

UNIVERSIDADE FEDERAL DE GOIÁS / INSTITUTO DE INFORMÁTICA

Exploração de Modelos Multimodais de Linguagem

Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua Portuguesa

Elisa Ayumi Masasi de Oliveira



UFG

UNIVERSIDADE
FEDERAL DE GOIÁS

UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)

ELISA AYUMI MASASI DE OLIVEIRA

Exploração de Modelos Multimodais de Linguagem

Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua Portuguesa

Goiânia
2025



UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TRABALHO DE CONCLUSÃO DE CURSO DE GRADUAÇÃO NO REPOSITÓRIO INSTITUCIONAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio do Repositório Institucional (RI/UFG), regulamentado pela Resolução CEPEC no 1240/2014, sem ressarcimento dos direitos autorais, de acordo com a Lei no 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo dos Trabalhos de Conclusão dos Cursos de Graduação disponibilizado no RI/UFG é de responsabilidade exclusiva dos autores. Ao encaminhar(em) o produto final, o(s) autor(a)(es)(as) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do Trabalho de Conclusão de Curso de Graduação (TCCG)

Nome(s) completo(s) do(a)(s) autor(a)(es)(as): ELISA AYUMI MASASI DE OLIVEIRA

Título do trabalho: Exploração de Modelos Multimodais de Linguagem

Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua Portuguesa

2. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador) Concorda com a liberação total do documento [X] SIM [] NÃO¹

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante: a) consulta ao(à)(s) autor(a)(es)(as) e ao(à) orientador(a); b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo do TCCG. O documento não será disponibilizado durante o período de embargo.

Casos de embargo:

- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro.

Obs.: Este termo deve ser assinado no SEI pelo orientador e pelo autor.



Documento assinado eletronicamente por **Elisa Ayumi Masasi De Oliveira, Discente**, em 10/01/2025, às 21:12, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Fernando Marques Federson, Professor do Magistério Superior**, em 15/01/2025, às 16:11, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5089548** e o código CRC **04792E93**.

Referência: Processo nº 23070.001553/2025-18

SEI nº 5089548

ELISA AYUMI MASASI DE OLIVEIRA

Exploração de Modelos Multimodais de Linguagem

Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua Portuguesa

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Orientador: Prof. Dr. Fernando Marques Federson

Goiânia

2025

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

OLIVEIRA, ELISA AYUMI MASASI DE

Exploração de Modelos Multimodais de Linguagem [manuscrito] :
Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua
Portuguesa / ELISA AYUMI MASASI DE OLIVEIRA. - 2025.
53 f.

Orientador: Prof. Dr. Fernando Marques Federson.
Trabalho de Conclusão de Curso (Graduação) - Universidade
Federal de Goiás, Instituto de Informática (INF), Inteligência
Artificial, Goiânia, 2025.

1. inteligência artificial. 2. modelos de linguagem. 3. multimodal. I.
Federson, Fernando Marques , orient. II. Título.

CDU 004

ELISA AYUMI MASASI DE OLIVEIRA

Exploração de Modelos Multimodais de Linguagem

Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua Portuguesa

Relatório final de Trabalho de Conclusão de Curso, apresentado à Universidade Federal de Goiás, como parte das exigências para a obtenção do título de Bacharel em Inteligência Artificial.

Data da Aprovação: 17 de dezembro de 2024.



Prof. Dr. Fernando Marques Federson
Orientador (INF-UFG)



Prof. Dr. Aldo André Díaz Salazar
Coordenador de TCC do BIA (INF-UFG)



Prof. Dr. Anderson da Silva Soares
Coordenador do BIA (INF-UFG)



Prof. Dr. Arlindo Rodrigues Galvão Filho
(INF-UFG)

ELISA AYUMI MASASI DE OLIVEIRA

Exploração de Modelos Multimodais de Linguagem

Avanços na Pesquisa e Desenvolvimento de MLLMs voltados à Língua Portuguesa

RESUMO

Este Relatório de Conclusão de Curso tem como objetivo reunir os resultados da minha jornada para me tornar um especialista em **Modelos de Linguagem Multimodal de Grande Escala (MLLMs)**. Uma ilustração e sua narrativa descrevem os períodos de trabalho. Os Apêndices contêm os Termos de Aceite de Entrega e os resultados obtidos durante cada período de trabalho.

Palavras-chave: inteligência artificial, modelos grandes de linguagem, geração automática de datasets.

ABSTRACT

This Course Completion Report aims to bring together the results of my journey to become an expert in **Large-Scale Multimodal Language Models (MLLMs)**. An illustration and its narrative describe the work periods. The Appendices contain the Delivery Acceptance Terms and the results obtained during each work period.

Keywords: artificial intelligence, large language models, automatic dataset generation.

Goiânia

2025

Minha Jornada



Semanas 1-3

Estudo sobre a evolução, arquitetura e o "estado da arte" dos MLLMs

Semanas 4-5

Comparação da performance, testes práticos e identificação de pontos críticos

Semanas 6-7

Definição do Modelo, tradução dos Datasets e participação na competição SemEval

Semanas 8-10

Treinamento e avaliação do Modelo

Elisa Ayumi Masasi de Oliveira

Especialista em: Modelos de Linguagem Multimodal de Grande Escala (MLLMs)

MINHA JORNADA

Nome: Elisa Ayumi Masasi de Oliveira

Especialidade: Modelos de Linguagem Multimodal de Grande Escala (MLLMs)

Objetivo deste documento

Durante o processo da disciplina Residência em IA, foram alcançados diversos resultados no desenvolvimento de um Modelo de Linguagem Multimodal de Grande Escala adaptado para o português. A cada semana, os avanços realizados foram formalizados por um Termo de Aceite de Entrega e avaliados por uma banca, considerando o planejado e o realizado para o período. Este documento tem como objetivo sintetizar os resultados obtidos, detalhando as etapas do estudo das arquiteturas multimodais, a adaptação dos datasets, o treinamento do modelo e sua validação, além de referenciar os Termos de Aceite de Entrega e seus documentos associados.

Minha Jornada

Minha jornada iniciou na **Semana 1** com a exploração da evolução e das arquiteturas dos Modelos de Linguagem Multimodal de Grande Escala (MLLMs). Baseando-me na survey "*The Evolution of Multimodal Model Architectures*", estudei os tipos *Standard Cross-Attention based Deep Fusion*, *Custom Layer based Deep Fusion*, *Non-Tokenized Early Fusion* e *Tokenized Early Fusion*, compreendendo suas vantagens, limitações e potenciais aplicações. Já na **Semana 2**, aprofundei-me nas principais limitações dessas arquiteturas e explorei o estado da arte dos modelos correspondentes. Este estudo destacou como desafios como escalabilidade e controle de fusão afetam o desempenho e a aplicabilidade prática dos modelos, conforme registrado no material disponibilizado no **Apêndice 1**. Na **Semana 3**, investiguei a performance dos modelos principais dentro dessas arquiteturas, analisando aspectos como eficiência computacional e aplicabilidade em tarefas específicas. Esses avanços foram fundamentais para a compreensão e seleção de métodos apropriados para integração multimodal.

