Catalogação na publicação
Seção de Catalogação e Classificação

S725a Sousa, Matheus Fernandes de.
Aplicação do modelo de linguagem BLIP-2 na geração automática de descrições em vídeos esportivos / Matheus Fernandes de Sousa. - João Pessoa, 2024.
14 f. : il.

Orientação: Yuri de Almeida Malheiros Barbosa.
Coorientação: Thaís Gaudencio do Rêgo.
TCC (Graduação) - UFPB/CI.

1. BLIP-2. 2. Métricas. 3. Descrições. 4. Vídeo. I. Barbosa, Yuri de Almeida Malheiros. II. Rêgo, Thaís Gaudencio do. III. Título.

UFPB/CI CDU 004.932.4

Elaborado por Michelle de Kássia Fonseca Barbosa - CRB-738

Aplicação do Modelo de Linguagem BLIP-2 na Geração Automática de Descrições em Vídeos Esportivos

Matheus F. Sousa[1], **Yuri de Almeida M. Barbosa**[2], **Thaís G. do Rêgo**[3]

[1]Centro de Informática – Universidade Federal da Paraíba (UFPB)
R. dos Escoteiros, s/n - Mangabeira, João Pessoa - PB, 58055-000

matheus.fernandes@academico.ufpb.br, yuri@ci.ufpb.br, gaudenciothais@gmail.com

Abstract. *The growing demand for audio description in sports videos and the need to make audiovisual content more accessible have driven the development of automatic technologies that can generate accurate and coherent descriptions of sports events. This study investigates the application of the BLIP-2 model in generating automatic descriptions of videos from various sports, such as goalball, volleyball, and soccer, aiming to capture essential visual details and provide consistent and precise descriptions. To assess the quality of these descriptions, evaluation metrics like METEOR, with scores up to 0,60, and ROUGE-L, reaching 0,63, were used, indicating semantic and structural alignment with manual references. While the model demonstrated effectiveness in captioning popular sports such as soccer, with consistent results and positive metrics, a limitation was observed in less widely covered sports, such as goalball, where the model showed difficulty in achieving accuracy. This is one of the study's key findings, highlighting the need for more training data for less popular sports.*

Resumo. *A demanda crescente por audiodescrição em vídeos esportivos e a necessidade de tornar o conteúdo audiovisual mais acessível impulsionam o desenvolvimento de tecnologias automáticas que possam gerar descrições precisas e coerentes de eventos esportivos. Este estudo investiga a aplicação do modelo BLIP-2 na geração de descrições automáticas de vídeos de diversas modalidades esportivas, como goalball, vôlei e futebol, com o objetivo de capturar detalhes visuais essenciais para fornecer descrições coerentes e precisas. Para avaliar a qualidade dessas descrições, foram utilizadas métricas como METEOR, com pontuações de até 0,60, e ROUGE-L, alcançando 0,63, que indicam correspondência semântica e estrutural com as referências manuais. Embora o modelo tenha demonstrado eficácia na legendagem de esportes como o futebol, com resultados consistentes e métricas positivas, observou-se uma limitação em esportes menos divulgados, como o goalball, onde o modelo apresentou dificuldades de precisão. Este é um dos principais resultados do estudo, destacando a necessidade de mais dados de treinamento para modalidades esportivas menos populares.*

1. Introdução

A acessibilidade em conteúdos audiovisuais constitui um aspecto fundamental para promover a inclusão social, especialmente no contexto esportivo, onde a emoção e a dinâmica das competições são intensas. Contudo, muitos eventos esportivos ainda carecem de descrições adequadas, especialmente em modalidades menos populares e em

transmissões ao vivo, onde a falta de infraestrutura e tecnologia para geração automática de descrições limita o acesso de pessoas com deficiências visuais ou auditivas. Isso inclui tanto competições locais quanto eventos paralímpicos, que frequentemente não possuem soluções robustas para audiodescrição em tempo real.. Aproximadamente 1 bilhão de pessoas no mundo vive com algum tipo de deficiência, e a falta de acessibilidade em mídias digitais representa um dos principais desafios enfrentados por essa população (World Health Organization, 2022).

