


# THÔNG TIN CHUNG CỦA BÁO CÁO

- Link YouTube video của báo cáo (tối đa 5 phút):  
[22521065 - 22521156 - Báo cáo đồ án cuối kỳ CS519.O21.KHTN - YouTube](#)
- Link slides (dạng .pdf đặt trên Github):  
[CS519.O21.KHTN/CS519.O21.KHTN.DeCuong.FinalReport.AIO.Slide.pdf at main · hphuoc0906/CS519.O21.KHTN \(github.com\)](#)
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*

<ul style="list-style-type: none"><li>• Họ và Tên: Đặng Hữu Phát</li><li>• MSSV: 22521065</li></ul> 	<ul style="list-style-type: none"><li>• Lớp: CS519.O21.KHTN</li><li>• Tự đánh giá (điểm tổng kết môn): 9/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 9</li><li>• Số câu hỏi QT của nhóm: 5</li><li>• Link Github: <a href="https://github.com/hphuoc0906/CS519.O21.KHTN">https://github.com/hphuoc0906/CS519.O21.KHTN</a></li><li>• Mô tả công việc và đóng góp của các nhân cho kết quả của nhóm:<ul style="list-style-type: none"><li>○ Đóng góp ý tưởng</li><li>○ Thực hiện khảo sát về các bài báo liên quan đến đề tài của nhóm.</li><li>○ Viết nội dung đề cương.</li><li>○ Làm Powerpoint.</li></ul></li></ul>
<ul style="list-style-type: none"><li>• Họ và Tên: Phan Hoàng Phước</li><li>• MSSV: 22521156</li></ul>	<ul style="list-style-type: none"><li>• Lớp: CS519.O21.KHTN</li><li>• Tự đánh giá (điểm tổng kết môn): 9/10</li><li>• Số buổi vắng: 0</li><li>• Số câu hỏi QT cá nhân: 9</li></ul>



- Số câu hỏi QT của nhóm: 5
- Link Github:  
<https://github.com/hphuoc0906/CS519.O21.KHTN>
- Mô tả công việc và đóng góp của các nhân cho kết quả của nhóm:
  - Đóng góp ý tưởng
  - Thực hiện khảo sát về các bài báo liên quan đến đề tài của nhóm.
  - Viết nội dung đề cương.
  - Làm poster.
  - Làm video youtube.

# ĐỀ CƯƠNG NGHIÊN CỨU

## TÊN ĐỀ TÀI (IN HOA)

BỘ DỮ LIỆU ĐA THỂ THỨC TIẾNG VIỆT VỀ HIỂU BIẾT VÀ  
LUẬN LÝ KHOA HỌC TỰ NHIÊN

## TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

A MASSIVE VIETNAMESE MULTIMODAL NATURAL SCIENCES  
UNDERSTANDING AND REASONING DATASET (VMNSU)