Na **Semana 4**, continuei explorando a performance de diferentes arquiteturas de MLLMs. Realizei análises comparativas dos modelos, focando em benchmarks de tarefas multimodais como *captioning*, *VQA* e geração de texto. Esses estudos foram registrados e ampliados no **Apêndice 2**, permitindo identificar características distintas entre eficiência e versatilidade. Na **Semana 5**, conduzi testes práticos com os modelos VL-T5 (*Standard Cross-Attention based Deep Fusion*) e LLaVA (*Non-Tokenized Early Fusion*), avaliando seu desempenho em tarefas reais. Esses testes consolidaram a decisão de priorizar o modelo LLaVA para projetos futuros, devido ao seu desempenho em interações multimodais e seguimento de instruções visuais, com um estudo voltado ao foco em possíveis melhorias e adaptações.

Nas **Semanas 6 e 7**, foram traduzidos os datasets necessários para o projeto, utilizando a API do Google Tradutor, com foco estratégico em conjuntos essenciais, como dados conversacionais e complexos, cujos detalhes estão disponíveis no **Apêndice 3**. Paralelamente, foi definido o objetivo principal do projeto: implementar um MLLM em português, utilizando o LLaVA como modelo de referência. Além disso, explorei técnicas de data augmentation para dados idiomáticos, baseando-me na survey “*A Survey on Data Synthesis and Augmentation for Large Models*”. Esses esforços consolidaram a base para os experimentos de further pre-training e para o estudo do fluxo de avaliação na *task* SemEval, permitindo avaliar a eficácia e a adequação dos dados ao modelo multilíngue.

Na **Semana 8**, finalizei a tradução dos datasets e conduzi uma avaliação qualitativa inicial das amostras, verificando sua adequação para o treinamento do modelo. Paralelamente, iniciei a preparação para o treinamento, revisando artigos relacionados para aprimorar as etapas de avaliação futura. Na **Semana 9**, após os ajustes necessários, dei início ao fine-tuning do modelo LLaVA 13B. Durante os testes, identifiquei limitações no hardware inicial e realizei a migração para uma GPU NVIDIA A100, garantindo maior eficiência computacional. Experimentos preliminares foram conduzidos com datasets simplificados, consolidando o fluxo que seria aplicado aos conjuntos traduzidos. Na **Semana 10**, o treinamento completo do modelo foi concluído com sucesso. Avaliações qualitativas confirmaram a capacidade do modelo em lidar com dados em português e gerar resultados satisfatórios em tarefas multimodais. As amostras geradas podem ser consultadas no **Apêndice 4**, e o modelo também está disponível para testes na plataforma Replicate.

Concluo que este projeto foi uma jornada significativa de aprendizado e desenvolvimento, resultando na adaptação bem-sucedida de um modelo MLLM para o português, utilizando o LLaVA como referência. A partir da análise inicial das arquiteturas até o treinamento final, cada etapa contribuiu para a construção de um modelo funcional e eficiente, capaz de lidar com tarefas multimodais complexas em um cenário multilíngue. A tradução estratégica de datasets e a implementação foram fundamentais para alcançar os resultados esperados. Esse trabalho não apenas demonstra o potencial de grandes modelos de linguagem adaptados ao português, mas também abre caminhos para futuras aplicações em tarefas que integram visão e linguagem.

APÊNDICE 1

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 18 de set. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Elisa Ayumi Masasi de Oliveira

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Nessa semana foi realizado um estudo sobre a evolução e arquitetura de Multimodal Large Language Models: [📄 180924 Estudos Residência - Elisa](#)

- Arquitetura e desenvolvimento dos modelos multimodais tipo A, B, C e D;
- Estudo inicial sobre os modelos propostos como próxima geração (Mamba e VL-Mamba Cobra)

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima semana, planejo:

- Realizar um estudo sobre as principais limitações enfrentadas pelos MLLMs;
- Estudo sobre o estado da arte.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

180924 Estudos - Arquiteturas.docx citado no Termo de Aceite de Entrega de 18 de setembro.

Estudos - Arquiteturas

Estudo sobre a evolução e arquitetura de Multimodal Large Language Models, e nos principais modelos atuais.

Baseado na survey: [The Evolution of Multimodal Model Architectures](#)

The Evolution of Multimodal Model Architectures

Shakti N. Wadekar
Purdue University
swadekar@purdue.edu

Abhishek Chaurasia
Chaos Industries Inc.
abhi@choasinc.com

Aman Chadha
Stanford; Amazon*
hi@aman.ai

Eugenio Culurciello
Purdue University
euge@purdue.edu

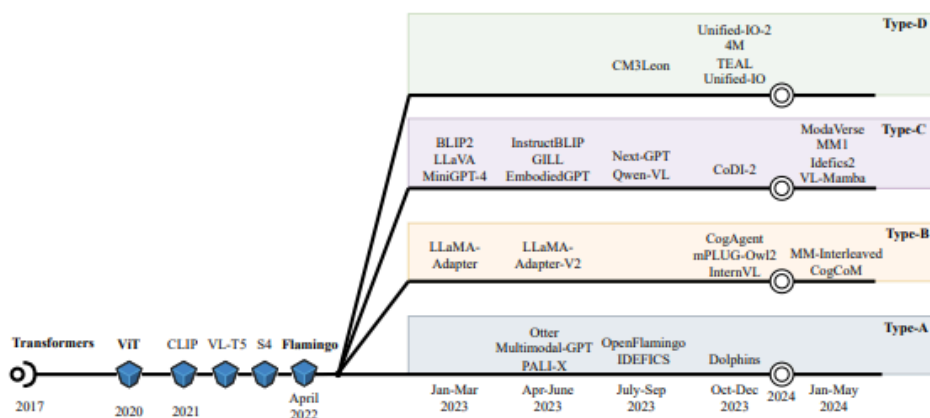
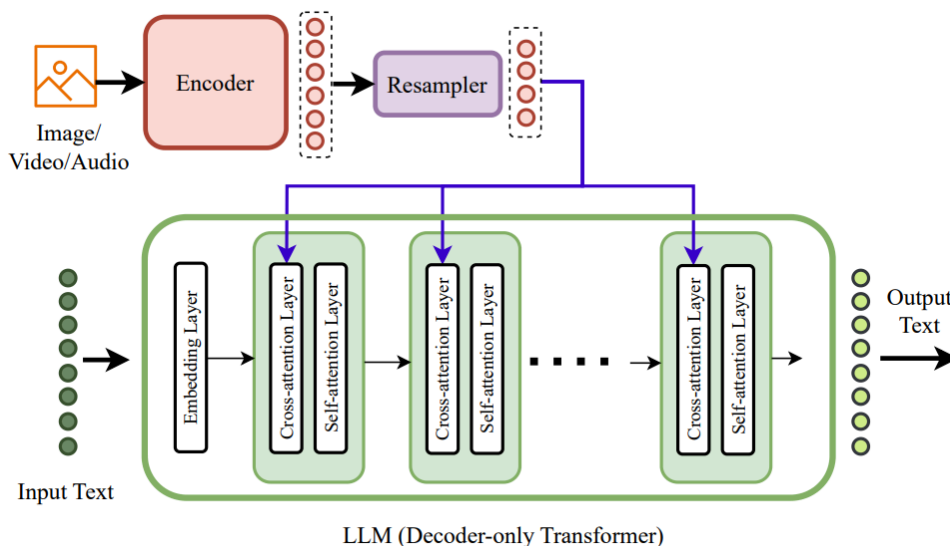


Figure 1: Development timeline of Multimodal models grouped in four proposed architecture types.

