# Análise da Atribuição de Significado Durante o Processo de Embedding em Modelos de Visão Computacional e Linguagem Natural Aplicados ao Meio Urbano

Graduando *Mateus Matzkin Gurza*, orientado por *Roberto Marcondes César Júnior*

**Resumo**

Este projeto de iniciação científica investiga o uso de visão computacional e inteligência artificial para analisar cenários urbanos, com foco na detecção e descrição de entidades em ambientes complexos. O objetivo é avaliar como diferentes modelos representam a semântica das imagens e identificar perdas de informação no processo de codificação para vetores compactos e na geração de descrições textuais. Utilizando dados reais de ambientes urbanos e suas particularidades, a pesquisa busca contribuir para o desenvolvimento de cidades inteligentes mais acessíveis, seguras e sustentáveis, otimizando o uso de IA no planejamento e gestão urbana.

## 1. Introdução

O desenvolvimento de tecnologias baseadas em visão computacional tem mostrado um crescimento significativo nos últimos anos, especialmente no contexto de aplicações urbanas. As cidades, como epicentros da vida moderna, estão cada vez mais dependentes da inovação tecnológica para lidar com desafios de mobilidade, acessibilidade e segurança. Nesse cenário, a inteligência artificial (IA) se tornou uma ferramenta importante para pensar o futuro da eficiência dos sistemas urbanos.

O presente projeto de iniciação científica se insere na temática da utilização de visão computacional e modelos de detecção e descrição para a análise das cidades e suas particularidades. O objetivo principal é aplicar e avaliar tanto modelos de Convolutional Neural Networks (CNN's)[1], usados em problemas de detecção e tracking[2], quanto modelos de descrição de imagem, em circunstâncias realistas, e comparar resultados, no que tange a capacidade desses diferentes modelos de representar o espaço visualizado, ampliando o domínio do uso da IA na exploração do meio urbano. Por meio dessa abordagem prática, espera-se obter dados e indicadores relevantes sobre os cenários de estudo e sobre a forma como diferentes modelos representam a semântica encontrada em cada imagem (neste caso, do ambiente urbano), e qual a perda de informação durante a abstração da imagem para a sua representação em vetor de "embedding". No processo, espera-se entender formas de aprimorar a maneira como empregamos algoritmos de detecção das entidades que interagem com a esfera sócio-urbana.

Este estudo pretende colaborar com o ecossistema de pesquisa centrado na viabilização de cidades mais inteligentes, acessíveis, sustentáveis, seguras e com maior mobilidade, através do uso de Inteligência Artificial. Especificamente, o projeto está inscrito na produção acadêmica

[1] Redes neurais convolucionais profundas

[2] Tracking seria o rastreamento do objeto em movimento em um vídeo

do grupo Creativision da Universidade de São Paulo. O núcleo é responsável por pesquisas avançadas nas áreas de visão computacional e modelos multimodais na USP. Considerando o exposto, o trabalho, aqui proposto, também terá como foco a interação com outras iniciativas no campo.

## 2. Revisão Bibliográfica

O intuito desta revisão bibliográfica é expor as linhas de atuação e iniciativas do mundo acadêmico no que tange a implementação de métodos de IA, com foco em visão computacional, nas cidades. Deseja-se delimitar o estado da arte dessa ferramenta e levantar casos de uso relevantes nesse contexto. Com isso feito, será possível, posteriormente, situar o trabalho proposto neste plano de pesquisa, entendendo onde ele se insere e a que fins contribuirá.

Para situar o leitor acerca dos diferenciais do uso de IA para o estudo do meio sócio-urbano, abordar-se-á, de antemão, um rápido exemplo. Imagine que se quer saber qual caminho conectando o ponto A ao ponto B é mais acessível para idosos. Uma forma antiquada de realizar isso seria fazer uma pessoa percorrer e avaliar todos esses percursos para depois comparar cada caminho. Esse método obsoleto tem como desvantagens: a interferência da subjetividade, dada a inclusão do avaliador humano e sua interpretação do que seria uma obstrução hostil a um idoso; o grande consumo de tempo, dada a necessidade de calcular parâmetros que permitam comparar os trajetos; e a ausência de uma padronização, dada a dificuldade de analisar igualmente todos os percursos. Outra maneira, mais moderna, de abordar esse problema seria utilizar visão computacional em vídeos desses trajetos para identificar obstáculos neles e comparar, de maneira automática, parâmetros obtidos de forma padronizada sobre todos os caminhos percorridos, podendo até mesmo obter descrições desses percursos usando LLMs[3]; indicando, por fim, qual o mais acessível.