Nos últimos anos, os avanços em inteligência artificial (IA), particularmente na área de IA generativa, têm aberto novas possibilidades para aprimorar a acessibilidade em conteúdos audiovisuais. Modelos de linguagem, como o BLIP-2, destacam-se por sua capacidade de gerar descrições textuais a partir de imagens, tornando conteúdos anteriormente inacessíveis mais inclusivos. O **BLIP-2** , conforme descrito por Li et al. (2023), combina técnicas de visão computacional e processamento de linguagem natural, permitindo a criação de descrições que são não apenas informativas, mas também contextualmente relevantes.

Ademais, a implementação de soluções de IA para a acessibilidade não apenas beneficia pessoas com deficiência visual, mas também enriquece a experiência de todos os espectadores. A inclusão de descrições automáticas pode proporcionar uma nova dimensão de engajamento, permitindo que um público mais amplo compreenda e aprecie as nuances das competições esportivas. À medida que a tecnologia avança, é crucial que as indústrias de mídia e entretenimento adotem essas inovações para garantir que todos tenham a oportunidade de participar e desfrutar do conteúdo esportivo de maneira equitativa.

O presente estudo tem como objetivo investigar a aplicação do modelo de linguagem BLIP-2, na geração automática de descrições para vídeos esportivos, avaliando a qualidade das descrições geradas, em comparação com aquelas elaboradas manualmente e com as produzidas por modelos de IA, como o ChatGPT. Para essa análise, utilizaremos métricas de avaliação como METEOR e ROUGE-L, amplamente reconhecidas na literatura para medir a qualidade de descrições geradas automaticamente [Banerjee e Lavie 2005; Lin 2004]. A relevância deste estudo reside na sua contribuição para a inclusão de pessoas com deficiência visual no universo esportivo, além de oferecer percepções sobre a eficácia de modelos de IA generativa na produção de conteúdo acessível. Essa tecnologia pode ser aplicada em diversos esportes, como *goalball*, futebol e vôlei, onde as ações exigem descrições precisas e detalhadas. O *goalball*, em particular, é fundamental nesse contexto, pois foi desenvolvido especificamente para atletas com deficiência visual, e a implementação de descrições automáticas pode melhorar significativamente a experiência de espectadores e participantes, promovendo uma maior inclusão e compreensão do jogo [Gavião de Almeida, 2012].

Especificamente, o trabalho pretende:

- Analisar a acurácia das descrições geradas pelo BLIP-2.
- Identificar as limitações do modelo na interpretação de cenas esportivas complexas.

2. Trabalhos relacionados

Bernardi et al. (2016) realizaram uma revisão sobre modelos de geração de descrições para imagens, abordando métodos de geração direta e de recuperação de descrições. Utilizando métricas como BLEU, METEOR e CIDEr, o estudo concluiu que os modelos de recuperação geram descrições mais precisas para imagens conhecidas, enquanto os modelos de geração direta são mais flexíveis, mas menos detalhados. O estudo é focado em imagens estáticas de conjuntos de dados como o MS-COCO.

Bianco et al. (2023) buscaram melhorar a descritividade de legendas por meio da fusão de descrições geradas por múltiplos modelos, incluindo BLIP-2 e OFA (*One for All*). Com métricas como METEOR, CIDEr e BLIPScore, os resultados indicaram que a fusão de descrições resulta em legendas mais ricas e precisas, aproximando-se da qualidade de descrições humanas.

Rao et al. (2024) desenvolveram o modelo *MatchVoice* para gerar comentários automáticos em partidas de futebol, utilizando um *pipeline* de alinhamento temporal multimodal para sincronizar comentários com eventos. Avaliado por métricas como METEOR e ROUGE-L, os resultados mostraram que o alinhamento temporal melhora significativamente a precisão e a relevância das legendas geradas.

Mkhallati et al. (2023) propuseram uma abordagem de *Single-anchored Dense Video Captioning* (SDVC) em vídeos de futebol, combinando detecção de ações com geração de legendas para criar comentários ancorados no tempo. A avaliação com métricas como BLEU e METEOR revelou que o alinhamento temporal melhora a correspondência entre eventos e descrições, fornecendo um nível de detalhamento importante para vídeos esportivos.