A evolução dos modelos multimodais tem sido marcada pelo desenvolvimento de arquiteturas que buscam integrar múltiplas modalidades, como texto, imagem, áudio e vídeo, de maneira eficiente e eficaz. A partir do documento, podemos identificar essa evolução através de quatro tipos principais de arquiteturas:

Arquiteturas:

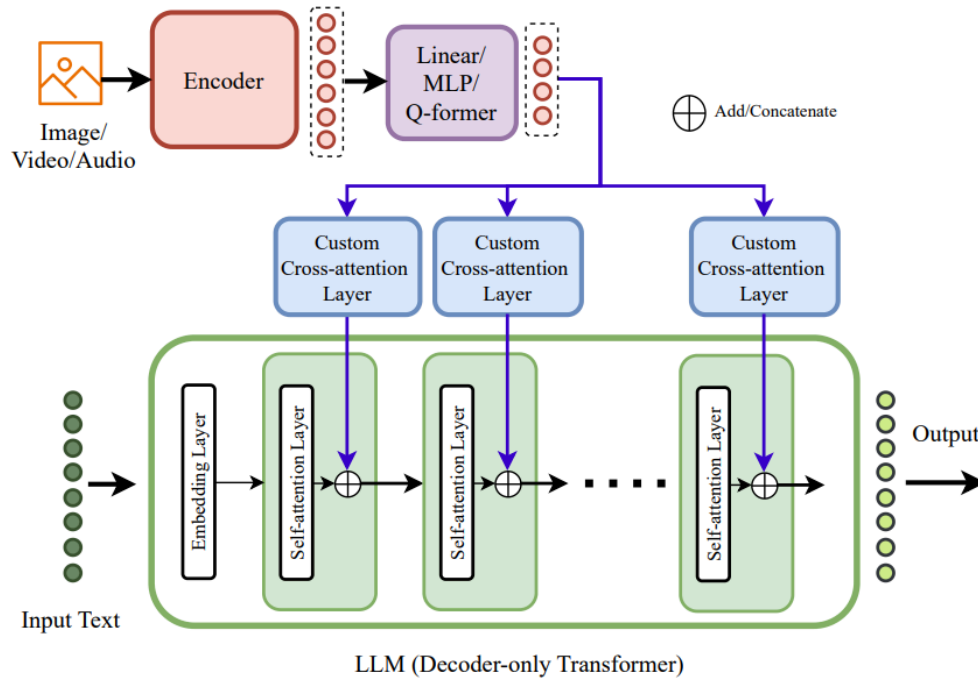
Type-A: Standard Cross-Attention based Deep Fusion (SCDF)



A arquitetura Tipo-A é caracterizada por integrar entradas multimodais nas camadas internas de um LLM pré treinado usando camadas de atenção cruzada padrão. Cada modalidade é processada por codificadores específicos e, em seguida, uma camada de resampling ajusta o número de tokens para se alinhar com as necessidades do LLM. Essa integração ocorre profundamente dentro do modelo, permitindo um controle refinado sobre a forma como as modalidades são combinadas. Modelos como Flamingo e OpenFlamingo ilustram essa abordagem, sendo pioneiros em integrar imagens e texto. Embora esse tipo de arquitetura ofereça uma fusão rica e detalhada, exige grandes quantidades de dados e recursos computacionais, tornando-se difícil de escalar e adaptar para adicionar novas modalidades, não sendo frequente em modelos mais modernos.

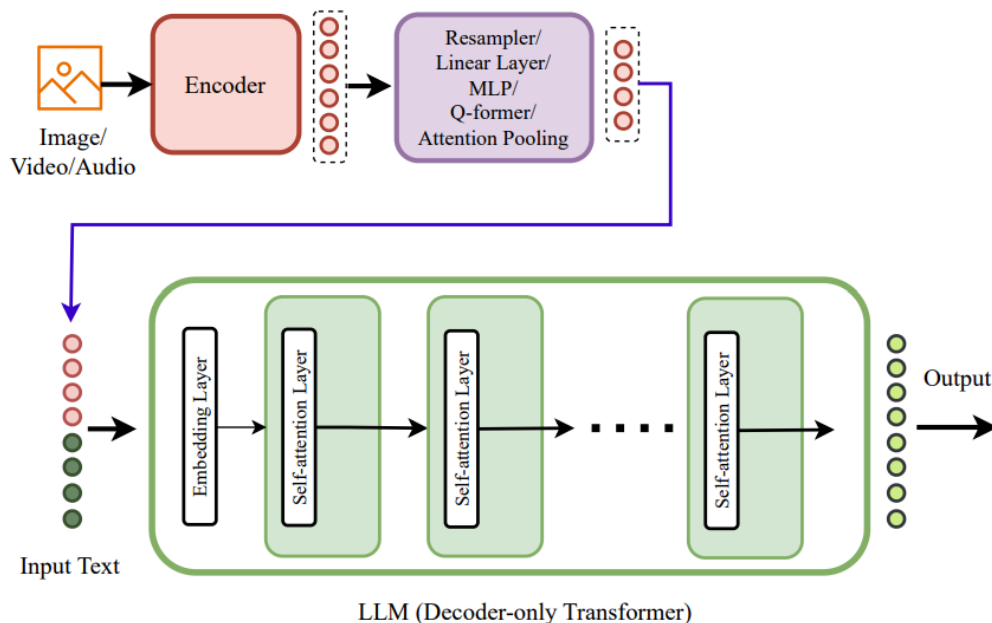
- Modelos pré treinados em next-text-token prediction e fine tuning contrastivo.

Type-B: Custom Layer based Deep Fusion (CLDF)