Assim, a IA aumenta a eficiência com a qual podemos analisar e identificar indicadores importantes sobre as cidades. Esse ganho em poder de análise se traduz em um potencial de entender e, portanto, impactar diversos aspectos do urbanismo através do design e controle urbano. Entre eles, a acessibilidade é um fator sob o qual a utilização da inteligência artificial pode ter grande efeito, servindo para guiar a construção de "human-centered smart cities"[4], conforme explicado em "The future of Urban Accessibility: The Role of AI"[2] (*FROEHLICH* et al., 2024). A capacidade da IA de gerar conhecimento para guiar decisões urbanísticas e logísticas, possibilitará a transformação do ambiente da cidade em um espaço de utilidade pública mais seguro e acessível. Para que se alcance esse objetivo, é necessário, então, através de machine learning, capturar indicadores importantes sobre as interações entre os elementos constitutivos do espaço estudado, por exemplo: calçadas, pedestres, carros, ruas e outras infraestruturas e agentes; para que assim seja possível usar os parâmetros encontrados para facilitar a tomada de decisões estratégicas sobre esse ambiente.

O estudo das redes de transporte, por exemplo, é um campo onde iniciativas direcionadas para a captura desses indicadores vem ocorrendo por meio de ferramentas baseadas em visão computacional, tal como os métodos de video-based "Automated Passenger Counting"(APCs) (*PRONELLO*; *GARZÓN*, 2020). Entre as infraestruturas que compõe o meio urbano, as redes de transporte podem ser facilmente associadas às temáticas de: sustentabilidade, pela emissão de gases de estufa; desigualdade socio-espacial, pela capacidade - ou sua ausência - de conectar pontos distantes; e segurança, pelas interações entre veículos e entre estes e pedestres. Visto isso, a aplicação de IA para analisar essas redes e obter parâmetros que permitam pensar em formas de melhorá-las é crucial para a consolidação de "Inteligent Transport Systems (ITS)"[5], uma das formas como, na atualidade, pode-se, segundo Pronello e Garzón (2023,

[3]Large Language Models

[4]Tradução Livre: cidades inteligentes centradas nas pessoas

[5]Tradução Livre: Sistemas de Transporte Inteligentes

p. 1) ”contribute to a safer, more efficient, and environmentally friendly transport network.”[6] e assim, ”make different aspects of urban mobility smarter...”[7].

Além das redes de transporte, a infraestrutura para pedestres tem uma posição central na forma como as pessoas interagem com as cidades. Entender os problemas de acessibilidade e segurança associados às condições físicas e distribuição espacial dessa infraestrutura é essencial para guiar a maneira como ela será repensada, em prol da resolução desses problemas, no futuro. Nesse cenário, métodos de mapeamento por imagens aéreas para a produção de dados geoespaciais sobre as calçadas, por exemplo, é uma das aplicações de visão computacional destinada à construção do entendimento mais claro desta face do meio urbano; entendimento, esse, que busca corroborar com o avanço de ”pedestrian-oriented city designs”e a melhora do ambiente para quem se desloca à pé na cidade. Em ”Towards global-scale crowd + AI techniques to map and assess sidewalks for people with disabilities”(*HOSSEINI* et al., 2023) pode-se enxergar esse esforço, por meio da aplicação de IA, direcionado ao mapeamento e avaliação da infraestrutura em foco.

Entende-se que o papel da Inteligência Artificial neste contexto é o de produzir eficientemente parâmetros e indicadores que tornem mais tangível a análise do espaço urbano. Facilitar a visualização e interpretação dessas informações capturadas pelos algoritmos acerca do ambiente é um fator que incentiva a criação de novas tecnologias. Exemplos dessa influência podem ser vistos na busca de melhorar o mapeamento de calçadas, capturando as conexões da rede que interliga elas, como no projeto ”Mapping the walk”(*HOSSEINI* et al., 2023), e até na na elaboração de novos frameworks de visualização, como no Urban Toolkit para visual analytics, que tem o objetivo de facilitar a integração e visualização de dados urbanos distintos (*MIRANDA* et al., 2024). Ambos projetos, entre muitos outros no campo, tem como intuito a consolidação de formas claras de utilizar a informação obtida por IA para a tomada de decisões de design urbano e logística, facilitando a descoberta de insights claros e acionáveis.