Em comum, esses trabalhos, assim como o presente estudo, utilizam métricas como METEOR para avaliar a precisão semântica e buscam melhorar a qualidade e detalhamento das descrições automáticas. Os estudos de Rao et al. (2024) e Mkhallati et al. (2023), assim como este trabalho, focam na geração de descrições temporais para vídeos esportivos, destacando a importância da captura precisa de ações em eventos dinâmicos. Entretanto, o presente trabalho também amplia o escopo ao analisar diferentes esportes, como o vôlei e, esportes menos conhecidos, como o *goalball*, oferecendo um desafio adicional devido à escassez de dados e ao caráter único desse tipo de conteúdo esportivo.

3. Processos metodológicos

O estudo foi conduzido utilizando uma combinação de técnicas de visão computacional e processamento de linguagem natural. A seguir, são detalhadas as ferramentas, infraestrutura, coleta e preparação de dados, processamento de vídeo e análise dos resultados.

A implementação foi realizada utilizando Python, com o suporte das bibliotecas Transformers (versão 4.44.2), para o processamento de linguagem natural, e cv2 (versão 4.10.0) para a manipulação de vídeos. O ambiente de desenvolvimento incluiu o uso do Google Colab, que forneceu os recursos computacionais necessários para a utilização do modelo.

Foram utilizadas GPUs NVIDIA Tesla T4 e/ou NVIDIA A100 no Google Colab Pro. Essas GPUs são essenciais para acelerar o treinamento e a inferência do modelo BLIP-2, oferecendo suporte para processamento paralelo e operações de alta performance.

3.1. Preparo e geração das descrições

Para avaliar a capacidade do modelo BLIP-2 na geração automática de descrições de eventos esportivos, foi adotada a seguinte metodologia:

3.1.1. Coleta dos dados

Para a coleta, foram utilizados vídeos de eventos esportivos disponíveis em canais oficiais de esportes no Youtube, como: Cazé TV, FIFA, Paralympics, Olympics e CONMEBOL Libertadores. Foram testados o total de 5 vídeos para cada modalidade esportiva, onde foram extraídos 3 quadros por segundo em um intervalo de 10 segundos por vídeo, totalizando 30 quadros por vídeo. Os vídeos utilizados possuem uma resolução de 640x360 pixels, oferecendo qualidade visual com menor uso de largura de banda, comparado a resoluções mais altas.

3.1.2. Preparação dos dados

Foi realizada a onfiguração do modelo BLIP-2 *(Bootstrapped Language Image Pre-training)*, desenvolvido pela equipe da Salesforce, ajustado para utilizar a quantização de 8 bits através da configuração *BitsAndBytesConfig*. Essa abordagem reduz significativamente o uso de memória, permitindo o processamento mais eficiente dos quadros, especialmente em dispositivos com recursos limitados. Para a instanciação do processador e do modelo, foram importados os componentes *Blip2Processor* e o *Blip2ForConditionalGeneration*, carregados a partir da biblioteca *Transformers*. O modelo utilizado é o *Salesforce/blip2-opt-2.7b*, que faz parte da série BLIP-2, conhecida por combinar pré-treinamento baseado em multimodalidade (texto e imagem), com capacidades avançadas de geração de linguagem. Esse modelo foi ajustado para suportar o processamento de imagens e a geração de descrições com eficiência, mesmo em cenários de baixa memória, graças à quantização de 8 bits e ao uso otimizado de CPU.

3.1.3. Processamento de vídeo

Os vídeos foram processados para capturar 3 quadros por segundo, permitindo a representação de momentos chave. Esse intervalo foi definido para equilibrar a quantidade de dados, sem sobrecarregar o processamento. Para cada quadro, foi gerada uma descrição automática utilizando o modelo BLIP-2, baseado exclusivamente no conteúdo visual, sem *prompts* manuais, destacando a capacidade do modelo em compreender a cena e gerar legendas contextuais.

3.1.4. Análise dos Resultados

Cada quadro foi exibido com sua descrição correspondente, permitindo uma análise visual clara da relação entre imagem e legenda gerada. As descrições foram também avaliadas quantitativamente, através das métricas METEOR e ROUGE-L, fornecendo uma visão

mais objetiva da precisão das legendas, em relação às referências esperadas. Exemplos de quadros gerados são exibidos no **Apêndice** deste trabalho.