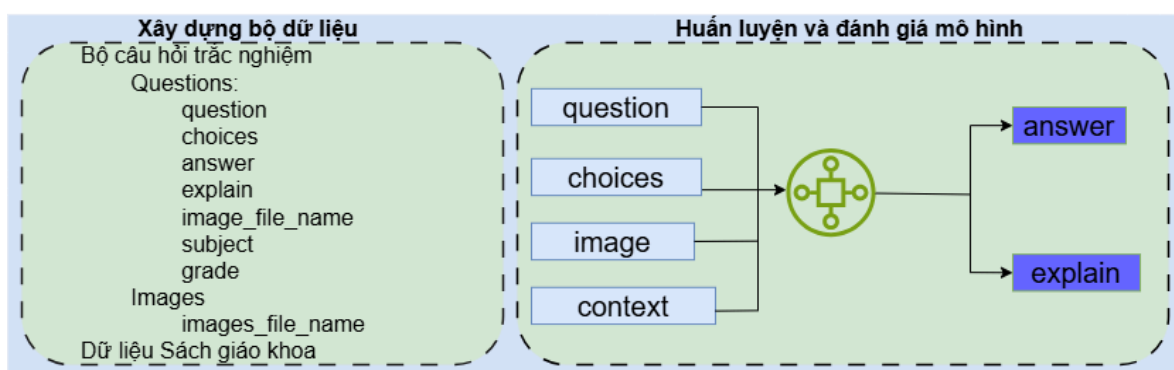
## TÓM TẮT

Bài toán VQA (Visual Question Answering) là một trong những thách thức mới, và nhận được rất nhiều sự quan tâm ở lĩnh vực multimodal trong những năm gần đây. Chính vì vậy, việc xây dựng các bộ dữ liệu đa thể thức trở nên hết sức quan trọng, đặc biệt là đối với ngôn ngữ tiếng Việt. Nhằm thúc đẩy sự phát triển của cộng đồng nghiên cứu ở Việt Nam trong lĩnh vực này, chúng tôi giới thiệu bộ dữ liệu đa thể thức tiếng Việt về hiểu biết và luận lý khoa học tự nhiên (VMNSU). Bộ dữ liệu sẽ chứa hơn 10.000 câu hỏi và nội dung sách giáo khoa trong các lĩnh vực về Toán học, Vật lý, Hóa học, Sinh học bằng tiếng Việt. Mỗi câu hỏi bao gồm dữ liệu dạng text và hình ảnh. Câu trả lời của các câu hỏi tương ứng sẽ theo dạng trắc nghiệm (multiple choice) và giải thích chi tiết từng bước để thu được kết quả cuối cùng. Dữ liệu từ sách giáo khoa sẽ được sử dụng làm cơ sở tri thức (knowledge base) hỗ trợ mô hình trong việc trả lời câu hỏi, giúp nâng cao độ chính xác, tin cậy và tính logic của các câu trả lời. Bên cạnh đó, bộ dữ liệu sẽ được dùng để đánh giá các phương pháp tiên tiến nhất hiện nay, từ đó đề xuất phương pháp, hướng tiếp cận mới để giải quyết những đặc trưng của bài toán.

## GIỚI THIỆU

Trong những năm gần đây, bài toán VQA (Visual Question Answering) đã trở thành một thách thức quan trọng và nhận được nhiều sự quan tâm trong lĩnh vực nghiên cứu trí tuệ nhân tạo. Trên thế giới, đã có một số công trình nghiên cứu lớn về chủ đề này.

Ví dụ, nhóm Xiang Yue đã phát triển Benchmark MMMU để đánh giá khả năng suy luận của các mô hình đa thể thức lớn (LMMs) và các mô hình ngôn ngữ lớn (LLMs) với hơn 11.000 câu hỏi từ các lĩnh vực như Khoa học, Y học, Nghệ thuật,... Bộ dữ liệu ScienceQA chứa hơn 20.000 câu hỏi từ các lĩnh vực Khoa học Tự nhiên, Xã hội và Ngôn Ngữ, áp dụng phương pháp Chain of Thought (CoT) trên các mô hình UnifiedQA và GPT-3 để cải thiện suy luận và giải quyết vấn đề. Benchmark ChartQA và tập kiểm tra đa nhiệm với 15.908 câu hỏi trắc nghiệm từ nhiều lĩnh vực cũng đã được phát triển. Tại Việt Nam, các nghiên cứu về mô hình đa thể thức vẫn còn rất hạn chế. Năm 2023, bộ dữ liệu OpenViVQA, do Nguyễn Hiếu Nghĩa và đồng nghiệp công bố, bao gồm hơn 11.000 ảnh với hơn 37.000 cặp câu hỏi và câu trả lời đã đánh dấu bước tiến quan trọng trong nghiên cứu ngôn ngữ và đa thể thức cho Tiếng Việt. Tuy nhiên, lĩnh vực này vẫn còn rất mới mẻ và thiếu hụt các bộ dữ liệu chất lượng cao, đặc biệt trong những lĩnh vực cụ thể như các môn khoa học tự nhiên. Để hỗ trợ sự phát triển của cộng đồng nghiên cứu tại Việt Nam trong lĩnh vực này, nghiên cứu này sẽ nghiên cứu, đóng góp bộ dữ liệu đa phương tiện bằng tiếng Việt về kiến thức và lập luận trong khoa học tự nhiên (VMNSU). Ngoài những câu hỏi trắc nghiệm đa dạng, phong phú nhiều cấp độ, bộ dữ liệu còn cung cấp nền tảng tri thức từ sách giáo khoa để hỗ trợ các nhà nghiên cứu phát triển các mô hình VQA, góp phần nâng cao độ chính xác, tin cậy và tính logic của các câu trả lời.