Considerando a importância da produção sistemática de indicadores precisos através dessas novas tecnologias, obter dados que descrevam com exatidão as particularidades de cada circunstância explorada no ambiente urbano é uma necessidade, visto que estes são cruciais para a construção de modelos mais exatos das experiências na cidade. Em outras palavras, a evolução da coleta de dados é essencial porque a qualidade deles e a disponibilidade de fontes multimodais são imperativos para a eficiência de algoritmos de aprendizado de máquina. No entanto, essa evolução não é uma tarefa fácil; o ambiente das cidades é repleto de obstáculos e condições adversas à aplicação de IA. Podemos ressaltar a poluição visual, sobreposição de objetos, perspectivas e ângulos de visão pouco convencionais, entre outros desafios. Projetos como o Streetaware (*PIADYK* et al., 2023), procuram, nesse contexto, disponibilizar dados de alta qualidade para a possibilitar a extração de insights mais claros sobre o espaço urbano e assim elevar o patamar da IA aplicada a este contexto. O Streetaware, especificamente, realiza essa tarefa para o cenário de interseções urbanas movimentadas, como em Brooklyn, Nova York. O projeto apresenta dados multimodais de alta resolução — incluindo áudio, vídeo e LiDAR — em três interseções, totalizando cerca de 8 horas de dados. Esses dados foram capturados por sensores sincronizados, oferecendo múltiplas perspectivas e modalidades de informação, possibilitando uma análise mais detalhada, como a descoberta de objetos oclusos, associação de eventos sonoros a suas representações visuais, rastreamento de objetos ao longo do tempo e medição da velocidade de pedestres. Esse tipo de recursos, como explicado, é muito valioso para estudos de interações urbano-ambientais e a aplicação de aprendizado de máquina nesse contexto.

Ademais, ainda sobre a captura de dados, coletá-los de forma realista, eficiente e personalizada para abordar os problemas que os métodos de IA se propuserem a mapear, é uma

[6]Tradução Livre: contribuir para uma rede de transporte mais segura, eficiente e ambientalmente amigável.

[7]Tradução Livre: tornar diferentes aspectos da mobilidade urbana mais inteligentes...

tarefa imprescindível para o avanço dos estudos do meio urbano. Projetos como o SideSeeing (*DAMACENO* et al., 2024), que apresenta um framework para coleta, processamento e análise de dados de uma perspectiva ”egocêntrica- baseada na perspectiva do pedestre e no ambiente como qual interage diretamente - representam o movimento de especialização e personalização dos dados para que apresentem mais precisamente a realidade das circunstâncias a serem analisadas através de visão computacional e modelos multimodais; No caso do SideSeeing, essa personalização busca aproximar a informação o máximo possível da vivência do humano inserido nos múltiplos cenários das cidades, entendendo as características dos percursos que tomam para ir para o hospital, por exemplo, capturando as nuances da perspectiva do individuo sujeito ao trajeto e, assim, possibilitando um análise mais detalhada das condições de acessibilidade de rotas importantes para a saúde pública e segurança.

Em suma, pesquisadores de diferentes áreas, impulsionados pela necessidade de repensar as cidades de forma mais inteligente para que estas atendam eficientemente a população, vem se dedicando a projetos que utilizam AI para viabilizar isso. Nesse cenário, a visão computacional emerge como uma ferramenta fundamental. Diversos estudos destacam o uso dela em soluções urbanas, com ênfase na análise de redes de transporte, infraestrutura para pedestres e mapeamento de espaços públicos. Pesquisas recentes abordam desde o monitoramento de tráfego até a avaliação da acessibilidade em calçadas, revelando como a tecnologia permite a análise da cidade em prol de melhorar a experiência cotidiana para indivíduos em diversas condições de vida.

Consolidado o entendimento de que a Visão Computacional é uma ferramenta de grande importância para a vida urbana, com aplicações crescentes e promissoras, torna-se evidente que o aprimoramento da capacidade de um modelo em compreender e representar integralmente uma imagem e suas informações tem impacto direto na sua utilidade. Em outras palavras, para avançar na aplicação da Inteligência Artificial na esfera socio-urbana, é essencial compreender como os modelos atribuem significado e relevância às informações contidas em uma imagem.

