NLI datasets performed well above random chance, suggesting that specific words in the hypotheses are related to a specific label. Furthermore, some studies ^[12] found instances of social bias in the examples of the SNLI dataset, while others ^[13] found wrong labels in the SICK dataset as well. To continue testing that idea, some researchers developed NLI stress tests to break top-performing NLI models, such as the BreakingNLI dataset ^[14]. Furthermore, additional adversarial datasets were developed with a focus on collecting examples that top-performing models predict wrong ^[15].

The protocol of collecting annotations for sentence pairs has been criticized by researchers in the field as not satisfactory ^{[16][17]}. The idea behind the criticisms is that one should not take the label most voted on by humans as the gold label and should not ignore other opinions. Pavlick et al. ^[18] believe that NLI needs a revision because the vague task of "do as a human would" is not in agreement with the fact that different humans can extract different conclusions from the same sentence pair. Therefore, they suggested a new type of measurement that takes into account the entire spectrum of opinions. One such dataset is ChaosNLI ^[17], which instead of gold labels, uses a distribution over a collective of human opinions.

Another way of improving NLI is by creating new and more diverse datasets that cover a quite broad range of linguistic phenomena. An example in this direction is the IMPLI dataset ^[19], which uses sentence pairs of figurative language and tests models on idiomatic expressions and metaphors. Another dataset is ANLI, which uses adversarial data to trick NLI models. However, some researchers argue that models could be implemented to purposely exploit such data for better scores ^[20]. One recent approach to creating datasets for NLI is WANLI ^[21], which takes advantage of the progress of natural language generation models to include them in the process of data creation together with human annotators. Additionally, human explanations of data have been used in the training process, with the aim of improving NLI, with good results ^[22].

2. Methods for Solving the Natural Language Inference Problem

One of the first attempts at NLI was the decomposable attention model (DAM) ^[23]. The authors of the DAM presented an attention model for NLI that decomposes the problem into sub-problems that can be parallelized. The model works in three steps: attend, compare, and aggregate. In the first step, the premise and the hypothesis are encoded to two vectors, *a* and *b*, and then an attention mechanism is applied between them. The attention mechanism finds the sub-phrases in *b* that are softly aligned with every word in *a*, as well as the sub-phrases in *a* that are aligned with every word in *b*. In the

second step, the words of *a* and their aligned sub-phrases in *b* are compared with a feed-forward neural network and are passed in a vector, the same as *b*. In the third step, the values in each vector are aggregated and passed to a final MLP with SoftMax for classification. In addition to the three steps, an optional attention step was presented, called intra-sentence attention. This step can take place before the attend step, and it calculates the self-attention of a sentence for better representation. The DAM was trained on the SNLI dataset with GloVe embeddings. It provided state-of-the-art results at that time: 86.3% on the test set for the basic steps and 86.8% with the inclusion of the intra-attention step.

Another popular model is enhanced sequential inference modeling (ESIM) [7]. The authors of ESIM focus on enhancing sequential models for inference. Their model is based on chaining LSTMs and applying attention mechanisms. In their paper, they also present a tree-LSTM model that encodes syntactic knowledge and can be used together with ESIM to form HIM (hybrid inference model). ESIM also works in three steps, similar to DAM: input encoding, local inference modeling and inference composition. In the first step, the premise and hypothesis are encoded with bidirectional LSTMs. The outputs of the biLSTMs are combined with an attention mechanism. The resulting vectors (the softly aligned representations of *a* and *b*) and the original vectors *a* and *b* are then further combined and concatenated for better representation. In the third step, the outputs of the second step are again fed into two biLSTMs. Then, instead of aggregating the resulting vectors, they compute the average and max pooling of both vectors, and the four resulting vectors are concatenated and fed in an MLP with a SoftMax for classification. The authors trained the model on SNLI with 300d GloVe embeddings with a batch size of 32, using the Adam optimizer with a learning rate of 0.0004 and a dropout rate of 0.5. The model provided the best result at that time on the SNLI test set (88.00%).