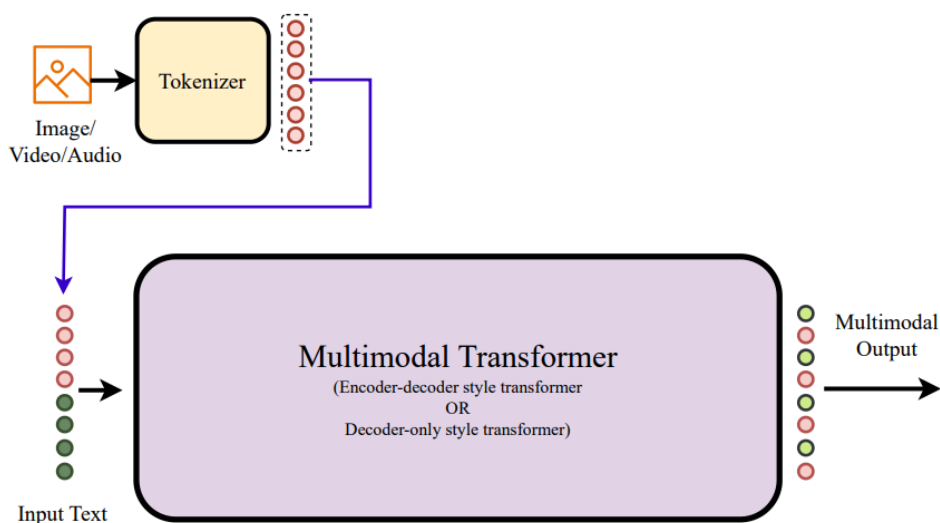
O Tipo-B também utiliza uma abordagem de deep fusion, mas diferencia-se do Tipo-A por incorporar camadas personalizadas para integrar as modalidades. Essas camadas customizadas permitem um controle ainda maior sobre como as modalidades são fundidas, ajustando especificamente o processo de integração para cada tarefa. Exemplos incluem LLaMA-Adapter e CogVLM, que utilizam camadas personalizadas de atenção cruzada para integrar informações visuais e textuais. Este tipo é mais eficiente e escalável do que o Tipo-A devido à natureza personalizável das camadas de conexão, mas ainda exige um conhecimento profundo das camadas internas do LLM. A arquitetura é mais flexível para adicionar novas modalidades, mas a complexidade de projetar camadas customizadas pode limitar sua aplicação.

Type-C: Non-Tokenized Early Fusion (NTEF)



O Tipo-C é a abordagem mais amplamente adotada, focando na fusão precoce de modalidades diretamente na entrada do modelo. Em vez de modificar as camadas internas do LLM, utiliza conectores como camadas lineares, MLPs ou Q-formers para integrar as saídas dos codificadores de modalidade. Isso simplifica a construção e o treinamento, tornando-o eficiente em termos de dados e recursos computacionais. Modelos como LLaVA, BLIP-2 e Qwen-VL exemplificam essa abordagem, empregando diferentes tipos de conectores para integrar imagens e texto. No entanto, essa simplicidade vem à custa de um controle menos refinado sobre a fusão das modalidades, o que pode não ser adequado para tarefas que exigem uma integração mais complexa.

Type-D: Tokenized Early Fusion (TEF)



A arquitetura Tipo-D diferencia-se ao tokenizar todas as modalidades, permitindo que sejam processadas como uma sequência unificada de tokens pelo LLM. Essa tokenização possibilita o uso de um objetivo de treinamento autorregressivo comum para todas as modalidades, tornando o modelo capaz de gerar saídas multimodais, como imagens e áudio, além de texto. Modelos como LaVIT e CM3Leon ilustram essa abordagem, utilizando tokenizadores específicos para converter imagens em tokens. Embora essa abordagem ofereça uma capacidade poderosa de geração multimodal, ela requer a criação de tokenizadores específicos para cada modalidade e um pré-treinamento significativo para alcançar um desempenho eficaz.

Conclusão

A evolução dos modelos multimodais é marcada pela transição de fusões profundas dentro das camadas internas dos modelos (Tipo-A e Tipo-B) para fusões mais precoces e eficientes (Tipo-C e Tipo-D). Enquanto os primeiros tipos focavam em integrar modalidades de forma mais detalhada dentro das estruturas dos modelos, as abordagens mais recentes se concentram em métodos eficientes e escaláveis para facilitar a integração de diferentes modalidades, permitindo uma aplicação mais ampla e versátil dos modelos multimodais.

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 26 de set. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Elisa Ayumi Masasi de Oliveira

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Nesta semana realizei estudos quanto às principais limitações enfrentadas pelas arquiteturas multimodais exploradas na semana anterior: [260924 Estudos - Limitações & SOTA](#)

- Type-A: Standard Cross-Attention based Deep Fusion (SCDF)
- Type-B: Custom Layer based Deep Fusion (CLDF)
- Type-C: Non-Tokenized Early Fusion (NTEF)
- Type-D: Tokenized Early Fusion (TEF)

Além disso, busquei estudar o estado da arte em MLLMs de cada arquitetura proposta, o que pode também ser verificado no documento supracitado.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Exploração da performance dos principais modelos estudados nesta semana.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go! ▾

260924 Estudos - Limitações & SOTA.docx citado no Termo de Aceite de Entrega de 26 de setembro

Estudos - Limitações & SOTA

Estudo sobre as principais limitações e os modelos estado da arte (SOTA) das arquiteturas exploradas.

180924 Estudos - Arquiteturas

Baseado na survey: [The Evolution of Multimodal Model Architectures](#)

Type-A: Standard Cross-Attention based Deep Fusion (SCDF)

- **Vantagens:**
 - Permite um controle detalhado sobre como as informações de diferentes modalidades fluem pelo modelo.
 - É totalmente treinável de ponta a ponta e utiliza camadas padrão de transformers, facilitando a fusão das modalidades (de dados).
 - Utiliza camadas de cross-attention padrão para a integração das modalidades.
- **Limitações:**
 - Necessita de grandes conjuntos de dados de treinamento e recursos computacionais significativos, tornando o processo complexo e difícil de escalar.
 - É desafiador adicionar novas modalidades depois da configuração inicial.
 - A complexidade aumenta significativamente ao aplicar um objetivo de treinamento autorregressivo além do texto.
- **SOTA:**
 - **Flamingo** e sua versão open-source **OpenFlamingo**
 - **Multimodal-GPT** (derivado do OpenFlamingo)
 - **VL-BART** e **VL-T5**

Type-B: Custom Layer based Deep Fusion (CLDF)

- **Vantagens:**
 - Similar ao Tipo-A, permite controle detalhado, mas usa camadas personalizadas para uma fusão de modalidades mais eficientes.
 - Mais eficiente computacionalmente em comparação ao Tipo-A, necessitando de menos dados de treinamento.
 - Escalável devido à natureza personalizável e eficiência computacional das camadas conectoras customizadas.
 - Adicionar novas modalidades é mais fácil devido ao mecanismo de controle (gating), que permite a adição direta de modalidades de entrada ao LLM.
- **Limitações:**

- Apesar de ser mais escalável que o Tipo-A, ainda enfrenta desafios quando comparado ao Tipo-C.
- Complexidade na implementação de objetivos autorregressivos para diversas modalidades.
- **SOTA:**
 - **LLaMA-Adapter e LLaMA-Adapter-V2**
 - **CogVLM**
 - **CogAgent**
 - **MoE-LLaVA**

Type-C: Non-Tokenized Early Fusion (NTEF)