Nesse contexto, dois grandes desafios emergem em relação à forma como as IAs interpretam os dados processados. De maneira geral, no campo do deep learning, destaca-se o problema da Baixa Explicabilidade. Esse desafio está relacionado à dificuldade de entender os processos que levam o modelo a tomar determinadas decisões. A falta de explicabilidade torna mais difícil identificar quais características das informações originais processadas influenciam os resultados obtidos, comprometendo a avaliação da confiabilidade do modelo.

Já no domínio específico da Visão Computacional, surge a questão da Atenção, que aborda como os modelos atribuem importância às diferentes partes de uma imagem. Nem sempre os aspectos mais relevantes para o significado de uma imagem são os centralizados no quadro ou os mais óbvios de descrever. Entretanto, para interpretar corretamente uma imagem, é crucial identificar as características principais que definem sua semântica.

Na aplicação de visão computacional em ambientes urbanos podemos pontuar mais alguns desafios com efeitos adversos à correta interpretação de imagens. Podem ser citados: poluição visual, ângulos de captação de dados não convencionais, dados que não refletem a realidade das circunstâncias sendo analisadas, sobreposição de objetos e condições climáticas ruins, como chuva ou baixa luminosidade.

Projetos recentes, como o Streetaware e o SideSeeing, ilustram o esforço em superar essas dificuldades, ressaltando a importância da adaptação das tecnologias para as especificidades do ambiente urbano, garantindo a captura realista e eficiente dos dados necessários para gerar insights acionáveis.

Por fim, este projeto propõe aprofundar o estudo sobre a aplicação de modelos de visão computacional no contexto urbano, com foco na detecção e interpretação de entidades em movimento nesse ambiente. O objetivo é investigar como a semântica desses cenários é definida e avaliar a quantidade e o tipo de informações perdidas durante o processamento, especialmente

no que diz respeito à capacidade de descrevê-los de forma precisa. A seguir, será apresentado o escopo da pesquisa, detalhando como este trabalho pretende contribuir para o avanço da inteligência artificial como ferramenta estratégica para a análise e compreensão das dinâmicas urbanas.

## 3. Escopo do trabalho

Este projeto de iniciação científica, inserido no ecossistema do das produções acadêmicas do núcleo Creativision da Universidade de São Paulo (USP), tem como objetivo avaliar a capacidade que modelos de visão computacional e LLMs, reconhecidos como estado da arte, tem de codificar e representar a semântica de imagens no processo de detecção de entidades e descrição de cenários do meio sócio-urbano. Em especial, o foco será a análise de dados provenientes de vídeos de regiões naturalmente movimentadas nas cidades. O intuito dessa iniciativa será estudar os conceitos e fundamentos de redes neurais convolucionais de detecção de objetos e modelos de linguagem natural que realizam o processo de embedding de imagens, sendo assim responsáveis pela representação da semântica do quadro processado. A pesquisa se propõem a, no percurso, avaliar o desempenho e aplicabilidade desses modelos de forma comparativa em condições marcadas pelas particularidades do meio urbano. Serão utilizados algoritmos de ponta atuais, como a versão mais recente do YOLO (You Only Look Once), EfficientDet ou SSD (Single Shot MultiBox Detector) e modelos open source de ponta, como os Llama ou os providos pela API do Keras, podendo também citar as redes neurais CLIP da OpenAI , para o processamento de cenários com automóveis e pedestres.

A análise será inicialmente realizada para casos base simplificados, ou seja, sem movimento de câmera e utilizando apenas vídeos dentre as modalidades de informação capturadas. Os dados utilizados serão aqueles captados pelo grupo Creativision e pela iniciativa StreetAware de NYU (New York University), os quais também estão sendo utilizados em outros projetos. Para os dados coletados pelo grupo Creativision, o método de obtenção é baseado em 3 câmeras fixas dispostas estrategicamente em diferentes perspectivas de um mesmo ponto; isso será feito para múltiplos locais movimentados do campus Butantã da USP. Este setup inicial permitirá entender os efeitos diretos - sem interferência de fatores como movimento da câmera ou informações não visuais - das condições do ângulo de visão na tarefa de detecção. Concomitantemente, esse caso base introduzirá as dificuldades intrínsecas do cenário da cidade em circunstâncias de baixa luminosidade, chuva, poluição visual, entre outros. Segue uma imagem na qual é possível identificar pontos de oclusão (barras metálicas do ponto de ônibus), ângulos de perspectiva não convencionais (visualização da parte traseira do caminhão), ambientes poluídos e objetos pequenos (carro e pessoa à distância no fundo da imagem). Vale ressaltar que problemas similares são encontrados nas capturas associadas à iniciativa StreetAware, com a particularidade de que o ambiente se refere à cidade de Nova York.