3.2. Métricas de Avaliação

As métricas de avaliação são essenciais para medir a qualidade das descrições geradas por modelos de linguagem. Neste estudo, utilizamos duas métricas principais: METEOR e ROUGE-L.

A METEOR *(Metric for Evaluation of Translation with Explicit Ordering)* avalia a precisão das descrições geradas ao compará-las com descrições de referência, considerando correspondência de palavras, sinônimos e a ordem das palavras. Essa métrica é útil para garantir que as descrições sejam semanticamente adequadas e fluentes. A pontuação do METEOR varia de 0 a 1, onde 0 indica que não há correspondência entre a tradução gerada e a referência e 1 indica uma correspondência perfeita [Denkowski and Lavie 2014].

A ROUGE-L *(Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence)* mede a similaridade entre as descrições geradas e as de referência, focando na subsequência mais longa. Isso permite avaliar a estrutura gramatical e a ordem das palavras, sendo especialmente valiosa em tarefas de resumo e geração de texto. O valor do ROUGE-L varia de 0 a 1. Um valor de 0 indica que não há correspondência entre a descrição e a referência, enquanto um valor de 1 indica uma correspondência perfeita [Lin and Och 2004].

3.3. Comparação das descrições geradas entre esportes

Para cada esporte avaliado (vôlei, futebol e *goalball*), foi realizada uma comparação entre as descrições geradas automaticamente pelo modelo BLIP-2 e um conjunto de referências criadas para cada tipo de evento. Essas descrições foram extraídas de partidas reais, como mostrado nas seções anteriores, e as métricas METEOR e ROUGE-L foram utilizadas para avaliar a precisão e fluência das descrições.

Além disso, foram geradas referências alternativas, utilizando o ChatGPT. Essas novas referências foram usadas para gerar um segundo conjunto de métricas, o que permitiu uma análise mais abrangente sobre a influência da criação de descrições utilizadas como referências, nos resultados das descrições geradas pelo BLIP-2.

Vôlei: As referências para o vôlei focaram tanto nas ações críticas do jogo, como bloqueios e ataques, como em ações generalistas. Exemplos de referências utilizadas incluem:

> *"The volleyball player attacks the ball with an impressive jump."* (O jogador de vôlei ataca a bola com um salto impressionante).
> *"A volleyball game is happening in an indoor arena."* (Um jogo de vôlei está acontecendo em uma arena coberta).

Futebol: Nas partidas de futebol, as referências se concentraram em momentos decisivos, como tentativas de gol e defesas. Exemplos de referências incluem:

"The soccer player is dribbling past an opponent." (O jogador de futebol está driblando um adversário).
"Two players are competing for possession of the ball." (Dois jogadores competem pela posse de bola).

***Goalball*:** Para *goalball*, as referências destacaram as jogadas estratégicas e a interação dos jogadores com o ambiente. Exemplos de referências incluem:

"The ball is being passed among the players." (A bola está sendo passada entre os jogadores).
"A player is defending the goal in a goalball game." (Um jogador está defendendo o gol em um jogo de *goalball*).

As métricas METEOR e ROUGE-L foram calculadas, tanto em relação às referências originais criadas manualmente, quanto às referências geradas pelo ChatGPT. Isso permitiu uma análise comparativa mais detalhada, evidenciando o impacto de diferentes fontes de referência, sobre a avaliação das descrições geradas pelo modelo BLIP-2.

4. Análise e avaliação dos resultados

Esta seção apresenta os resultados obtidos a partir do percurso metodológico descrito anteriormente. As análises buscam identificar padrões de desempenho, falhas e potencialidades do modelo, além de examinar a influência das diferentes fontes de referência nas avaliações métricas. Também são discutidas possíveis limitações e aprimoramentos a serem considerados para futuras aplicações.

4.1. Média e desvio padrão para os 5 vídeos testados de cada modalidade

Para avaliar a precisão das descrições geradas pelo modelo BLIP-2 em diferentes esportes, foram calculadas as métricas de METEOR e ROUGE-L para os 5 vídeos de cada modalidade: futebol, vôlei e *goalball*. As tabelas abaixo apresentam a média e o desvio padrão dessas métricas, tanto para as referências manuais quanto para as geradas pelo ChatGPT (indicadas como GPT), permitindo uma análise comparativa da qualidade das descrições.