- **Vantagens:**
 - Arquitetura modular que facilita a construção e treinamento de modelos de forma eficiente.
 - Requer menos dados de treinamento e recursos computacionais em comparação com os Tipos A, B e D.
 - Não exige conhecimento detalhado das camadas internas do LLM, apenas da interface para integração.
 - Altamente escalável devido ao design modular.
 - Adição de modalidades simplificada através de camadas aprendíveis como Linear/MLP/Q-former ou camadas customizadas.
 - Menor número de parâmetros treináveis comparado a outros tipos, tornando-o eficiente em termos computacionais para treinamento.
- **Limitações:**
 - Diferentes entradas de modalidades são fundidas apenas na entrada do decodificador, o que pode gerar problemas ao treinar com um objetivo autorregressivo para modalidades não textuais.
 - Dificuldade ou incapacidade de controlar como as informações de diferentes modalidades (como texto, imagem ou áudio) interagem e se combinam ao longo das camadas do modelo.
- **SOTA:**
 - **LLaVA e suas variantes: LLaVA-Med, LLaVAR, LLaVA-1.5, LLaVA-Phi, LLaVA-NeXT**
 - **PaLM-E**
 - **MiniGPT-v2**
 - **DetGPT**
 - **PandaGPT**

Type-D: Tokenized Early Fusion (TEF)

- **Vantagens:**

- Arquitetura simplificada devido à tokenização das modalidades de entrada e saída.
- Todas as modalidades são treinadas usando um objetivo autorregressivo padrão, o que é vantajoso para a integração de modalidades diversas.
- **Limitações:**
 - Desafios na criação de um tokenizador universal ou de tokenizadores específicos para cada modalidade.
 - Necessita de grandes conjuntos de dados de treinamento e recursos computacionais substanciais, de forma similar ao Tipo-A.
 - Assim como o Tipo-C, carece de controle detalhado sobre o fluxo de informações de modalidades.
 - As entradas de modalidades são fundidas diretamente na entrada do modelo principal, sem processamento adicional nas camadas internas.
- **SOTA:**
 - **LaVIT**
 - **TEAL**
 - **CM3Leon**
 - **VL-GPT**

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 3 de out. de 2024





Participantes da Entrega [matriculados em Residência em IA]:

Elisa Ayumi Masasi de Oliveira

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Comparação da performance e características dos principais modelos das arquiteturas de Multimodal Large Language Models estudadas anteriormente:

031024 Estudos - Performance Modelos

- Type-A: Standard Cross-Attention based Deep Fusion (SCDF)
 - Flamingo  Flamingo.pdf
 - VL-T5  VL-T5.pdf
- Type-B: Custom Layer based Deep Fusion (CLDF)
 - LLaMA-Adapter-V2  LLaMA-Adapter-V2.pdf
 - CogVLM  CogVLM.pdf

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Continuar o estudo da performance dos principais modelos das arquiteturas:

- Type-C: Non-Tokenized Early Fusion (NTEF)
- Type-D: Tokenized Early Fusion (TEF)

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: 

APÊNDICE 2

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 9 de out. de 2024





Participantes da Entrega [matriculados em Residência em IA]:

Elisa Ayumi Masasi de Oliveira

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Continuação da comparação da performance e características dos principais modelos das arquiteturas de Multimodal Large Language Models estudadas anteriormente:

03-091024 Estudos - Performance Modelos (o documento utilizado na semana anterior foi incrementado)

- Type-C: Non-Tokenized Early Fusion (NTEF)
 - LLaVA  LLaVA.pdf
 - PaLM-E  PaLM-E.pdf
- Type-D: Tokenized Early Fusion (TEF)
 - LaVIT  LaVIT.pdf
 - CM3 Leon  CM3Leon.pdf

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima semana, seguindo os estudos das arquiteturas, o objetivo será:

- Realizar testes com os modelos estudados, analisando suas limitações e pontos fortes conforme discutidos nos artigos correspondentes.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: 

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 17 de out. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Elisa Ayumi Masasi de Oliveira

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Nessa semana, assim como planejado, foram realizados testes práticos em alguns modelos das arquiteturas:

03-09-161024 Estudos - Performance Modelos

(o documento foi incrementado com a sessão de Testes)

- Type-A: Standard Cross-Attention based Deep Fusion (SCDF)
 - VL-T5
- Type-C: Non-Tokenized Early Fusion (NTEF)
 - LLaVA

Sendo possível observar o desempenho individual e comparado entre os modelos, dessa forma, orientando a decisão de utilizar o LLaVA para a exploração de trabalhos futuros.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima semana, planejo explorar a arquitetura do modelo LLaVA, com o objetivo de identificar pontos de melhoria. Pretendo focar em entender as fraquezas do modelo atual e, com base nessas observações, considerar ajustes ou otimizações.

Além disso, estou avaliando a possibilidade de implementar algo no cenário imersivo, a fim de aprimorar a eficácia e a capacidade de adaptação do modelo, buscando integração com projetos pessoais anteriores.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: Go! ▾

03-09-161024 Estudos - Performance Modelos.docx citado nos Termos de Aceite de Entrega de 03, 09 e 17 de outubro.

Estudos - Performance dos Modelos

Estudo sobre a performance dos principais modelos de cada uma das arquiteturas exploradas.

180924 Estudos - Arquiteturas

260924 Estudos - Limitações & SOTA

Type-A: Standard Cross-Attention based Deep Fusion (SCDF)

Comparação entre os modelos **Flamingo** e **VL-T5**:

Análise do **Flamingo**:

- O Flamingo integra um codificador de visão com um modelo de linguagem pré-treinado, intercalando processamento de texto e dados visuais, usando camadas de cross-attention para conectar as modalidades.
- Aprendizado com Poucos Exemplos: Capaz de se adaptar rapidamente a novas tarefas usando apenas um pequeno conjunto de exemplos, superando modelos ajustados especificamente para determinadas tarefas em vários benchmarks.
- Processamento Flexível de Entradas: Lida com sequências intercaladas de texto e imagens ou vídeos de forma arbitrária, se tornando adequado para tarefas que envolvem múltiplas imagens ou diálogos visuais (como foi comentado com o professor Cedric, no último gate).
- Desempenho de Última Geração: Alcança resultados superiores em diversos benchmarks com poucos exemplos específicos da tarefa e desempenha bem em tarefas abertas de visão e linguagem.


Análise do **VL-T5**:


- O VL-T5 utiliza T5/BART como backbone e adiciona embeddings visuais para lidar com tarefas de visão e linguagem por meio da geração de texto.
- Arquitetura Unificada: Utiliza o mesmo modelo para diferentes tarefas de visão e linguagem, facilitando o aprendizado multitarefa.
- Foco na Geração de Texto: Ao enquadrar todas as tarefas como problemas de geração de texto, aproveita os modelos de linguagem pré-treinados para lidar com a diversidade de saídas.
- Generalização: Demonstra melhor desempenho em questões com respostas raras, indicando uma forte capacidade de generalização.

Comparação do desempenho:

- VL-T5 tem como objetivo uma abordagem unificada e lida com diversas tarefas de visão e linguagem de forma eficaz, sem arquiteturas específicas para cada tarefa. Ele atinge um desempenho comparável aos modelos estado da arte em vários benchmarks de visão e linguagem.
- Flamingo destaca-se no aprendizado com poucos exemplos, permitindo rápida adaptação a novas tarefas com uma quantidade mínima de exemplos, superando modelos estado da arte ajustados em diversos casos. Ele mostra resultados sólidos em tarefas baseadas tanto em imagem quanto em vídeo com menos dados de treinamento.