The invention of transformers bred a new generation of models for NLI. The first popular adaptation was BERT (bidirectional encoding representations from transformers) [10]. BERT is a bidirectional language representation model that uses the architecture of the transformer model, more specifically, the encoder part of the transformer model. BERT is pre-trained on large corpora of unlabeled text that can then be fine-tuned on specific data. During pre-training, BERT performs two tasks: masked language modeling (MLM) and next-sentence prediction (NSP). MLM is essentially how BERT manages to learn in a bidirectional manner. The task of MLM includes masking a percentage of the input at random and then trying to figure out the correct word given its context (the words to the left and right of the mask). In the NSP task, BERT is given two sentences and must predict if one sentence (logically) follows the other. Due to its similarity, NSP is supposed to help with performance on NLI tasks. BERT is trained on the BookCorpus and the English Wikipedia (800 M + 2.5 B words). The input of BERT is a sequence of a maximum of 512 tokens. The first token is always the special (CLS) token, which encodes the classification, and sentences in the input are separated with the special (SEP) token. BERT uses WordPiece embeddings with a vocabulary of 30,000 words. In addition to token embeddings, BERT uses segment embedding, which matches a token with the sentence it appears in, and positional embedding, which tracks the position of the embeddings. There are two BERT sizes available: BERT-Base (110 M params) and BERT-Large (345 M params). BERT has demonstrated new state-of-the-art results on many NLP tasks, including NLI.

Another popular transformer model is RoBERTa (Robustly Optimized BERT Approach) [24]. The authors of RoBERTa believed that BERT was undertrained, so they presented a more optimized approach to training BERT called RoBERTa. The main changes regard the hyperparameters of pre-training, the task of MLM, the removal of the NSP task and the usage of more pre-training data. RoBERTa was trained on BookCorpus, Wikipedia, CC-NEWS, OPENWEBTEXT, and STORIES, totaling 148 GB of uncompressed text. The change in the MLM tasks is the introduction of dynamic masking, which creates a new masking pattern for each input sequence. In addition to dynamic masking, training data are replicated 10 times so that many masking patterns are created. They removed the NSP loss after carrying out several experiments that showed that NSP hurts performance. Finally, they used Byte-Pair Encoding, which encodes bytes instead of Unicode characters, resulting in a larger vocabulary of 50,000 words. As with BERT, RoBERTa comes in two sizes: Base and Large. RoBERTa-Base is comprised of $L = 12$ layers, a hidden size of $H = 768$, and $A = 12$ attention heads (110 M params). RoBERTa-Large has 355 M parameters ($L = 24$, $H = 1024$, $A = 16$) and has demonstrated state-of-the-art results on many NLP tasks, surpassing BERT.

A problem with BERT and ROBERTA is that they have hundreds of millions of parameters. This puts many restrictions on the training process as it requires a lot of GPU memory and is very time-consuming. Researchers from Google Research and Toyota Technological Institute in Chicago developed a different version of BERT called ALBERT (A Lite BERT) [25]. ALBERT is a light version of BERT that uses significantly fewer parameters. ALBERT uses the same architecture as BERT but makes three important distinctions. First, the vocabulary embedding is decomposed into two smaller ones. Second, parameters are shared across all layers, and third, NSP is replaced by a sentence-order prediction task. In BERT and RoBERTa, the embedding size is tied with the hidden size, $E = H$. The authors believe this is suboptimal, as H encodes context-dependent information and E encodes context-independent information; thus, H should be bigger than E , $H \gg E$. Therefore, instead of projecting the one-hot vectors onto H , they first project them on a smaller matrix, E , and

then E is projected onto H . This way, the parameters are reduced from $O(V \times H)$ to $O(V \times E + E \times H)$, which is critical when $H \gg E$. Next, parameter sharing is used as it provides a smoother change from layer to layer. Lastly, NSP was replaced by SOP, a task that focuses on inter-sentence coherence. SOP gives positive feedback when one sentence follows the other but negative when the same two sentences are inserted with their order switched. There are four versions of ALBERT: base, large, xlarge, and xxlarge. They have 11, 17, 60, and 235 million parameters, respectively. ALBERT's xxlarge version has demonstrated new state-of-the-art results, surpassing RoBERTa on several NLI tasks.

