Discrete & Continuous Models
& Applied Computational Science
ISSN 2658-7149 (Online), 2658-4670 (Print)

2024, 32 (2) 234–241
http://journals.rudn.ru/miph

Research article

UDC 519.6
PACS 07.05.Tp,
DOI: 10.22363/2658-4670-2024-32-2-234–241 EDN: CUCXTY

Developing a computer system for student learning based on vision-language models

Eugeny Yu. Shchetinin[1], Anastasia G. Glushkova[2], Anastasia V. Demidova[3]

[1] *Financial University under the Government of the Russian Federation, 49 Leningradsky Ave, Moscow, 125993, Russian Federation*
[2] *Endeavor, London W4 5HR, Chiswick Park, 566 Chiswick High Road, United Kingdom*
[3] *RUDN University, 6 Miklukho-Maklaya St, Moscow, 117198, Russian Federation*

(received: November 6, 2023; revised: January 20, 2024; accepted: February 7, 2024)

Abstract. In recent years, artificial intelligence methods have been developed in various fields, particularly in education. The development of computer systems for student learning is an important task and can significantly improve student learning. The development and implementation of deep learning methods in the educational process has gained immense popularity. The most successful among them are models that consider the multimodal nature of information, in particular the combination of text, sound, images, and video. The difficulty in processing such data is that combining multimodal input data by different channel concatenation methods that ignore the heterogeneity of different modalities is an inefficient approach. To solve this problem, an inter-channel attention module is proposed in this paper. The paper presents a computer vision-linguistic system of student learning process based on the concatenation of multimodal input data using the inter-channel attention module. It is shown that the creation of effective and flexible learning systems and technologies based on such models allows to adapt the educational process to the individual needs of students and increase its efficiency.

Key words and phrases: deep learning, vision-language learning model, neural networks-transformers, through-channel attention module

Citation: Shchetinin E. Y., Glushkova A. G., Demidova A. V., Developing a computer system for student learning based on vision-language models. *Discrete and Continuous Models and Applied Computational Science* **32** (2), 234–241. doi: 10.22363/2658-4670-2024-32-2-234–241. edn: CUCXTY (2024).

1. Introduction

The history of the development and creation of computer learning models follows from the creation of models of machine translation of text, Hidden Markov chain models, then Recurrent Neural Networks (RNN) were used for this purpose for a long time [1–3]. However, as the flow of information increases, the quality of its processing using then recurrent neural networks decreased significantly, because to obtain a complete and coherent final text it is not enough to translate individual sentences, it is necessary to consider its overall context. Also, recurrent networks required sequential computations, which limited the ability to effectively use modern GPUs for model training.

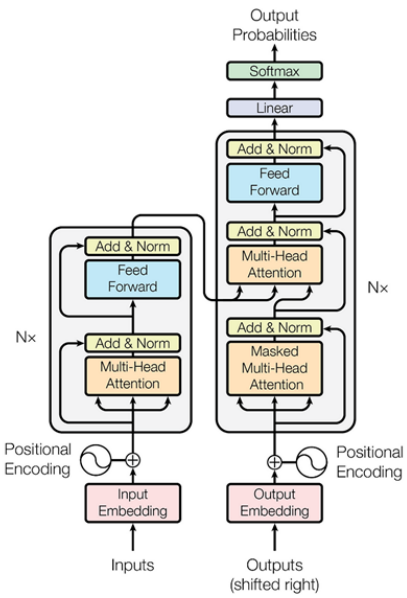

Figure 1. Transformer architecture

To fix the problem, an "attention mechanism" was developed to focus on important parts of the text [4, 5]. With its help, the neural network evaluates which position of the incoming sequence is the most important for a particular position of the sequence in the output. In the process of developing the architecture of recurrent networks, scientists from Google Research and Google Brain came up with a more technologically advanced family of deep learning architectures – transformer neural networks (transformers). It combines parallel data processing, the possibility of pre-training models and a wide application of the attention mechanism. A transformer neural network consists of two sets of layers – encoders and decoders, which contain multiple layers.

This paper presents a computer system-based vision-language model of student learning process based on combining the multimodal input data through a through-channel attention module. It is shown that the creation of effective and flexible learning systems and technologies based on such models allows to adapt the educational process to the individual needs of students and increase its efficiency.

2. Modeling the learning process using Vision-Language Models

To create a vision-language computer system of student learning, this paper used the Transformers architecture, that's most recent and successful approach in the field of natural language processing. Our model consists of several layers of transformers, each layer consisting of multiple attention mechanisms – input, internal and output. Input attention mechanisms allow the model to model and assign importance to different parts of the input data. Internal attention mechanisms allow the model to analyze the interactions between different data elements. Output attention mechanisms allow the model to generate responses based on the information received. The result is shown in Fig. 1.