Estudo realizado com base nos seguintes materiais:

 Flamingo.pdf

 VL-T5.pdf

Testes VL-T5

Inputs utilizados:

“describe the image”



Output:

city

Inputs utilizados:

“describe the image”

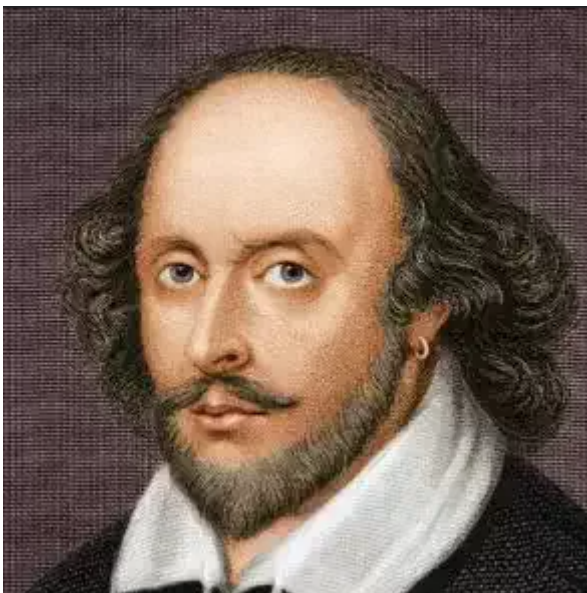


Output:

in

Inputs utilizados:

“describe the image”



Output:

a

Inputs utilizados:

“describe the image”



Output:
of
Inputs utilizados:
“what is in the picture?”



Output:
dog

Type-B: Custom Layer based Deep Fusion (CLDF)

Comparação entre os modelos **LLaMA-Adapter-V2** e **CogVLM**:

Análise do **LLaMA-Adapter-V2**:

- Melhora a compreensão visual integrando modelos especializados (como sistemas de legenda e OCR) durante a inferência, sem custos adicionais de treinamento.
- Nos benchmarks e habilidades de seguimento de instruções, o LLaMA-Adapter V2 mostra melhorias em relação ao seu antecessor (LLaMA-Adapter) e apresenta desempenho competitivo em comparação com outros modelos como MiniGPT-4 e LLaVA.
- Demonstra um ajuste equilibrado para instruções visuais e um forte aprendizado de instruções visuais, com bom desempenho em tarefas de Visual Question Answering (VQA) enquanto mantém a compreensão da linguagem.


Análise do **CogVLM**:

- Alcança resultados de estado da arte em 17 benchmarks de tarefas multimodais, incluindo image captioning, VQA e tarefas de grounding visual.
- Nos benchmarks de Vision Large Language Model (LVLM), o CogVLM demonstra liderança significativa, especialmente em tarefas complexas de raciocínio visual, localização de objetos e tarefas baseadas em texto.

Comparação do desempenho:

- CogVLM foca em uma fusão profunda dos dados visuais e linguísticos dentro de um modelo de linguagem pré-treinado, alcançando desempenho de estado da arte em uma ampla variedade de benchmarks multimodais. O modelo se destaca especialmente em tarefas como geração de legendas para imagens, VQA e benchmarks de LVLM.
- LLaMA-Adapter V2 é projetado com eficiência em mente, visando transformar grandes modelos de linguagem em seguidores de instruções visuais através de estratégias eficientes de ajuste de parâmetros. Embora não alcance o mesmo domínio abrangente de benchmarks como o CogVLM, o LLaMA-Adapter V2 equilibra efetivamente o ajuste para instruções visuais e a capacidade do modelo de linguagem, mostrando forte compreensão multimodal em tarefas que integram visual e textual.

Estudo realizado com base nos seguintes materiais:

 LLaMA-Adapter-V2.pdf

 CogVLM.pdf

Type-C: Non-Tokenized Early Fusion (NTEF)

Comparação entre os modelos **LLaVA** e **PaLM-E**:

Análise do **LLaVA**:

- É projetado para conversas multimodais e seguimento de instruções, utilizando modelos de visão e linguagem para realizar tarefas como descrições detalhadas de imagens, raciocínio complexo e visual question answering.
- LLaVA obteve alto desempenho em diálogos multimodais, especialmente em configurações de chatbot, e atingiu resultados fortes em benchmarks como o Science QA
- O enfoque no ajuste por instrução permite ao LLaVA se destacar em interações humano-AI, chat visual e no manuseio de conversas multimodais com descrições precisas e detalhadas de imagens.

Análise do **PaLM-E**:

- Se destaca em tarefas de embodied reasoning, como manipulação robótica, planejamento de ações e tomada de decisões, além de tarefas tradicionais de visão e linguagem, como VQA e legenda de imagens.
- É aplicado em guiar robôs em tempo real, resolver desafios de manipulação robótica e tarefas de percepção e raciocínio visual.

Comparação do desempenho:

- O PaLM-E se destaca em tarefas de raciocínio no mundo real, como manipulação robótica e planejamento, que dependem de dados de sensores físicos e tomada de decisões sequenciais. Ele é excelente em controle robótico e entendimento visual complexo, especialmente quando ancorado no contexto físico.
- O LLaVA, por sua vez, é mais otimizado para tarefas de seguimento de instruções e conversas multimodais. Ele se sai muito bem em raciocínio visual e em diálogos baseados em imagens, assim como em responder perguntas complexas relacionadas a imagens.

Estudo realizado com base nos seguintes materiais:

 LLaVA.pdf

 PaLM-E.pdf

Testes LLaVA

Inputs utilizados:

“describe the image”



Output:

The image is a panoramic view of a large city, showcasing its impressive skyline and the surrounding water. The cityscape features a mix of tall buildings, some of which are skyscrapers, and a bridge that connects different parts of the city. The city is situated next to a large body of water, which adds to the overall beauty of the scene.

There are several boats visible in the water, scattered throughout the scene, adding to the bustling atmosphere of the city. The combination of the city's architecture, the water, and the boats creates a dynamic and vibrant image of urban life.

Inputs utilizados:

“describe the image”

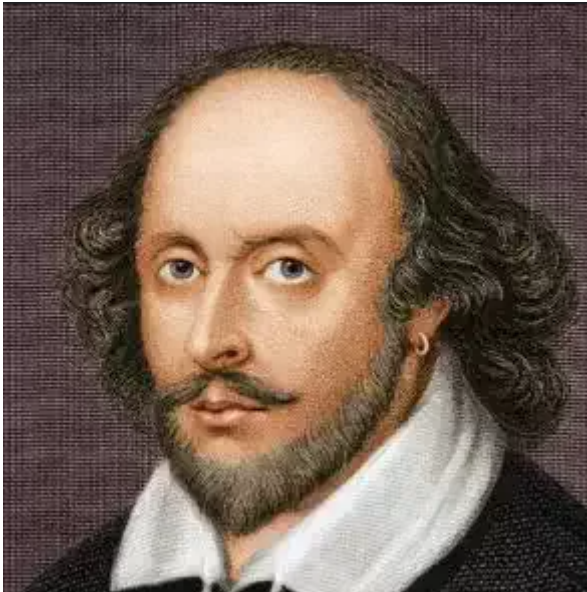


Output:

The image is a colorful and vibrant graphic design that features a sun with sunglasses on it. The sun is positioned in the center of the image, surrounded by a variety of colors and patterns. The design is visually appealing and likely intended to convey a sense of fun and creativity.