Figura 1: Imagem representativa de captura realizada, pelo Creativision, nas condições explicadas.

O objetivo inicial da aplicação dos modelos será identificar e descrever o ambiente urbano e os agentes que interagem com ele. Após caracterizar a cena, composta principalmente pela presença de pessoas e veículos, a pesquisa buscará avaliar se as descrições geradas representam, de forma fiel, as partes relevantes da imagem e seu significado, além de identificar quais informações se perdem durante o processo de codificação da imagem em vetores de embedding e na descrição da cena urbana.

Para alcançar esses objetivos, serão estabelecidos frameworks claros de avaliação, juntamente com métricas específicas que permitam comparar o desempenho dos modelos na tarefa de codificar a semântica de uma imagem em diferentes cenários. Essa abordagem possibilitará uma análise crítica e abrangente dos resultados, fornecendo meios para quantificar a perda de informação ao longo do processo de caracterização do ambiente urbano.

Como dito anteriormente, este projeto se insere em um contexto de colaboração e interdisciplinaridade, interagindo com outras pesquisas em andamento no Creativision. Essa sinergia permitirá a troca de informações e técnicas entre projetos que também exploram a coleta e interpretação de dados no ambiente urbano, incluindo o mapeamento das características das infraestruturas urbanas. Com isso em mente, existem perspectivas de expansão dos casos abordados nesta pesquisa para abranger os frameworks e ferramentas em desenvolvimento no Laboratório, como por exemplo o SideSeeing. Assim, dada a interlocução com outros esforços na área de IA, associados à rede integrada do Creativision e acadêmicos parceiros, espera-se que este trabalho contribua para o avanço da aplicação de visão computacional em ambientes urbanos, ajudando a identificar limitações e potencialidades dos algoritmos existentes e, também, buscando fornecer insights para o desenvolvimento de abordagens mais robustas e adaptáveis ao contexto das cidades.

## 4. Método

A detecção de objetos é uma tarefa central na Visão Computacional, com o objetivo de identificar e localizar elementos de interesse em imagens ou vídeos. Esse processo, amplamente explorado na literatura, geralmente utiliza caixas delimitadoras para circunscrever os objetos detectados, às quais são atribuídos rótulos que indicam suas classes. Para realizar essa tarefa de forma eficiente, modelos de redes neurais profundas, como o YOLO (You Only Look Once), são amplamente empregados devido à sua abordagem de ponta a ponta. Esses modelos combinam tarefas de classificação, que identificam a classe de um objeto, e regressão, que determina a posição e o tamanho das caixas delimitadoras. Uma vez treinados, esses modelos são capazes de processar novas imagens e vídeos, gerando listas de detecções acompanhadas de suas respectivas pontuações de confiança.

Apesar de sua eficácia, a detecção de objetos enfrenta diversos desafios, como variações de escala, oclusões parciais, alterações na aparência dos objetos (cor, textura, forma ou ângulo de visão) e a qualidade das imagens capturadas. Esses fatores afetam diretamente a capacidade do modelo de identificar os elementos mais relevantes de uma cena. No contexto urbano, isso se torna ainda mais complexo devido à densidade e à diversidade das interações presentes, como a coexistência de veículos, pedestres e estruturas físicas. Esses fatores afetam diretamente o processo de codificação e a interpretação semântica das imagens e superar esses desafios é essencial para criar representações precisas das cenas urbanas e facilitar sua posterior descrição e análise. Em outras palavras, descrever corretamente um cenário urbano depende da capacidade do modelo de identificar e priorizar os componentes mais relevantes na cena.

No cerne deste projeto está o uso de espaços de embedding para traduzir imagens complexas em representações vetoriais compactas. Embeddings são vetores numéricos que capturam as características mais relevantes de uma imagem, simplificando sua estrutura original em uma forma que pode ser comparada e analisada. Esses embeddings servem como base para descrever o conteúdo das imagens por meio de Modelos de Linguagem de Grande Escala

(LLMs), que geram descrições textuais detalhadas das cenas. O objetivo da pesquisa é compreender como essas representações vetoriais preservam (ou perdem) informações críticas durante o processo de codificação e como isso impacta a fidelidade das descrições textuais geradas.