The complexity and multimodality of learning data is such that instead of combining the multimodal input data through channel concatenation, which ignores the heterogeneity of different modalities, we propose a cross-channel attention module. This paper presents a computer system-based vision-language model of student learning process based on combining the multimodal input data through a through-channel attention module. Proposed model was trained on a large set of texts containing information from different subject areas. Each text is broken down into character sequences of several hundred (e.g., 40000) characters in length to feed the model. This approach allows the model to be trained on longer contexts and to consider the full amount of information when generating responses.

3. Model training

A teacher-guided learning method is used to train the model, where for each input sequence of symbols, there is a corresponding output sequence of symbols that represents the correct answer or the next step in student learning. While training the model, the input symbols are gradually fed into the model one by one, and the model generates predictions for the next symbol in the sequence. For each symbol generated, the model calculates the error and adjusts its parameters to reduce the error. This process is repeated over several epochs until the model reaches a given level of accuracy.

Advantages of a transformer-based model:

1. Flexibility: Transformers allow the modeling of long sequences of characters, which is especially important when processing texts and contexts in an educational setting. A Transformer-based model is able to capture the relationships between different data elements and generate responses that are tailored to the learning context.
2. Better context processing: Transformers allow the model to process many consecutive characters simultaneously using the attention mechanism. This allows the model to consider all previous symbols when generating a response, which is especially important in the case of student learning, where context can be critical to understanding and aiding learning.
3. Generating high quality responses: Transformers have the advantage of generating accurate and grammatically correct output sequences of symbols. This is especially important for educational tasks where accuracy and clarity of responses are critical.
4. Adaptability and Personalization: The Transformer-based model easily adjusts its parameters based on individual student needs to provide personalized learning. The model can adaptively respond to the student's proficiency level and progress, providing more appropriate and individualized responses and prompts.

Additional steps and research include determining the appropriate architecture and hyperparameters of the model, selecting the right data, optimizing training, and evaluating and validating the model using quality indicators of the original student learning. Despite the challenges, developing a Transformer-based vision-language model for student learning has great potential to improve the educational process [6]. This model can help students in getting more personalized learning, quick access to information, and effective comprehension of the material. Developing a language model for teaching students English based on Transformers can be very useful in an educational setting. Such a model can help students to master various aspects of English including grammar, vocabulary, reading, listening, and writing.

4. An example of a vision-language model for teaching students to learn English

Let's present an example of the language model of teaching students English based on transformers:

1. Problem Statement: The model was trained on a large corpus of English texts that can be collected from various sources - books, articles, blogs, etc. The goal of the model is to teach the student correct grammatical constructions, improve his/her vocabulary, develop reading and comprehension skills in English text, and help develop writing skills in English.
2. Input data: for model training, we use English text data, which we split into character sequences of several hundred (e.g., 40000) characters in length. Thus, each input sequence is a set of characters preceding a certain point in the text.
3. Model Architecture: The model consists of several layers of transformers, where each layer contains attention mechanisms. Input symbols are fed into the model sequentially, and at each step the model generates a prediction for the next symbol in the sequence.
4. Model training: The model is trained using a teacher approach, where for each input sequence of symbols, there is a corresponding correct output sequence of symbols. The model parameters are adjusted based on the error computed between the generated and correct output sequences. The goal is to minimize this error over multiple training epochs.
5. Answer generation and evaluation: after training the model, students can ask questions or provide texts in English and the model will generate answers or sentences using their knowledge of English. These responses can be evaluated based on correct grammatical constructions, lexical variety, comprehensibility and cohesion.

Such a model based on transformers can help students improve their English skills. Here are some possible examples of how to use the model:

- Generating answers to grammar questions: A student can ask a question about the correct use of a certain grammatical construction or tense, and the model can generate the correct answer with explanations and examples.
- Help with reading and comprehension of English texts: Students can provide a text to read, and the model is able to offer explanations of difficult words or phrases and provide translations or additional contextual materials for better understanding.
- Support for writing skills: Students can provide their English texts for revision and commentary. The model can offer corrections, suggestions for improving style and variation of phrases to help the student develop their writing skills.
- Exam Preparation: The model can help students prepare for English exams by providing practice in writing essays, composing narratives, answering grammar exercises, and reading and listening tests.

One practical example of using a Transformer-based model to teach students English could be to use it in online courses or educational platforms. The model can be integrated into a system where students can ask questions, write essays or do exercises and the model will provide feedback and hints to improve skills. For example, students can submit their essays for revision and the model will analyze the text, detect errors in grammar, and suggest correction options. It can point out difficult grammatical constructions and provide explanations about their usage. The model can also suggest different phrases or expressions to improve the writing style and expand the student's vocabulary.