Inputs utilizados:

“describe the image”



Output:

The image features a portrait of a man with a beard and mustache, wearing a white collar. He appears to be an older gentleman, possibly a famous person or a historical figure. The man's hair is curly, and he is wearing a white shirt. The portrait is set against a purple background, which adds a touch of elegance to the scene.

Inputs utilizados:

“describe the image”



Output:

The image features a group of red flamingos made from paper, standing together on a white surface. There are a total of nine flamingos in the scene, with some of them being larger and others smaller. The flamingos are positioned in various orientations, creating a sense of depth and movement in the image. The paper flamingos are arranged in a way that they appear to be interacting with each other, giving the impression of a lively gathering.

Type-D: Tokenized Early Fusion (TEF)

Comparação entre os modelos **CM3Leon** e **LaViT**:

Análise do **CM3Leon**:

- É um modelo autoregressivo multi-modal baseado em decoder, focado tanto na geração de imagens a partir de texto quanto na geração de texto a partir de imagens.
- Utiliza um tokenizador de imagem que converte imagens em tokens discretos, que são então processados pelo modelo da mesma maneira que os tokens de texto.
- Otimizado para eficiência computacional, usando 5 vezes menos recursos de computação em comparação com outros modelos de ponta, como o PARTI, e ainda assim consegue resultados competitivos em termos de qualidade de geração de imagem (com FID de 4.88 no benchmark MS-COCO).


Análise do **LaVIT**:

- O LaVIT também é um modelo que tenta integrar imagens e texto de maneira unificada, tratando imagens como uma "linguagem estrangeira" ao traduzi-las em tokens discretos, que são compreensíveis por grandes modelos de linguagem. Ele usa uma abordagem de tokenização dinâmica para comprimir informações visuais de maneira eficiente, adaptando o número de tokens com base na complexidade da imagem.
- Também demonstra excelentes capacidades de geração de imagens, com um FID de 7.4 no benchmark MS-COCO, superando vários outros modelos LLMs multimodais. Ele pode gerar imagens não apenas a partir de prompts de texto, mas também com combinações de imagem e texto como entrada, oferecendo uma flexibilidade maior na geração.

Comparação do desempenho:

- O CM3Leon se destaca por sua eficiência no uso de recursos computacionais, alcançando desempenho de ponta com menos dados e menos treinamento em comparação com outros modelos.
- O LaVIT inova ao introduzir uma tokenização visual dinâmica, que adapta a quantidade de tokens de imagem com base na complexidade do conteúdo visual, otimizando o uso de recursos e melhorando a eficiência.
- Ambos os modelos são versáteis, mas o CM3Leon parece ser mais voltado para tarefas que exigem controle detalhado e edição de imagens, enquanto o LaVIT oferece uma abordagem mais integrada e eficiente para integração de dados multimodais complexos.

Estudo realizado com base nos seguintes materiais:

 [CM3Leon.pdf](#)

 [LaVIT.pdf](#)

APÊNDICE 3

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.


Data da Reunião (“gate”) de aprovação: 30 de out. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

Elisa Ayumi Masasi de Oliveira

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

De acordo com o planejamento das semanas anteriores, a partir do estudo da arquitetura do modelo escolhido, foi definido o objeto de estudo do projeto, que abrange os seguintes pontos:

- Objetivo: Implementar o modelo MLLM em português.
- Modelo de Referência:  LLaVA-OneVision.pdf ;
- Conjunto de Dados: Descrito em: [docs/Data.md](#)
 - [llava_instruct_150k](#)
 - [llava_instruct_80k](#)
 - [conversation_58k](#)
 - [detail_23k](#)
 - [complex_reasoning_77k](#)
- Validação do modelo: [SemEval-2025 Task 1](#)
 - Subtask A - Static Images
 - Subtask B - Image Sequences (or Next Image Prediction)

Adicionalmente, iniciei a tradução do primeiro dataset, com a biblioteca [py-googletrans](#).

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima semana será necessário realizar a tradução dos demais datasets que foram utilizados no treino do modelo.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

LEONARDO ALVES: Em análise! ▾

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 7 de nov. de 2024

Participantes da Entrega [matriculados em Residência em IA]:


ELISA AYUMI MASASI DE OLIVEIRA

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

De acordo com o planejamento da Semana anterior, foi identificado que a tradução dos dados documentados seria essencial para possibilitar um further pre-training no modelo LLaVA One Vision, com o objetivo de alcançar um cenário multilíngue para a língua portuguesa. Dessa forma, nesta semana, foram traduzidos os seguintes dados, de maneira estratégica, focando em dados conversacionais:

- [llava_instruct_150k](#) (sendo traduzido neste instante)
- [llava_instruct_80k](#) (sendo traduzido neste instante)
- [conversation_58k](#)
- [detail_23k](#)
- [complex_reasoning_77k](#)

Além disso, como forma de complementar a proposta, estou estudando data augmentation para dados idiomáticos:

 [A Survey on Data Synthesis and Augmentation for Large.pdf](#)

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima entrega, serão realizados os experimentos de further pre-training com o objetivo também de validar se os dados serão suficientes.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: 

APÊNDICE 4

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 13 de nov. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

ELISA AYUMI MASASI DE OLIVEIRA

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Dando continuidade ao trabalho feito na Semana anterior, a tradução dos datasets foi encerrada, e foi realizada uma etapa de avaliação qualitativa em parte das amostras.

[301024 Estudos - LLaVA](#)

Iniciando o treino do modelo, onde acabou ocorrendo a necessidade de trocar o modelo utilizado, do LLaVA-OneVision para o LLaVA.

Para a etapa de avaliação estou estudando os artigos publicados pelas equipes que participaram de edições anteriores.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Para a próxima Semana os objetivos serão:

- Realizar a etapa de avaliação do modelo;
- Levantar mais datasets, visando um treino mais extensivo.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

CEDRIC LUIZ DE CARVALHO: [Go!](#)

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 18 de set. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

ELISA AYUMI MASASI DE OLIVEIRA

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Durante essa Semana:

- Após os testes iniciais de fine-tuning utilizando a GPU 4090, identificou-se a necessidade de um hardware com maior capacidade computacional.
- Diante disso, foi realizada a migração para a GPU A100.
- Na última Semana, os experimentos foram avaliados em um dataset de menor complexidade, buscando a replicação desse fluxo nos datasets que foram traduzidos em marcos anteriores.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Replicar o fluxo para o conjunto de dados traduzido.

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

ACEITE DA ENTREGA:

LEONARDO ANTÔNIO ALVES: Em análise! ▾

Termo de Aceite de Entrega

Objetivo deste documento

Este documento faz parte do Processo da disciplina Residência em IA e tem como objetivo formalizar o aceite da entrega considerando o planejado e o realizado para o período.

Data da Reunião (“gate”) de aprovação: 5 de dez. de 2024

Participantes da Entrega [matriculados em Residência em IA]:

ELISA AYUMI MASASI DE OLIVEIRA

Entrega: [descrever a ENTREGA: requisitos e produtos gerados: links para textos, códigos, vídeos etc.]