A transformação de uma imagem em um vetor de embedding envolve, inevitavelmente, uma simplificação das informações originais. Nesse processo, parte da semântica pode ser perdida, e compreender a extensão dessa perda é um objetivo central do projeto. Para quantificar essa perda, a pesquisa fará uso de métricas baseadas no conceito de entropia. Conceitos como a entropia de Shannon e a divergência de Kullback-Leibler, provenientes da Teoria da Informação, por exemplo, serão usados para medir a perda das informações dentro dos embeddings, indicando o grau de compactação das características mais relevantes das imagens e, consequentemente as possíveis falhas dos modelos em representar corretamente a semântica de determinados cenários.

Os embeddings gerados também serão usados como entrada para os LLMs, que traduzirão essas representações em descrições textuais das cenas urbanas. A qualidade das descrições será avaliada comparando-se os textos gerados e as imagens correspondentes. A ideia é produzir uma espécie de métrica de entropia cruzada por meio da qual, usando as distâncias euclidianas entre vetores de embedding de N descrições e as distâncias euclidianas para os embeddings das N imagens originais e comparando-as, será possível quantificar a informação perdida por meio das variações nas distâncias.

O processo completo será conduzido em quatro etapas principais: (1) coleta de dados, utilizando imagens e vídeos capturados por iniciativas como StreetAware e Creativision; (2) processamento das imagens com modelos para gerar embeddings; (3) transformação desses embeddings em descrições textuais por meio de LLMs; e (4) avaliação dos resultados com base em métricas de entropia, divergência de Kullback-Leibler e similaridade ou diferença de distâncias euclidianas entre os vetores.

A análise detalhada dessas etapas permitirá uma compreensão mais profunda da relação entre a qualidade dos embeddings, a precisão das descrições textuais e o impacto dessas representações no entendimento de cenas urbanas. Esse método não apenas oferecerá insights sobre a eficiência dos modelos utilizados, mas também estabelecerá uma base sólida para o desenvolvimento de sistemas mais avançados e confiáveis para análise urbana baseada em visão computacional e inteligência artificial.

## 5. Cronograma

Nessa seção serão abordadas as etapas do projeto de pesquisa e sua distribuição ao longo do próximos trimestres. Vale ressaltar que novas etapas e aprofundamentos podem ser adicionadas ao cronograma. A extensão do projeto leva em consideração tanto essas possibilidades quanto a contribuição com outras atividades dentro do grupo Creativision.

| Cronograma Trimestral | $1^o$ | $2^o$ | $3^o$ | $4^o$ | $5^o$ | $6^o$ | $7^o$ | $8^o$ |
|---|---|---|---|---|---|---|---|---|
| Estudo dos fundamentos e conceitos básicos de LLMs e CNNs | x | | | | | | | |
| Obtenção dos videos e bases dados | x | | | | | | | |
| Produção de scripts e framework de geração de embeddings | x | x | | | | | | |
| Teste dos scripts e avaliação de resultados iniciais | | x | x | | | | | |
| Estudo e elaboração das métricas de avaliação de entropia e perda de informação | | | | x | x | | | |
| Reavaliação e melhoria da estrutura e framework testados | | | | | x | x | | |
| Aplicação em larga escala ao banco de dados de videos do StreetAware | | | | | | x | x | |
| Avaliação dos resultados e redação do relatório científico | | | | | | | | x |

- **Estudo dos fundamentos e conceitos básicos**

  O aluno iniciará o projeto com o estudo de conceitos fundamentais de aprendizado de máquina, incluindo redes neurais convolucionais (CNNs), funções de perda, otimização por descida do gradiente, e métricas de avaliação de desempenho. Paralelamente, serão explorados conceitos relacionados a Modelos de Linguagem de Grande Escala (LLMs) para melhor compreender como esses modelos podem ser aplicados na análise de imagens e vídeos.

- **Obtenção dos vídeos e bases de dados**

  Serão coletados os dados necessários para o desenvolvimento do projeto, utilizando vídeos e imagens provenientes do grupo Creativision (USP) e da iniciativa StreetAware (NYU). Essa etapa incluirá a organização e o pré-processamento dos dados, com foco em garantir a qualidade e relevância das amostras para os experimentos propostos.