In addition, the model can help students perceive and understand English text. By providing reading text, students can get explanations for difficult words, phrases, and expressions. The model can suggest synonyms, antonyms, or examples of how words are used in different contexts. Using the attention mechanism, the model can highlight key points in the text and help students understand its content. The model can offer supplementary materials related to topics from the text so that students can extend their knowledge and improve their comprehension.

Also, the model can offer reading and listening exercises, analyze students' responses, and provide feedback on their performance. It can help students develop listening skills, detect errors in grammar

or vocabulary by automatically checking answers. In addition to educational platforms, the model can be integrated into mobile apps that students can use to study and practice anywhere and anytime they want. This allows students to access English language learning instruction and support even outside the classroom.

One of the important functions of the computer vision-language model for English language learning is the ability to personalize learning [7, 8]. The model can consider the proficiency level and needs of each student by providing individualized recommendations and feedback. It can analyze student's mistakes and suggest appropriate material to correct them. This allows students to receive individualized instruction and focus on their weaknesses [9]. In addition, the model can have interactive dialog features, allowing students to ask questions and receive direct answers from the model. This can be useful for clarifying unclear concepts, overcoming difficulties, or getting additional explanations on topics that students find challenging.

Interactive dialog can also be used to practice speaking skills. Students can ask questions or ask the model to comment on their statements in English. This will help them practice grammar, correct pronunciation, and the ability to express themselves in English. Overall, the Transformer-based model for teaching students English can greatly enrich the learning process and improve the effectiveness of language learning. It can provide personalized support, give feedback, offer supplementary materials, and help to develop both writing, reading, listening, and speaking skills.

5. Discussion of the results

The development of a computerized language model for teaching English language learners based on Transformers is complex and requires a large amount of training data. In addition, continuous refinement and updating of the model is an important part of the process so that it can adapt to the changing needs and requirements of students. It is also important to account for the diversity of language skill levels and student needs so that the model can provide individualized support and adapt to each student. Despite the challenges in development, the Transformer-based model for teaching English language learners has significant potential. It can provide individualized and effective feedback, help students master grammar and vocabulary, understand, and analyze English texts, and develop writing skills. In the future, with further research and development, such models can become an integral part of the educational process. They can help students overcome language barriers, provide more flexible and personalized learning, and increase the effectiveness of English language learning.

Artificial intelligence extends the possibilities of the educational process from the creation of a digital university to the training of specific disciplines. Artificial intelligence becomes the basis for the creation of new means and tools for learning, in particular chatbots that can be used as simulators in the learning system. The most popular and convenient in teaching a foreign language in a higher education institution is the GPT chatbot based on a language model (GPT-3,5; GPT-4, etc. [9]). Implementation of GPT models is carried out not only by OpenAI company, but also by other developers both foreign (BERT by Google, Turing NLG by Microsoft) and domestic (Yandex company announced in February 2023 the release of ChatGPT-style – "YaLM 2.0", "Sber" company trained GPT-3 model on the Russian language corpus, creating ruGPT, widely used in the Russian-speaking field). GOOGLE has also launched its analog of the Bard chatbot, based on the principle of artificial intelligence.

This chatbot has all chances to become a valuable tool for teaching English to students. It can be used to practice speaking, grammar, vocabulary, and writing skills, learn and use cultural context, and foster group collaboration skills. With the help of technology such as GPT, English language

learning can become more accessible and engaging. However, GPT should be used in conjunction with traditional language learning methods, such as classroom instruction and language immersion, and with clear guidelines for students. When using ChatGPT in the classroom, it is important to set rules and expectations for student behavior. Students should be encouraged to use ChatGPT as a tool for learning, but not rely on it as the sole source of language learning. Teachers should also monitor students' interactions with ChatGPT to ensure that students are using it productively and that its use does not lead to the substitution of an understanding of AI-assisted language learning with the use of chatbot capabilities to enhance students' performance in their own language acquisition [10].

6. Conclusion

The development of a computer vision-language model for student learning is a powerful tool for improving the educational process. This model can analyze and understand the learning context, generating accurate and grammatically correct answers, and adapting its parameters according to individual student needs. However, it should be noted that developing and optimizing such a model is a challenging task that requires large computational power and voluminous training data. Also, this is only a general concept of developing a vision-language model based on transformer architecture for student learning. Such projects require deep knowledge in natural language processing, machine learning and neural networks [10].