Hình 1. Tổng quan quá trình xây dựng và đánh giá trên tập VMNSU

## MỤC TIÊU

1. Xây dựng bộ dataset gồm các câu hỏi đa thể thức (bao gồm cả văn bản và hình

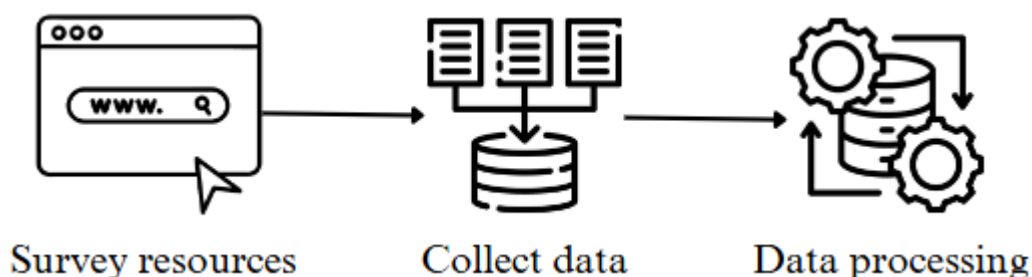
ảnh) trong lĩnh vực Toán, Vật lý, Hóa học và Sinh học theo hình thức đa lựa chọn (multiple choice); nội dung sách giáo khoa của Bộ Giáo dục và Đào tạo Việt Nam.

2. Đánh giá các phương pháp tiên tiến hiện nay trên bộ dữ liệu đã được xây dựng.
3. Nghiên cứu và đề xuất phương pháp, hướng tiếp cận mới trong lĩnh vực đa thể thức nhằm giải quyết bài toán.

## NỘI DUNG VÀ PHƯƠNG PHÁP

### Nội dung 1: Xây dựng bộ datasets.

Phương pháp thực hiện:



Hình 2. Quy trình xây dựng dataset

### Giai đoạn 1: Tìm kiếm và khảo sát nguồn tài nguyên có sẵn

Đầu tiên, để xây dựng một bộ dữ liệu đáng tin cậy và phản ánh đúng mục đích nghiên cứu, việc tìm kiếm các nguồn tài nguyên chất lượng là vô cùng cần thiết. Mục tiêu của cuộc khảo sát này là tìm kiếm và lựa chọn những nguồn dữ liệu phù hợp, dựa trên tính đa dạng và chất lượng.

- Để xây dựng bộ dữ liệu đáng tin cậy và phản ánh đúng mục đích nghiên cứu, việc tìm kiếm các nguồn dữ liệu chất lượng từ các trang web uy tín như **Vietjack**, **Tracnghiemhay**, **Tech12h**, và sách giáo khoa chính thức của Bộ Giáo Dục và Đào Tạo Việt Nam sẽ được tiến hành. Việc này nhằm đảm bảo rằng dữ liệu phản ánh sự đa dạng của kiến thức và kỹ năng cần thiết.