A more recent approach is ERNIE 3.0 [26], developed by researchers at Baidu. The authors saw a problem with the data used for training in popular models. They believed that the text used was plain and did not incorporate linguistic and word knowledge. Another problem they found was that the models were trained in an auto-regressive way, which, according to J. Devlin et al., worsens performance on downstream tasks [6]. Their proposal, ERNIE, is a unified framework to train large-scale models on a big corpus of text data as well as a knowledge graph. ERNIE combines the auto-regressive network and auto-encoding network so that both NLU and NLG are achieved. The model was tested on many Chinese NLP tasks and achieved first place on the SuperGLUE benchmark. ERNIE uses a shared network as the backbone to capture universal lexical and syntactic information, which is called the universal representation module, and it is built with a multi-layer Transformer-XL. ERNIE also uses a task-specific representation module, which is also a multi-layer transformer XL that is used to capture the top-level semantic representation for different task paradigms. While pre-training ERNIE, the authors used several pre-training tasks. The two word-aware tasks were knowledge-masked language modeling and document language modeling. The first masks phrases and named entities for the model to predict. The second is a pre-training task in which a traditional language model is used for generative purposes. Two structure-aware tasks were used: sentence reordering, in which the model tries to recreate a sentence given the segments of the sentence in random order, and sentence distance, which is an extension of the NSP task. The final task used was the universal knowledge-text prediction task. This task requires unstructured text and knowledge graphs. The way it works is, given a triple from the graph and a sentence from an encyclopedia, the model tries to predict the relation in the triple from the sentence. ERNIE 3.0 was tested on Chinese versions of NLI datasets and demonstrated new state-of-the-art results on OCLI and XNLI. In more detail, ERNIE demonstrated 82.75% accuracy on the OCNLI development set compared to the previous 78.80% accuracy exhibited by RoBERTa. On XNLI, the accuracy achieved on the test set was 83.77%, which is a smaller increase from the former best accuracy of 83.09%.

The pathways language model (PaLM) is a recent contribution to NLU and NLG developed by engineers at Google [27]. It is a 540 billion parameter language model based on a transformer trained on 6144 TPU v4 chips. It brought impressive results with few-shot learning as well as with fine-tuning specific NLP tasks. PaLM uses the transformer architecture, using only the decoder and some additional architectural differences. The authors preferred swiGLU activation functions over ReLU for the MLP. They also used a different formulation in each transformer block for faster training speeds at large scales. They used RoPE embeddings, shared input-output embeddings and a SentencePiece vocabulary with 256 k tokens. The model was pre-trained on 780 billion tokens. The data were taken from filtered webpages, books, Wikipedia, news articles, source code and social media conversations, which comprise 50% of the total data. They created three versions of the model: an 8 B parameter model with 32 layers, 16 attention heads, and a hidden layer size of 4096; a 62 B parameter model with 64 layers, 32 attention heads, and a hidden layer size of 8192; and finally, a 540 B parameter model with 118 layers, 48 attention heads and a hidden layer size of 18,432. The model was evaluated on 29 benchmarks, including SuperGLUE and ANLI. After being fine-tuned on SuperGLUE, the model performed close to SOTA results, and it currently stands 3rd on the leaderboard. On ANLI, the largest model exhibits 56.9% accuracy with few-shot learning.

ST-MoE (stable and transferable mixture-of-experts) is a recent approach to tackle NLU [28]. Developed by researchers at Google Brain, it is a 269-billion-parameter sparse model that manages to achieve state-of-the-art results in many NLP tasks. One of its main advantages is that it avoids the usual training instabilities often encountered in sparse models. Training instability was the main focus of work in the paper, and the authors tried to tackle it from many angles. They proposed a new type of loss called router z-loss, which they found to improve stability without degrading the quality of the model. They followed the traditional approach of pre-training on large data and fine-tuning downstream tasks. In the fine-tuning phase, they noticed overfitting issues on two SuperGLUE tasks. To answer this problem, they updated only a subset of model parameters during fine-tuning. The model was tested on many NLU tasks, including several NLI tasks. On RTE, the model demonstrates 93.5% accuracy, and on the R3 test set of ANLI, it exhibits an impressive 74.7% accuracy. The model currently holds first place on the leaderboard of the SuperGlue benchmark.