Tabela 1. Métricas gerais dos vídeos de futebol

Métrica	Média	Desvio Padrão
METEOR	0,59	0,10
ROUGE-L	0,31	0,15
METEOR (GPT)	0,24	0,06
ROUGE-L (GPT)	0,26	0,08

Tabela 2. Métricas gerais dos vídeos de vôlei

Métrica	Média	Desvio Padrão
METEOR	0,47	0,13
ROUGE-L	0,12	0,09
METEOR (GPT)	0,13	0,05
ROUGE-L (GPT)	0,18	0,06

Tabela 3. Métricas gerais dos vídeos de *goalball*

Métrica	Média	Desvio Padrão
METEOR	0,24	0,09
ROUGE-L	0,08	0,02
METEOR (GPT)	0,15	0,06
ROUGE-L (GPT)	0,05	0,02

No futebol, as descrições manuais alcançaram uma média de METEOR de 0,59 e ROUGE-L de 0,31, com baixos desvios padrão (0,10 e 0,15, respectivamente), sugerindo alta precisão semântica e estrutural, além de consistência entre as descrições geradas e as referências manuais. Em contraste, as descrições de referência geradas pelo ChatGPT apresentaram uma queda significativa, com METEOR de 0,24 e ROUGE-L de 0,26, indicando uma menor precisão e fluência.

No vôlei, embora as médias de METEOR (0,47) e ROUGE-L (0,12) tenham sido boas, o desvio padrão mais elevado (0,13 em METEOR e 0,09 em ROUGE-L) reflete uma maior variabilidade na precisão das descrições automáticas, sugerindo que o modelo pode não capturar de forma consistente a dinâmica das partidas. As descrições do ChatGPT também ficaram abaixo do esperado, com METEOR de 0,13 e ROUGE-L de 0,18.

Já no *goalball*, o desempenho foi significativamente mais fraco. As descrições automáticas tiveram baixa correspondência semântica e estrutural, com média de METEOR de 0,24 e ROUGE-L de 0,08, refletindo que o modelo enfrentou dificuldades para reconhecer as ações específicas desse esporte menos comum. O desvio padrão baixo (0,09 em METEOR e 0,02 em ROUGE-L) indica consistência nas falhas, apontando para uma lacuna na generalização do modelo. As descrições de referência geradas pelo ChatGPT também apresentaram resultados inferiores, sugerindo a necessidade de mais treinamento em dados específicos para esportes menos populares, como o *goalball*.

4.2. Análise individual de vídeos experimentados

Nesta subseção, são apresentados e discutidos os resultados obtidos a partir das descrições automáticas geradas pelo modelo BLIP-2 para alguns vídeos selecionados de esportes como vôlei, futebol e *goalball*.

BLIP-2

Quadro	Descrição	METEOR	ROUGE-L	METEOR (gpt)	ROUGE-L (gpt)
	*a volleyball player is jumping up to hit the ball*	0,51	0,63	0,39	0,11
	*a group of men playing volleyball in a stadium*	0,48	0,11	0,11	0,24
	*a volleyball game is being played in front of a large crowd*	0,44	0,10	0,11	0,20

Figura 1. Traduções: “um jogador de vôlei está saltando para bater na bola”; “um grupo de homens jogando vôlei em um estádio”; “um jogo de vôlei está sendo jogado diante de uma grande multidão”. (As métricas-GPT referem-se às pontuações calculadas com base nas referências geradas pelo ChatGPT). Fonte: Autor (2024).

BLIP-2

Quadro	Descrição	METEOR	ROUGE-L	METEOR (gpt)	ROUGE-L (gpt)
	*the women's volleyball team is hugging each other*	0,35	0,22	0,24	0,24
	*a group of women playing volleyball in a court*	0,55	0,11	0,11	0,24
	*a volleyball game is being played in front of a crowd*	0,44	0,10	0,11	0,21

Figura 2. Traduções: “a equipe feminina de vôlei está se abraçando”; “um grupo de mulheres jogando vôlei em uma quadra”; “um jogo de vôlei está sendo jogado diante de uma multidão”. (As métricas-GPT referem-se às pontuações calculadas com base nas referências geradas pelo ChatGPT). Fonte: Autor (2024).