- **Produção de scripts e framework de geração de embeddings**

  Durante este período, serão desenvolvidos scripts e frameworks para geração de embeddings de imagens, utilizando redes neurais convolucionais e ferramentas modernas de visão computacional. A implementação se concentrará na extração de representações vetoriais compactas que preservem as informações semânticas dos quadros analisados.

- **Teste dos scripts e avaliação de resultados iniciais**

  Após a implementação inicial dos scripts, será realizada uma bateria de testes com dados controlados para validar os frameworks criados. Esses testes avaliarão a capacidade dos modelos de gerar embeddings precisos e confiáveis, permitindo ajustes e refinamentos necessários na estrutura.

- **Estudo e elaboração das métricas de avaliação de entropia e perda de informação**

  Com base na teoria da informação, será feita uma análise detalhada para definir métricas como entropia de Shannon e divergência de Kullback-Leibler, que servirão para quantificar a perda de informação no processo de codificação das imagens. Esse estudo incluirá tanto a formulação das métricas quanto sua validação com dados reais.

- **Reavaliação e melhoria da estrutura e framework testados**

  Após o desenvolvimento das métricas, os frameworks serão reavaliados e otimizados com base nos resultados obtidos nos testes. Serão implementadas melhorias nos algoritmos e ajustados os processos de extração de embeddings e geração de descrições semânticas.

- **Aplicação em larga escala ao banco de dados de vídeos do StreetAware**

  Os modelos e frameworks otimizados serão aplicados ao banco de dados de vídeos da iniciativa StreetAware, utilizando diferentes cenários urbanos. Essa aplicação em larga escala permitirá avaliar a robustez e a generalização das técnicas desenvolvidas em ambientes mais complexos e variados.

- **Avaliação dos resultados**

  Os resultados obtidos serão avaliados de forma crítica, utilizando métricas quantitativas (como entropia e divergência de Kullback-Leibler) e qualitativas (análise da qualidade das descrições textuais geradas por LLMs). Será dada ênfase à identificação de padrões de perda de informação e limitações dos modelos em cenários urbanos.

- **Redação do relatório científico**

  Com o término das análises, será produzido um relatório científico detalhado, documentando as metodologias utilizadas, os resultados obtidos, as conclusões e as perspectivas para trabalhos futuros. Esse relatório servirá como a principal entrega do projeto.

## 6. Justificativa e Motivação

Considerando os pontos levantados durante a revisão bibliográfica, esta seção apresenta os motivos que fundamentam a pesquisa proposta, justificando-a frente ao cenário atual do uso de inteligência artificial (IA) em aplicações relacionadas ao urbanismo. Atualmente, a visão computacional, associada a outros métodos de IA, destaca-se como uma ferramenta poderosa para aprofundar o entendimento das dinâmicas urbanas. Essa capacidade se dá, principalmente, pelas seguintes características: a produção padronizada e sistemática de análises sobre os dados capturados; a automação do processamento de informações complexas; e a consolidação de resultados analíticos isentos de subjetividade humana, permitindo uma abordagem mais precisa e objetiva.

No contexto urbano, essas dinâmicas incluem as interações entre diferentes entidades e infraestruturas, como pedestres, veículos, calçadas e redes de transporte. Tais interações definem experiências que variam entre os indivíduos, dependendo de fatores como idade, condições físicas, e contextos socioeconômicos e espaciais. Entender essas interações é essencial para planejar e construir cidades que sejam acessíveis, seguras, sustentáveis e inclusivas. A acessibilidade, por exemplo, pode variar amplamente dependendo da mobilidade física de uma pessoa, enquanto questões como segurança no trânsito e conforto ambiental também impactam grupos sociais de formas distintas. Nesse sentido, a capacidade de medir e analisar essas condições de forma confiável é essencial para a tomada de decisões baseadas em dados na engenharia e no design do meio socio-urbano.