Na Semana 10, foi realizado o treinamento bem sucedido do modelo [LLaVA 13B](#). Assim como constatado na Semana anterior, foi necessário migrar para a NVIDIA A100 80GB, onde o treinamento foi bem sucedido, como descreve o documento: [301024 Estudos - LLaVA](#)

Algumas amostras podem ser vistas no documento supracitado, mas o modelo também pode ser testado na plataforma [Replicate](#). Com base nisso, estou realizando uma avaliação qualitativa dos resultados gerados.

Planejamento: [descrever o que pretende fazer para realizar a próxima ENTREGA]

Finalizar a escrita do documento final de conclusão de curso (TCC).

Observação: [caso precise fazer alguma observação, de qualquer “natureza”]

Por questões familiares, não será possível realizar a entrega presencial desse documento.

ACEITE DA ENTREGA:


CEDRIC LUIZ DE CARVALHO: Go! ▾

301024 Estudos - LLaVA.docx citado no Termo de Aceite de Entrega de 5 de dezembro.

Estudos - LLaVA

Estudo sobre os experimentos com o modelo LLaVA:

Baseado nos artigos:

Modelo original  [LLaVA.pdf](#)

Visual Instruction Tuning

Haotian Liu^{1*}, Chunyuan Li^{2*}, Qingyang Wu³, Yong Jae Lee¹

¹University of Wisconsin–Madison ²Microsoft Research ³Columbia University

<https://llava-vl.github.io>

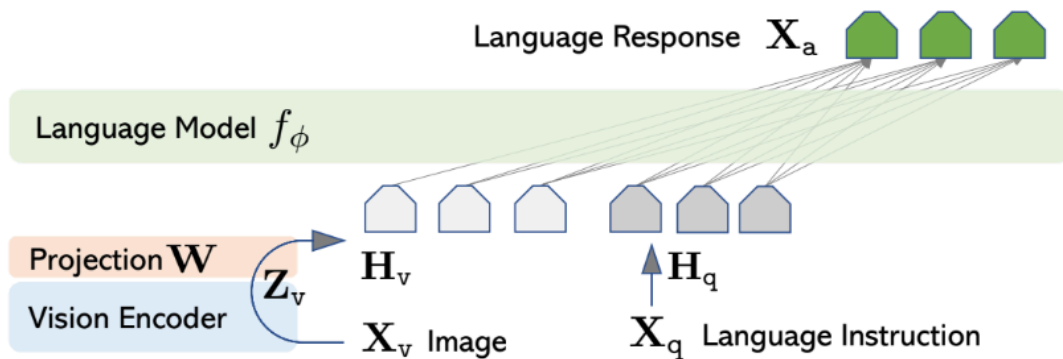


Figure 1: LLaVA network architecture.

[Datasets Originais:](#)

Data file name	Size
llava_instruct_150k.json	229 MB
llava_instruct_80k.json	229 MB
conversation_58k.json	126 MB
detail_23k.json	20.5 MB
complex_reasoning_77k.json	79.6 MB

Amostras do dataset original:

```
{
  "id": "000000379143",
  "image": "000000379143.jpg",
  "conversations": [
    {
      "from": "human",
      "value": "What is the man doing in the back of the pickup truck?\n<image>"
    },
    {
      "from": "gpt",
      "value": "The man is sitting on top of piled objects or belongings loaded into the back of
a pickup truck."
    },
    {
      "from": "human",
      "value": "Is the man holding anything in his hands?"
    },
    {
      "from": "gpt",
      "value": "Yes, the man is holding a beer in his hand while sitting on top of the objects in
the back of the pickup truck."
    }
  ]
}
```

Datasets traduzidos:

Amostras do dataset traduzido:

```
{
  "id": 148540,
  "image": "000000148540.jpg",
  "conversations": [
    {
      "from": "human",
      "value": "<imagem>\nAs pessoas estão saindo de um trem ou ônibus na imagem?"
    },
    {
      "from": "gpt",
      "value": "As pessoas estão saindo de um trem na imagem."
    },
    {
      "from": "human",
      "value": "Você pode descrever as cores do trem?"
    },
    {
      "from": "gpt",
      "value": "O trem é azul e rosa."
    },
    {
      "from": "human",
      "value": "O que está acontecendo na plataforma quando o trem para?"
    },
    {
      "from": "gpt",
      "value": "Quando o trem para na plataforma, a porta se abre e as pessoas começam a sair do trem. Entre eles, há um casal carregando suas bagagens."
    }
  ]
}
```

O fine-tuning do modelo [LLaVA 13B](#) foi realizado seguindo as recomendações oficiais presentes no repositório do GitHub do projeto. Abaixo estão os detalhes específicos da configuração utilizada:

1. Learning Rate: O modelo foi ajustado com uma taxa de aprendizado de **$2e-4$** . Essa escolha é consistente com as práticas recomendadas para treinamento estável e eficiente de grandes modelos de linguagem.
2. Comprimento Máximo do Modelo: O parâmetro `model_max_length` foi configurado para **2048**, permitindo que o modelo processe sequências extensas de texto, o que é especialmente útil em tarefas que requerem contextos mais longos.
3. Número de Épocas: O treinamento foi realizado por apenas 1 época, conforme sugerido nas diretrizes, visando minimizar o risco de overfitting e garantir que o modelo continue a generalizar bem em dados não vistos.
4. Hardware Utilizado: O fine-tuning foi executado em uma GPU **NVIDIA A100 com 80GB de VRAM**, um hardware adequado para lidar com as demandas de memória e processamento associadas a modelos de grande escala como o LLaVA 13B.

Essa configuração foi projetada para equilibrar eficiência e desempenho, garantindo que o modelo fosse ajustado de maneira otimizada para a tarefa específica sem comprometer sua capacidade geral. A implementação seguiu as instruções fornecidas no repositório oficial do LLaVA, garantindo reprodutibilidade e aderência às melhores práticas da comunidade.

Algumas amostras produzidas pelo modelo podem ser vistas a seguir:

Prompt:

“Descreva a imagem:”



Output do modelo:

“A borboleta azul e vermelha está parada em um pedaço de areia, próximo a uma colina. A borboleta é o principal foco da imagem, com sua vibrante cor e design. O céu é visível no fundo, criando um cenário pitoresco para a borboleta.”

Prompt:

“Qual é a cor da quadra de tênis da imagem?”



Output do modelo:

“A quadra de tênis da imagem é azul.”

Prompt:

“Que horas eram no momento em que a foto foi tirada?”



Output do modelo:

“A foto foi tirada durante o pôr do sol, pois a luz do sol está se pondo no horizonte.”

Prompt:

“O que esta acontecendo na imagem?”



Output do modelo:

“Na imagem, há uma erupção vulcânica em curso, com lava derramando-se de uma montanha e formando um rio de lava. A erupção está ocorrendo em uma montanha rochosa, e a lava está fluindo em direção a uma área de terra mais plana.”

O modelo treinado também pode ser testado na plataforma [Replicate](#).