As descrições automáticas analisadas na Figura 1 apresentaram um bom desempenho em precisão semântica, com uma pontuação de METEOR de 0,51, indicando boa correspondência com as descrições manuais. No entanto, a pontuação ROUGE-L de 0,63 sugere algumas variações na estrutura das frases em comparação com as referências. Por outro lado, as descrições geradas pelo ChatGPT mostraram um desempenho inferior, com menor precisão e fluência.

Na Figura 2, os resultados foram semelhantes, com a maior pontuação de METEOR atingindo 0,55, sugerindo uma boa correspondência com os eventos visuais. Contudo, as pontuações de ROUGE-L indicam diferenças na estrutura linguística entre as descrições automáticas e as manuais, apesar da correção semântica.

BLIP-2

Quadro	Descrição	METEOR	ROUGE-L	METEOR (gpt)	ROUGE-L (gpt)
	*two soccer players are fighting for the ball*	0,60	0,47	0,13	0,12
	*a soccer player is kicking the ball while another player is running after it*	0,55	0,52	0,19	0,09
	*a soccer player is diving to catch the ball*	0,45	0,56	0,50	0,11

Figura 3. Traduções: "dois jogadores de futebol estão disputando a bola"; "um jogador de futebol está chutando a bola enquanto outro jogador corre atrás dela"; "um jogador de futebol está se lançando para pegar a bola". (As métricas-GPT referem-se às pontuações calculadas com base nas referências geradas pelo ChatGPT). Fonte: Autor (2024).

BLIP-2

Quadro	Descrição	METEOR	ROUGE-L	METEOR (gpt)	ROUGE-L (gpt)
	*a soccer player is kicking the ball in front of other players*	0,57	0,57	0,21	0,10
	*a soccer player is sliding into the goal*	0,49	0,59	0,26	0,12
	*a soccer player is laying in the ground after a goal*	0,42	0,50	0,18	0,20

Figura 4. Traduções: "um jogador de futebol está chutando a bola na frente de outros jogadores"; "um jogador de futebol está deslizando em direção ao gol"; "um jogador de futebol está deitado no chão após um gol". (As métricas-GPT referem-se às pontuações calculadas com base nas referências geradas pelo ChatGPT). Fonte: Autor (2024).

Na Figura 3, a descrição *"two soccer players are fighting for the ball"* apresentou a melhor pontuação de METEOR (0,60) e ROUGE-L (0,47), o que sugere uma forte correspondência entre a descrição gerada e a referência manual, tanto no conteúdo semântico quanto na estrutura da frase. As descrições subsequentes, como *"a soccer player is kicking the ball while another player is running after it"*, também tiveram um bom desempenho, especialmente em METEOR (0,55), indicando que o modelo capturou bem a essência da ação no vídeo. Na Figura 4, as descrições continuaram a apresentar bons resultados, com *"a soccer player is kicking the ball in front of other players"* alcançando 0,57 em METEOR e 0,57 em ROUGE-L, o que destaca uma similaridade considerável entre a descrição gerada e a referência.

BLIP-2

Quadro	Descrição	METEOR	ROUGE-L	METEOR (gpt)	ROUGE-L (gpt)
	*a man is standing in front of a net while another man is playing a volleyball game*	0,22	0,12	0,14	0,08
	*a badminton match is being played in an indoor arena*	0,21	0,07	0,10	0,05
	*brazilian men's volleyball team in action during the olympic games*	0,22	0,07	0,11	0,04

Figura 5. Traduções:"um homem está em pé em frente a uma rede enquanto outro homem joga uma partida de vôlei"; "uma partida de badminton está sendo jogada em uma arena coberta"; "a equipe masculina de vôlei do Brasil em ação durante os Jogos Olímpicos". (As métricas-GPT referem-se às pontuações calculadas com base nas referências geradas pelo ChatGPT). Fonte: Autor (2024).