### Giai đoạn 2: Thu thập dữ liệu thô

Phương pháp thực hiện:

Sau khi các nguồn dữ liệu tiềm năng được xác định, kỹ thuật crawl sẽ được sử dụng

để thu thập dữ liệu từ các trang web trực tuyến. Dữ liệu trắc nghiệm sẽ được thu thập bao gồm văn bản, hình ảnh, đáp án đúng, và giải thích chi tiết cho từng câu hỏi. Đồng thời, các công cụ chuyên dụng sẽ được sử dụng để trích xuất dữ liệu từ các sách giáo khoa phiên bản PDF.

### **Giai đoạn 3: Tiền xử lý và lựa chọn dữ liệu:**

- Dữ liệu thu thập được sẽ được tiền xử lý nhằm loại bỏ dữ liệu nhiễu, không liên quan, và các bản ghi bị thiếu thông tin quan trọng. Dữ liệu trùng lặp sẽ được xóa bỏ để tránh sự chồng chéo và tối ưu hóa hiệu suất phân tích. Các dạng dữ liệu không phù hợp sẽ được chuẩn hóa bằng cách chuyển đổi định dạng và điều chỉnh các giá trị không hợp lệ. Sau cùng, dữ liệu chất lượng cao sẽ được chọn lọc dựa trên các tiêu chí như tính đầy đủ, nhất quán, và tính đại diện.

### **Nội dung 2: Huấn luyện, đánh giá các mô hình tiên tiến hiện nay trên bộ dữ liệu.**

- Thực hiện khảo sát và lựa chọn các mô hình mã nguồn mở tiên tiến trên Tiếng Anh như **OpenFlamingo2-9B**, **BLIP-2 FLAN-T5-XXL**, **InstructBLIP**, **LLaVA**,... để áp dụng vào bài toán VQA trên Tiếng Việt.

### **Đánh giá mô hình:**

- Một câu trả lời được coi là đúng nếu nó khớp với câu trả lời đúng (gold answer) cả phần trắc nghiệm và phần giải thích. Đối với phần trắc nghiệm, câu trả lời đúng là khi lựa chọn được dự đoán (A, B, C, D) trùng khớp với lựa chọn đúng. Đối với phần giải thích, câu trả lời đúng là khi câu trả lời dự đoán chứa câu trả lời đúng dưới dạng chuỗi hoặc số sau khi đã chuẩn hóa (chuyển về chữ thường và chuyển đổi số nếu có thể).
- **Độ chính xác (accuracy)** được tính bằng cách lấy tổng số câu trả lời đúng chia cho tổng số câu hỏi. Công thức cụ thể là:

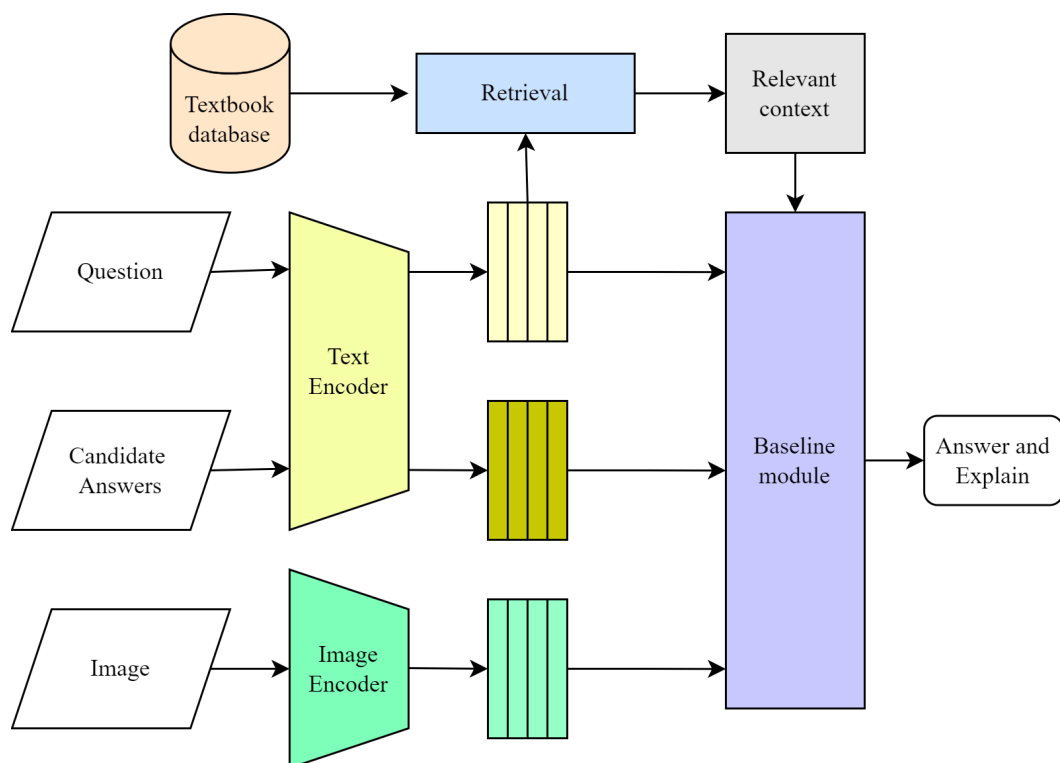
$$Accuracy = \frac{\text{Số câu trả lời đúng}}{\text{Tổng số câu hỏi}}$$

### **Nội dung 3: Đề xuất phương pháp, hướng tiếp cận mới.**

Phương pháp nghiên cứu:

Sau khi đã hoàn thành việc huấn luyện và đánh giá các phương pháp hiện có, sẽ bắt đầu bước phân tích để đề xuất hướng tiếp cận mới trên bài toán, khai thác những đặc trưng của Tiếng Việt để nâng cao hiệu suất của mô hình.

- Phân tích kết quả của các mô hình đã được huấn luyện và đánh giá để hiểu rõ hơn về những ưu điểm và hạn chế của chúng trong việc suy luận đa thể thức trên bộ dữ liệu VMNSU.
- Dựa trên những phân tích kết quả, đề xuất cải tiến cho các phương pháp hiện tại và đề xuất các phương pháp mới hoặc hướng tiếp cận mới có thể cải thiện hiệu suất của mô hình trong bài toán suy luận đa thể thức. Sử dụng text encoder như **BERT**, **XLNet** và **ViT**, **CLIP** như image encoder đưa vào model.
- Thực hiện thử nghiệm và đánh giá các phương pháp, hướng tiếp cận mới để xác định khả năng giải quyết bài toán suy luận đa thể thức và so sánh với các phương pháp hiện tại.



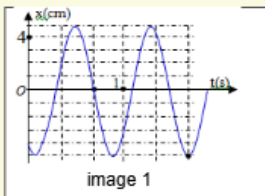
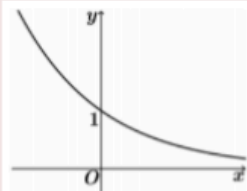
Hình 3. Mô tả tổng quan phương thức hoạt động của phương thức đề xuất

## KẾT QUẢ MONG ĐỢI

(Viết kết quả phù hợp với mục tiêu đặt ra, trên cơ sở nội dung nghiên cứu ở trên)

## 1. Bộ dữ liệu VMNSU:

Bộ dữ liệu sẽ chứa hơn 10.000 câu hỏi trong các lĩnh vực Toán học, Vật lý, Hóa học, Sinh học trên Tiếng Việt. Mỗi câu hỏi sẽ được kèm theo hình ảnh minh họa, như bảng số liệu, hình mô tả thí nghiệm, đồ thị,... giúp nâng cao tính đa thể thức và trực quan của dữ liệu. Bên cạnh đó, bộ dữ liệu cũng sẽ bao gồm các văn bản bài học được rút trích từ sách giáo khoa của Bộ Giáo dục và Đào tạo.

