Proposta de Projeto de Iniciação Científica

Predizendo níveis de pobreza utilizando imagens de satélites e formulários socieconômicos.

João Pedro Donaire Albino

Orientador:
Prof. Dr. João Paulo Papa

Departamento de Computação, Faculdade de Ciências,
Universidade Estadual Paulista Júlio de Mesquita Filho
Av. Eng. Luiz Edmundo Carrijo Coube, 14-01 - Vargem Limpa,
CEP 17033-360 - Bauru - SP.

22 de Julho de 2018

1 Introdução

A pobreza é uma das chagas complexas e importantes de nossa sociedade atualmente, pois se trata de um problema dificilmente controlado e trabalhado por métodos efetivos de combate [1]. Atrelado a isso também temos uma escassez de dados relevantes que possam mensurar com consistência indicadores de qualidade de vida e de poder aquisitivo da população. Esses dois fatores desencadeiam um notável esforço para mapear e diagnosticar o cenário da pobreza em determinadas regiões, assim perpetuando o problema da má distribuição de renda no país [2].

Não obstante, existem alguns estudos e métodos científicos que procuram aplicar a Tecnologia de Dados ao mapeamento da pobreza e, consequentemente, ao combate da mesma. Nesta linha de pesquisa temos alguns trabalhos exemplares como o estudo da utilização de Dados Móveis para verificar a qualidade de vida de terminada população [3] ou também a análise de mapas noturnos a fim de obter a intensidade luminosa e correlacioná-la ao índice de poder aquisitivo da região estudada [4].

Recentemente, Neal Jean, et al. desenvolveram um sistema que prediz níveis de pobreza utilizando imagens de satélites e dados de formulários socioeconômicos. Essa técnica obteve resultados promissores quando, ao avaliar mapas de países africanos como Uganda e Nigéria, por exemplo, pode explicar 75% da variação nos índices econômicos e, consequentemente, agrupar e identificar mais facilmente regiões onde a pobreza possui níveis alarmantes.

Por fim, o presente projeto tem a finalidade de aplicar tal método para expandir e otimizar os estudos no mapeamento e qualificação de índices de pobreza.

2 Justificativa

Em setembro de 2015, líderes dos 193 países membros da Organização das Nações Unidas (ONU) aprovaram um plano global de desenvolvimento sustentável, com o objetivo de melhorar os indicadores econômicos, sociais e ambientais para as próximas gerações [5]. Algo relevante para a proposta de pesquisa em questão é averiguar que a primeira dessas metas configura-se em eliminar todas as formas de pobreza no mundo. Essa decisão foi tomada ao levar em consideração que cerca de 705,5 milhões de pessoas vivem atualmente na extrema pobreza [6].

Para exemplificar através de dados locais, uma das principais dificuldades do Brasil reside em direcionar recursos para as população mais pobre [7]. Mesmo que através de formulários e pesquisas se torna possível identificar regiões com altos índices de pobreza, sofremos em obter dados relevantes que mensuraram, indicam e agrupam regiões por níveis de renda ou outras indicadores importantes.

Outro ponto relevante para se levar em consideração na proposta do trabalho diz respeito à quantidade de investimentos necessários para realizar uma pesquisa intensiva bem esclarecedora sobre indicadores de pobreza. Se levar-

mos em consideração países que não possuem recursos abundantes, obter dados relevantes se torna algo difícil. Segundo o próprio Instituto Brasileiro de Geografia e Estatística (IBGE), o orçamento do Censo de 2010 realizado no país fora calculado em R$ 1,677 bilhão [8].

Analisando tais fatos, chega-se em uma investida para inferir indicadores precisos utilizando dados *open-source* gratuitos já disponíveis na *Internet*. Esse é um dos tópicos que a Tecnologia de Dados que há algum tempo vem sendo trabalhado e será tratado dentro desta pesquisa.

2.1 Convolutional Neural Network

As Redes Neurais Convolucionais (ConvNets ou CNNs) são uma categoria de Redes Neurais capazes de alcançar resultados extremamente positivos em reconhecimentos de imagens baseando-se em dados complexos e utilizando apenas aprendizagem supervisionada [9]. Um exemplo da utilização de CNNs estão na identificação de rostos, objetos e sinais de trânsito, além da identificação remota em robôs e carros.

A CNN é caracterizada por um processamento de 3 passos totalmente interligados, por esse fato são chamadas de redes (Network, em inglês). O primeiro desses passos, denominado Convolação (*Convolutional*, em inglês), é o momento em que se recebe uma composição matricial baseada em pixels e cores e, por consequência, cria-se uma nova camada por um filtro desejado. O filtro pode ser direcionado a identificar bordas, limites, objetos, entre outros elementos, dependendo da finalidade do processo.

Em um segundo momento, a camada originada pelo filtro do passo anterior passa por uma nova filtragem chamada ReLU que substitui por zero todos os valores de pixel negativos no mapa de recursos. Em seu Ãºltimo passo, ocorre o agrupamento espacial (Spatial Pooling, em inglês) que mantém as informações mais importantes reduzindo a dimensionalidade de cada mapa de recursos.

Todos esses passos ao final possibilitam a formação de um resultado de saída (*output*, em inglês) de prescrição que visa identificar elementos presentes na imagem de entrada.