BLIP-2

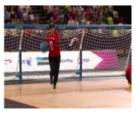
Quadro	Descrição	METEOR	ROUGE-L	METEOR (gpt)	ROUGE-L (gpt)
	*two men playing a game of handball in front of a crowd*	0,18	0,07	0,11	0,05
	*a group of people playing soccer on a court*	0,15	0,08	0,11	0,05
	*a man kneeling on the floor while playing volleyball*	0,26	0,06	0,13	0,04

Figura 6. Traduções: "dois homens jogando uma partida de handebol diante de uma multidão";"um grupo de pessoas jogando futebol em uma quadra"; "um homem ajoelhado no chão enquanto joga vôlei". (As métricas-GPT referem-se às pontuações calculadas com base nas referências geradas pelo ChatGPT). Fonte: Autor (2024).

A descrição *"a man is standing in front of a net while another man is playing a volleyball game"*, na Figura 5, teve uma pontuação METEOR de 0,22 e ROUGE-L de 0,12, indicando uma correspondência limitada com o conteúdo correto. Isso sugere que o modelo confundiu o esporte, descrevendo uma ação de vôlei em vez de *goalball*. De maneira semelhante, a segunda descrição *"a badminton match is being played in an indoor arena"* obteve pontuações ainda mais baixas, com METEOR de 0,21 e ROUGE-L de 0,07, reforçando a dificuldade do modelo em reconhecer corretamente as ações esportivas específicas e o esporte praticado. Na Figura 6, observou-se um desempenho semelhante. A descrição *"two men playing a game of handball in front of a crowd"* teve um METEOR de 0,18 e ROUGE-L de 0,07, refletindo a confusão com outro esporte (handebol).

4.3. Considerações finais

Os resultados obtidos indicam que o modelo BLIP-2 apresentou desempenho satisfatório em esportes populares, como vôlei e futebol, com boas pontuações em métricas como METEOR e ROUGE-L, sugerindo que o modelo consegue capturar as principais ações de forma semântica. No entanto, enfrentou limitações significativas em esportes menos populares, como o *goalball*, onde as descrições geradas foram menos precisas e frequentemente confusas, indicando uma menor familiaridade do modelo com as particularidades desse esporte. Isso reforça a necessidade de melhorar o treinamento com dados mais diversos e específicos para garantir maior acurácia em contextos variados.

Além disso, as pontuações mais baixas em métricas como ROUGE-L sugerem que, apesar de captar corretamente as ações principais, o modelo BLIP-2 ainda apresenta desafios na estruturação fluente e coesa das descrições geradas. O modelo também demonstrou dificuldades em distinguir entre diferentes modalidades esportivas, o que reforça a importância de aprimorar o reconhecimento contextual e visual.

Para aprimorar o desempenho do BLIP-2, é importante expandir o conjunto de dados de treinamento com mais vídeos de diferentes esportes e em maior variedade de cenários. Além disso, testar outros modelos mais avançados podem trazer informações valiosas. A integração de outras métricas de avaliação, como BLEU e CIDEr, pode complementar a análise do desempenho, oferecendo uma avaliação mais completa da precisão e fluência das descrições. Essas abordagens futuras têm o potencial de aumentar a aplicabilidade e a eficácia do modelo BLIP-2 em tarefas complexas de descrição de vídeos esportivos, com foco em melhorar a acessibilidade e a categorização automatizada do conteúdo.

5. Referências

Lin, C.-Y. and Och, E. H. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the 42nd Annual Meeting of ACL (ACL 2004), Barcelona, Spain.

Denkowski, M. and Lavie, A. (2013). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Callison-Burch, C., Koehn, P., Monz, C., and Zaidan, O., editors, Proceedings of the Eighth Workshop on Statistical Machine Translation, pages 45–51. Association for Computational Linguistics.

LI, X., WANG, Z., LIU, H., WANG, Y., & WANG, C. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

LI, J., LI, D., SAVARESE, S., & HOI, S. (2022). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models.

TANIGUCHI, Yasufumi et al. Generating live soccer-match commentary from play data. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2019. p. 7096-7103.

Souza, M. R. A. D. (2020). Esportes na era do streaming: uma análise da transmissão e consumo de eventos esportivos na Internet.