References

1. MacCartney, B.; Manning, C.D. An Extended Model of Natural Logic. In Proceedings of the Eight International Conference on Computational Semantics, Tilburg, The Netherlands, 7–9 January 2009; Association for Computational Linguistics: Toronto, ON, Canada, 2009; pp. 140–156.
2. de Marneffe, M.-C.; Pado, S.; Manning, C.D. Multi-word expressions in textual inference: Much ado about nothing? In Proceedings of the 2009 Workshop on Applied Textual Inference, ACL-IJCNLP 2009, Suntec, Singapore, 6 August 2009; pp. 1–9.
3. Dagan, I.; Glickman, O.; Magnini, B. The PASCAL Recognising Textual Entailment Challenge. In Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment, Proceedings of the First PASCAL Machine Learning Challenges Workshop, MLCW 2005, Southampton, UK, 11–13 April 2005; Quiñonero-Candela, J., Dagan, I., Magnini, B., d'Alché-Buc, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3944.
4. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C. A large annotated corpus for learning natural language inference. arXiv 2015, arXiv:1508.05326.
5. Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. arXiv 2018, arXiv:1804.07461.
6. Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; Bowman, S.R. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. arXiv 2020, arXiv:1905.00537.
7. Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; Inkpen, D. Enhanced LSTM for Natural Language Inference. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1657–1668.
8. Chen, Q.; Zhu, X.; Ling, Z.-H.; Inkpen, D.; Wei, S. Neural Natural Language Inference Models Enhanced with External Knowledge. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2406–2417.
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. arXiv 2017, arXiv:1706.03762, 03762.
10. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv 2019, arXiv:1810.04805.
11. Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; van Durme, B. Hypothesis only baselines in natural language inference. arXiv 2018, arXiv:1805.01042.
12. Rudinger, R.; May, C.; Van Durme, B. Social Bias in Elicited Natural Language Inferences. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, Valencia, Spain, 4 April 2017.
13. Kalouli, A.-L.; Real, L. Correcting Contradictions. Computing Natural Language Inference Workshop. 2017, pp. 1–6. Available online: <https://aclanthology.org/W17-7205.pdf> (accessed on 10 January 2023).
14. Glockner, M.; Shwartz, V.; Goldberg, Y. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia, 15–20 July 2018; pp. 650–655.
15. Nie, Y.; Williams, A.; Dinan, E.; Bansal, M.; Weston, J.; Kiela, D. Adversarial NLI: A New Benchmark for Natural Language Understanding. arXiv 2019, arXiv:1910.14599.
16. Fort, K. Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects; Wiley-ISTE: New York, NY, USA, 2016; p. 196. ISBN 978-1-84821-904-5.
17. Nie, Y.; Zhou, X.; Bansal, M. What Can We Learn from Collective Human Opinions on Natural Language Inference Data? arXiv 2020, arXiv:2010.03532.
18. Pavlick, E.; Kwiatkowski, T. Inherent Disagreements in Human Textual Inferences. Trans. Assoc. Comput. Linguistics 2019, 7, 677–694.
19. Stowe, K.; Utama, P.; Gurevych, I. IMPLI: Investigating NLI Models' Performance on Figurative Language. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Guilin, China, 24–25 September 2022; pp. 5375–5388.
20. Bowman, S.R.; Dahl, G.E. What Will it Take to Fix Benchmarking in Natural Language Understanding? arXiv 2021, arXiv:2104.02145.
21. Liu, A.; Swayamdipta, S.; Smith, N.A.; Choi, Y. WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation. arXiv 2022, arXiv:2201.05955.

22. Stacey, J.; Belinkov, Y.; Rei, M. Supervising Model Attention with Human Explanations for Robust Natural Language Inference. In Proceedings of the AAAI Conference on Artificial Intelligence, Enter Virtual Venue, 22 February–1 March 2022; Volume 36, pp. 11349–11357.
23. Parikh, A.; Täckström, O.; Das, D.; Uszkoreit, J. A Decomposable Attention Model for Natural Language Inference. arXiv 2016, arXiv:1606.01933.
24. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv 2019, arXiv:1907.11692.
25. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv 2020, arXiv:1909.11942.
26. Sun, Y.; Wang, S.; Feng, S.; Ding, S.; Pang, C.; Shang, J.; Liu, J.; Chen, X.; Zhao, Y.; Lu, Y.; et al. ERNIE 3.0: Large-scale Knowledge Enhanced Pre-training for Language Understanding and Generation. arXiv 2021, arXiv:2107.02137.
27. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling Language Modeling with Pathways. arXiv 2022, arXiv:2204.02311.
28. Zoph, B.; Bello, I.; Kumar, S.; Du, N.; Huang, Y.; Dean, J.; Shazeer, N.; Fedus, W. ST-MoE: Designing Stable and Transferable Sparse Expert Models. arXiv 2022, arXiv:2202.08906.

Retrieved from <https://encyclopedia.pub/entry/history/show/127404>