2.2 Transfer Learning

Outro conceito importante para a pesquisa se refere á Transferência de Aprendizagem (TL). Basicamente, TL nos permite lidar com base de dados diferentes alavancando os dados rotulados já¡ existentes de alguma tarefa ou domínio relacionado. Em suma, o objetivo dessa teoria é melhorar a aprendizagem na tarefa de destino agregando ao conhecimento da tarefa de origem [10].

Esse conceito no projeto auxilia na avaliação dos resultados pela CNN os correlacionando aos dados socieconômicos obtidos por questionários abertos, como Avaliação de Padrões de Vida (Living Standards Measurement Study ou LSMS) ou os próprios dados gerados pelo Censo de 2010 do IBGE.

3 Metologia

Como explicado anteriormente, o principal objetivo do projeto é continuar desenvolvendo estudos pré-existentes na área. Desta forma, podemos dividir a pesquisa em:

1. Treinamento de um modelo CNN para classificação de imagens;
2. Aplicar o modelo treinado de CNN em imagens de satélites noturnas e diurnas;
3. Correlacionar características aprendidas no passo 2 com as análises dos questionários socieconômicos.
4. Produzir indicadores e relações entre os dados estudados das imagens em questão.

Na primeira etapa da pesquisa é necessário um modelo bem treinado e preparado para identificar elementos em imagens de satélite, possibilitando no final um sistema útil para identificação de níveis de pobreza. Portanto, em uma primeira instância, deve-se utilizar um modelo de CNN treinado a fim de identificar características básicas de imagens. A linguagem escolhida para construir o sistema será o Python. Tal escolha baseia-se na alta demanda da mesma atualmente e na quantidade de bibliotecas que auxiliam nos estudos de *Machine Learning*.

Para o segundo momento, deve-se desenvolver uma aprendizagem em relação a imagens de mapas diurnos e noturnos. Primeiro, extrai-se a intensidade luminosa do mapas diurnos para, em seguida, identificar a correlacionar essa aprendizagem às imagens de mapas diurnos.

Serão utilizados os seguintes bancos de dados imagéticos: Mapas de satélites diurnos (Google Static Maps API) e de satélites noturnos (National Geophysical Data Center, Version 4 DMSP-OLS Nighttime Lights Time Series).

Em um terceiro momento, será necessário correlacionar as atribuições possíveis com os mapas aos dados socioeconômicos já obtidos pelos questionários. Para a análise desses dados, a pesquisa utilizará dados de formulários brasileiros de pesquisa de consumo fornecidos por instituições que fornecem dados abertos, como o próprio IBGE. Por fim, o produto final do projeto

Por fim, será feito uma avaliação final dos dados obtidos. Será avaliado se a classificação pelas imagens de fato condizem com os dados brutos obtidos pelos formulários.

4 Cronograma

O objetivo desta sessão é formalização do plano de distribuição das etapas do projeto. O projeto tem a previsão de um ano de realização. A Tabela 1 apresenta esse cronograma.

ID	Tarefas	Meses											
		1	2	3	4	5	6	7	8	9	10	11	12
1	Estudo sobre Machine Learning: CNN												
2	Estudo sobre Transfer Learning												
3	Estudo sobre utilização de LSMS												
4	Construção de modelo de CNN												
5	Implementação do primeiro treinamento de imagens												
6	Aperfeiçoamento do modelo de CNN												
7	Aplicação do modelo CNN em imagens de satélites												
8	Aperfeiçoamento do modelo de CNN												
9	Construção do modelo de obtenção de dados da Avaliação de Padrões de Vida (LSMS)												
10	Aplicação do modelo CNN nos dados da Avaliação de Padrões de Vida												
11	Validação dos resultados												
12	Escrita de relatório parcial												
13	Escrita de artigos científicos e relatório final												

Figura 1: Cronograma de atividades da presente proposta

Referências

[1] G. Alejandro, "Medição da pobreza: o que tem na linha?" *Centro de Pobreza internacional*, 2004.

[2] R. Barros, R. Henriques, and R. Mendonça, "A estabilidade inaceitável: Desigualdade e pobreza no brasil," *Instituição de Pesquisa Econômica Aplicada*, p. 29, 2001.

[3] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, 2015.

[4] J. Henderson, A. Storeygard, and D. Weil, "Measuring economic growth from outer space," *American Economic Review*, 2012.

[5] "Onu: Países chegam a acordo sobre nova agenda de desenvolvimento pós-2015," 2015, [Online; Acessado 20 de Julho 2018].

[6] "Global extreme poverty," 2017, [Online; Acessado 10 de Julho 2018]. [Online]. Available: https://ourworldindata.org/extreme-poverty

[7] R. Lazarotto, "Disitribuição de renda no brasil - uma análise pós-plano real," p. 83, 2009.

[8] "Operação censitária," 2018, [Online; Acessado 21 de Julho 2018]. [Online]. Available: https://censo2010.ibge.gov.br/materiais/guia-do-censo/operacao-censitaria.html

[9] J. Henderson, A. Storeygar, and D. Weil, "Measuring economic growth from outer space," *American Economic Review*, 2012.

[10] L. Torrey and J. Shavlik, "Transfer learning," p. 22, 2009.