Siqueira, G. G., Saraiva, C. L., & Winckler, C. (2023). Organização das ações em sistemas ofensivos no goalball. In Universidade Federal de Juiz de Fora, Universidade Paulista, Universidade Federal de São Paulo.

Anagnostopoulou, A., Gouvea, T. S., & Sonntag, D. (2023). Enhancing journalism with AI: A study of contextualized image captioning for news articles using LLMs and LMMs. In DFKI German Research Center for Artificial Intelligence, Applied Artificial Intelligence, Carl von Ossietzky University Oldenburg.

Bianco, S., Celona, L., Donzella, M., & Napoletano, P. (2023). Improving image captioning descriptiveness by ranking and LLM-based fusion. In Department of Informatics, Systems and Communication, University of Milano-Bicocca.

Kilickaya, M., Erdem, A., Ikizler-Cinbis, N., & Erdem, E. (2023). Re-evaluating automatic metrics for image captioning. In Hacettepe University Computer Vision Lab, Department of Computer Engineering, Hacettepe University.

Santaella, L., & Kaufman, D. (2023). A inteligência artificial generativa como quarta ferida narcísica do humano. In Pontifícia Universidade Católica de São Paulo, São Paulo, Brasil.

BERNARDI, R. (2016). Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures. Journal of Artificial Intelligence Research, 55, 409-442.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Language Technologies Institute, Carnegie Mellon University.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In Hugging Face, Brooklyn, USA.

Gavião de Almeida, J. J. (2012). Os processos auto-organizacionais do goalball. Revista Brasileira de Ciências do Esporte, 34(3), 741-760.

Rao, A., Wu, Y., Li, S., & Liu, Y. (2024). MatchVoice: Temporal Alignment for Automatic Soccer Commentary Generation. In Proceedings of the Association for Computational Linguistics (ACL).

Mkhallati, H., Cioppa, A., Giancola, S., Ghanem, B., & Van Droogenbroeck, M. (2023). SoccerNet-Caption: Dense Video Captioning for Soccer Broadcasts Commentaries. In Computer Vision and Pattern Recognition (CVPR).

Apêndice A - exibição dos quadros com as descrições do modelo

Abaixo são exibidos alguns dos quadros com as descrições geradas pelo modelo. Cada quadro exibe a descrição que o modelo retornou mediante o quadro, seguido das métricas METEOR e ROUGE-L.

Frame 3: a volleyball player is jumping up to hit the ball

METEOR score: 0.51
ROUGE-L score: 0.63

Figura 7. Tradução: “um jogador de vôlei está saltando para bater na bola”.

Frame 24: the volleyball players are celebrating after a point

METEOR score: 0.39
ROUGE-L score: 0.35

Figura 8. Tradução: “Os jogadores de vôlei estão comemorando depois de um ponto.”

Frame 16: a soccer player is kicking the ball while another player is trying to block it

METEOR score: 0.54
ROUGE-L score: 0.50

Figura 9. Tradução: “Um jogador de futebol está tentando chutando a bola enquanto outro jogador está tentando o bloquear.”

Frame 22: two soccer players are celebrating after a goal

METEOR score: 0.46
ROUGE-L score: 0.35

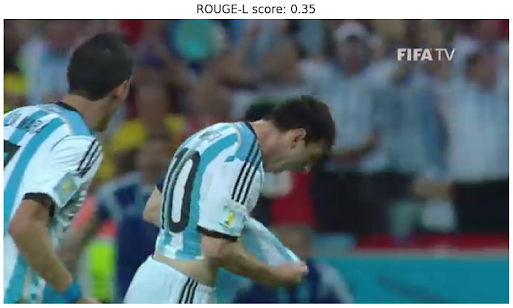


Figura 10. Tradução: “Dois jogadores de futebol estão celebrando depois de um gol.”

Frame 17: a group of people playing soccer on a court

METEOR score: 0.15
ROUGE-L score: 0.08

Figura 11. Tradução: “Um grupo de pessoas estão jogando futebol em uma quadra.”

Frame 1: two people in green shirts playing soccer on the court

METEOR score: 0.29
ROUGE-L score: 0.07

Figura 12. Tradução: “Duas pessoas de camisas verdes estão jogando futebol em uma quadra.”