Author Contributions: Conceptualization, Shchetinin E.; methodology, Shchetinin E.; software, Shchetinin E.; writing—review and editing Demidova A.; Demidova A.; supervision, Demidova A.; project administration, Glushkova A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing is not applicable.

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

References

1. Devlin, J., Chang, M., Lee, K. & K., T. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* 2018.
2. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. *Attention is All you Need* in *Advances in Neural Information Processing Systems* (eds Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S. & Garnett, R.) **30** (Curran Associates, Inc., 2017), 5998–6008.
3. Liu Y. andOtt, M., Goyal N. andDu, J., Joshi, M., Chen, D., Levy, O., Lewis M. andZettlemoyer, L. & V., S. *RoBERTa: A Robustly Optimized BERT Pretraining Approach* 2019.
4. Clark, E. & Gardner, M. *Simple and Effective Multi-Paragraph Reading Comprehension* 2018.
5. Klein, G., Kim, Y., Deng, Y., Senellart, J. & Rush, A. *OpenNMT: Open-Source Toolkit for Neural Machine Translation* in *Proceedings of ACL 2017, System Demonstrations* (eds Bansal, M. & Ji, H.) **28** (Association for Computational Linguistics, Vancouver, Canada, July 2017), 67–72. doi:10.18653/V1/P17-4012.
6. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. *Improving language understanding by generative pre-training* https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
7. Nogueira, R. & Cho, K. *Passage Re-ranking with BERT* 2019.
8. Schröder, S., Niekler, A. & Potthast, M. *Revisiting Uncertainty-based Query Strategies for Active Learning with Transformers* 2021.

9. Yang, F., Wang, X., Ma, H. & Li, J. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. *BMC Medical Informatics and Decision Making* **21,** 141–157. doi:10.1186/s12911-021-01459-0 (2021).
10. Rashid, M., Höhne, J., Schmitz, G. & Müller-Putz, G. A Review of Humanoid Robots Controlled by Brain-Computer Interfaces. *Frontiers in Neurorobotics,* 1–28 (2020).

Information about the authors

Shchetinin, Eugeny Yu.—Doctor of Physical and Mathematical Sciences, lecturer of Department of Mathematics (e-mail: riviera-molto@mail.ru, ORCID: 0000-0003-3651-7629, ResearcherID: O-8287-2017, Scopus Author ID: 16408533100)

Glushkova, Anastasia G.—researcher (e-mail: aglushkova@endeavorco.com, ORCID: 0000-0002-8285-0847, Scopus Author ID: 57485591900)

Demidova, Anastasia V.—Candidate of Physical and Mathematical Sciences, Assistant professor of Department of Probability Theory and Cyber Security (e-mail: demidova-av@rudn.ru, ORCID: 0000-0003-1000-9650)

UDC 519.6
PACS 07.05.Tp,
DOI: 10.22363/2658-4670-2024-32-2-234–241 EDN: CUCXTY

Разработка компьютерной системы обучения студентов на основе визуально-лингвистических моделей

Е. Ю. Щетинин[1], А. Г. Глушкова[2], А. В. Демидова[3]

[1] *Финансовый университет при Правительстве Российской Федерации, Ленинградский пр-т, д. 49, Москва, 125993, Российская Федерация*
[2] *Эндевор, ш. Чизвик, д. 566, Чизвик Парк, Лондон W4 5HR, Великобритания*
[3] *Российский университет дружбы народов, ул. Миклухо-Маклая, д. 6, Москва, 117198, Российская Федерация*

Аннотация. В последние годы методы искусственного интеллекта получили большое развитие в различных областях, в частности в образовании. Разработка компьютерных систем для обучения студентов является важной задачей и может значительно улучшить процесс обучения студентов. Разработка и внедрение методов глубокого обучения в образовательный процесс приобрели огромную популярность. Наиболее успешными среди них являются модели, учитывающие мультимодальный характер информации, в частности сочетание текста, звука, изображений и видео. Сложность обработки таких данных состоит в том, что объединение мультимодальных входных данных различными методами конкатенации каналов, игнорирующих неоднородность разных модальностей, является неэффективным подходом. Для решения этой проблемы в работе предложен междуканальный модуль внимания. В статье представлена компьютерная визуально-лингвистическая система процесса обучения студентов, основанная на объединении мультимодальных входных данных с использованием междуканального модуля внимания. Показано, что создание эффективных и гибких систем и технологий обучения на основе таких моделей позволяет адаптировать образовательный процесс к индивидуальным потребностям обучающихся и повысить его эффективность.

Ключевые слова: глубокое обучение, модель визуально-лингвистического обучения, нейронные сети-трансформеры, модуль сквозного внимания