A inteligência artificial se apresenta como um recurso indispensável para abordar esses desafios, permitindo a coleta, processamento e análise sistemática de grandes volumes de dados urbanos. Modelos de IA podem identificar padrões e nuances que escapam à análise humana, além de fornecer insights em tempo real, automatizar fluxos de informações e gerar métricas quantitativas e qualitativas sobre o espaço urbano. Essas métricas possibilitam representar numericamente, por meio de índices, ou categoricamente, por classificações, as condições e fenômenos urbanos. Isso torna mais tangível a análise de cenários urbanos e facilita a comunicação de informações críticas para a tomada de decisão. Tarefas como essas são

difíceis ou impraticáveis sem a utilização de aprendizado de máquina, o que destaca o papel da IA como ferramenta estratégica para capturar e compreender as dinâmicas das cidades.

Diante da necessidade de compreender as interações humanas com o meio urbano para torná-lo mais equitativo e sustentável, as capacidades analíticas da IA tornam-se cada vez mais relevantes. No entanto, para que essas tecnologias sejam plenamente integradas ao planejamento urbano, é crucial compreender os desafios associados à sua aplicação. Essa compreensão permitirá superar limitações, melhorar a eficácia dos modelos e orientar a evolução das ferramentas existentes.

A presente pesquisa se justifica por sua contribuição no sentido de consolidar o uso sistêmico de inteligência artificial como método para promover avanços na análise urbana. O foco em modelos de visão computacional de última geração e LLMs, visa explorar e otimizar sua aplicação em tarefas específicas de detecção e descrição de cenários urbanos, abrindo espaço para entender o que é necessário aprimorar no processo de atribuição de significado por modelos de IA aplicados ao meio urbano. Além disso, busca-se expandir o entendimento sobre a avaliação e a aplicabilidade desses modelos no contexto real, particularmente em condições adversas comuns nas cidades, como baixa luminosidade, oclusões e poluição visual. Assim, este trabalho não apenas promete contribuir para o desenvolvimento de técnicas de IA mais robustas e adaptáveis, mas também apoia a criação de ferramentas que forneçam insights críticos para o planejamento e a gestão de cidades inteligentes, inclusivas e eficientes, alinhadas às demandas de um futuro urbano sustentável.

## Referências


[1] R. J. P. Damaceno et al. *SideSeeing: A multimodal dataset and collection of tools for sidewalk assessment*. 2024. arXiv: 2407.06464 [cs.CV]. URL: https://arxiv.org/abs/2407.06464.

[2] Jon E Froehlich et al. "The Future of Urban Accessibility: The Role of AI". Em: (2024).

[3] Maryam Hosseini et al. "Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery". Em: *Computers, Environment and Urban Systems* 101 (2023), p. 101950. ISSN: 0198-9715. DOI: https://doi.org/10.1016/j.compenvurbsys.2023.101950. URL: https://www.sciencedirect.com/science/article/pii/S0198971523000133.

[4] Maryam Hosseini et al. "Towards global-scale crowd+ AI techniques to map and assess sidewalks for people with disabilities". Em: *arXiv preprint arXiv:2206.13677* (2022).

[5] Deepak Kumar Jain et al. "Robust multi-modal pedestrian detection using deep convolutional neural network with ensemble learning model". Em: *Expert Systems with Applications* 249 (2024), p. 123527.

[6] Yangke Li e Xinman Zhang. "Multi-modal deep learning networks for RGB-D pavement waste detection and recognition". Em: *Waste Management* 177 (2024), pp. 125–134.

[7] Fabio Miranda et al. "The Urban Toolkit: A Grammar-Based Framework for Urban Visual Analytics". Em: *IEEE Transactions on Visualization and Computer Graphics* 30.1 (2024), pp. 1402–1412. DOI: 10.1109/TVCG.2023.3326598.

[8] Yurii Piadyk et al. "Streetaware: A high-resolution synchronized multimodal urban scene dataset". Em: *Sensors* 23.7 (2023), p. 3710.

[9] Caio Pieroni et al. "Big data for big issues: Revealing travel patterns of low-income population based on smart card data mining in a global south unequal city". Em: *Journal of Transport Geography* 96 (2021), p. 103203.

[10] Cristina Pronello e Ximena Rocio Garzón Ruiz. “Evaluating the Performance of Video-Based Automated Passenger Counting Systems in Real-World Conditions: A Comparative Study”. Em: *Sensors* 23.18 (2023). ISSN: 1424-8220. DOI: 10.3390/s23187719. URL: https://www.mdpi.com/1424-8220/23/18/7719.

[11] Claudio T Silva et al. “Integrated Analytics and Visualization for Multi-modality Transportation Data”. Em: (2019).