PHYSICS	MATH
<p><b>Question:</b> Một con lắc lò xo được treo vào một điểm cố định đang dao động điều hòa theo phương thẳng đứng. Hình bên là đồ thị biểu diễn sự phụ thuộc ly độ <math>x</math> của vật <math>m</math> theo thời gian <math>t</math>. Tần số dao động của con lắc lò xo có giá trị là</p>  <p>Choices: A. 1,5 Hz    <u>B. 1,25 Hz</u>    C. 0,5 Hz    D. 0,8 Hz</p> <p><b>Explain:</b> Mỗi ô có khoảng thời gian là <math>1/3</math> s Từ đồ thị ta có 3 ô (từ ô thứ 2 đến ô thứ 5 có <math>5T/4 = 1</math> s): <math>\frac{5T}{4} = 1s \Rightarrow T = 0,8s \Rightarrow f = \frac{1}{T} = \frac{1}{0,8} = 1,25Hz</math>. <math>\Rightarrow</math> Chọn đáp án B</p> <p><b>Subject:</b> Vật lý <b>Grade:</b> Thi thử THPT QG</p>	<p><b>Question:</b> Hàm số nào sau đây mà đồ thị có dạng như hình vẽ bên dưới?</p>  <p>Choices: A. <math>y = \ln x</math> B. <math>y = (\sqrt{2})^x</math> <u>C. <math>y = (\frac{1}{e})^x</math></u> D. <math>\log_{\frac{1}{e}} x</math></p> <p><b>Explain:</b> Dựa vào đồ thị ta thấy hàm số xác định trên <math>\mathbb{R}</math> nên loại đáp án A, D. Lại có: Đồ thị hàm số nghịch biến trên <math>\mathbb{R}</math> nên chọn đáp án C.</p> <p><b>Subject:</b> Toán <b>Grade:</b> Thi thử THPT Quốc gia</p>

Hình 4. Ví dụ một số mẫu trong tập dữ liệu thu được.

## 2. Kết quả các mô hình baseline dùng để đánh giá

Huấn luyện đánh giá các mô hình baseline với độ chính xác trên 30%. Từ đó đánh giá khả năng giải quyết bài toán suy luận đa thể thức của các mô hình trên Tiếng Việt.

## 3. Phương pháp, hướng tiếp cận mới

Phân tích kết quả và đề xuất: Dựa trên kết quả thử nghiệm, phân tích những ưu điểm và hạn chế của các phương pháp và mô hình đã thử nghiệm. Đề xuất hướng phát triển tiếp theo, có thể là cải thiện các phương pháp hiện tại hoặc đề xuất các phương pháp mới để cải thiện hiệu suất trong bài toán suy luận đa thể thức trong Tiếng Việt.

## TÀI LIỆU THAM KHẢO (Định dạng DBLP)

[1]. Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruofei Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu,



Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, Wenhua Chen:

MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. CoRR abs/2311.16502 (2023)

[2] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, Enamul Hoque: ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. ACL (Findings) 2022: 2263-2279

[3] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, Ashwin Kalyan:

Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. NeurIPS 2022

[4]. Nghia Hieu Nguyen, Duong T. D. Vo, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen: OpenViVQA: Task, dataset, and multimodal fusion models for visual question answering in Vietnamese. Inf. Fusion 100: 101868 (2023)

[5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, Ludwig Schmidt:

OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. CoRR abs/2308.01390 (2023)

[6] Junnan Li, Dongxu Li, Silvio Savarese, Steven C. H. Hoi:

BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023: 19730-19742

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, Steven C. H. Hoi:

InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. NeurIPS 2023

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee:

Visual Instruction Tuning. NeurIPS 2023

- [9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186
- [10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, Quoc V. Le: XLNet: Generalized Autoregressive Pretraining for Language Understanding. NeurIPS 2019: 5754-5764
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR 2021
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021: 8748-8763