SmartSummy um sistema inteligente para resumos em português de vídeos na língua inglesa

SmartSummy a smart system for portuguese summarizations from english videos

Keomas Monteiro *[1]

[1] *Universidade Federal de Sergipe, DCOMP, São Cristovão, SE, Brasil.*

Resumo

A Inteligência Artificial e Aprendizado de Máquina têm avançado rapidamente e possibilitado a criação de soluções inovadoras. Uma delas, SmartSummy, é proposta nesse trabalho que utiliza modelos pré-treinados, Whisper e LED para gerar resumos em português de vídeos em inglês. Com essa ferramenta, é possível economizar tempo e facilitar o acesso a informações importantes, tornando o processo de aprendizado mais eficiente.

Palavras-chave: Aprendizado de Máquina. Sumarização. Combinação de modelos.

Abstract

Artificial Intelligence and Machine Learning have advanced rapidly and enabled the creation of innovative solutions. One of them is SmartSummy, presented in this article which uses pre-trained models, Whisper and LED to generate summaries in Portuguese from videos in English. With this tool, you can save time and facilitate access to important information, making the learning process more efficient.

Keywords: Machine Learning. Summarization. Ensemble Models.

1 Introdução

O uso de Inteligencia Artificial (IA) vem crescendo rapidamente nos últimos anos, impulsionado pela disponibilidade crescente de dados, avanços na tecnologia e demanda por automação e eficiência. De acordo com um relatório da consultoria McKinsey (CAFFERATA, 2023), em 2020 cerca de 50% dos líderes empresariais entrevistados afirmaram ter implementado pelo menos uma solução de IA em suas empresas. Grande parte dessa soluções são decorrentes de um ramo da IA denominado Aprendizagem de Máquina (ou Machine Learning).

A Aprendizagem de Máquina, permite que sistemas aprendam e se aprimorem continuamente, sem a necessidade de serem explicitamente programados. Dentro desse campo, a Aprendizagem Profunda (ou Deep Learning) tem sido um dos tópicos mais estudados, utilizando redes neurais artificiais profundas para processar e analisar grandes quantidades de dados, permitindo que esses modelos aprendam a representar os dados de forma hierárquica e identificar padrões complexos. Essa abordagem tem sido amplamente aplicada em áreas como reconhecimento de imagem, processamento de fala, tradução automática e muitas outras (LECUN; BENGIO; HINTON, 2015).

Entre os principais modelos de Aprendizagem Profunda, podemos citar as Redes Neurais Convolucionais (CNNs), que são utilizadas em visão computacional para reconhecimento de imagens e vídeos; as Redes Neurais Recorrentes (RNNs), que são empregadas em processamento de linguagem natural e análise de séries temporais; e as Redes Generativas Adversariais (GANs), que são utilizadas em tarefas de geração de conteúdo. Esses modelos inteligentes podem ser combinados e oferecer soluções para as mais variadas necessidades, com o objetivo de trazer eficiência e agilidade ao ser humano.

O presente trabalho apresenta uma combinação de modelos de aprendizagem profunda para executar a tarefa de criar textos de resumos em portugues de vídeos na língua inglesa. Tal solução

*Email: keomas@academico.ufs.br

Textolivre
Linguagem e Tecnologia

DOI: 0009-0007-9656-9478

Seção:
Artigos

Autor Correspondente:
Keomas Monteiro

Recebido em:
3 de maio de 2023
Aceito em:
3 de maio de 2020
Publicado em:
3 de maio de 2023

pode ser útil em áreas como jornalismo, pesquisa científica e relatórios de negócios, onde há muitas informações para serem processadas rapidamente. A solução proposta por este trabalho foi batizada e SmartSummy.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta um referencial teórico sobre os modelos utilizados pelo SmartSummy, a Seção 3 apresenta como o SmartSummy foi desenvolvido, a Seção 4 apresenta os resultados e a Seção 5 a conclusão.

2 Modelos Utilizados

O SmartSummy é um sistema de inteligência artificial que recebe um vídeo em inglês e gera um resumo em texto em português. Para alcançar esse resultado, são utilizados dois modelos: o Whisper e o Longformer-Encoder-Decoder (LED).

O Whisper é um modelo de processamento de voz que utiliza redes neurais convolucionais profundas para melhorar o reconhecimento de fala em ambientes com ruído (RADFORD et al., 2022). Ele é capaz de extrair informações relevantes da fala e filtrar o ruído ambiente, tornando o reconhecimento de voz mais preciso e eficiente. Esse modelo foi escolhido para lidar com a entrada de áudio do vídeo recebido pelo SmartSummy.

Nesse sentido, tem-se ainda o Longformer-Encoder-Decoder (LED) é um modelo de aprendizado profundo que foi projetado especificamente para a tarefa de resumir textos. Ele utiliza uma arquitetura de codificador-decodificador para gerar resumos precisos e concisos a partir de textos longos e complexos. O LED tem sido um modelo bastante eficiente para essa tarefa, superando outros modelos de resumo de texto, como o BART e o T5 em diversas métricas de desempenho (BELTAGY; PETERS; COHAN, 2020). Existem várias abordagens para a construção de modelos de resumo de texto, os modelos geralmente começam segmentando o texto em frases ou parágrafos e, em seguida, classificando essas unidades de acordo com sua importância para o conteúdo geral do texto. Uma das técnicas mais comuns para classificar a importância de frases é chamada de "sumarização extrativa". Nessa abordagem, o modelo examina cada frase do texto e atribui uma pontuação com base em fatores como a frequência de palavras-chave, a presença de informações importantes e a relevância para o contexto geral do texto. As frases com as pontuações mais altas são então selecionadas para compor o resumo final. Outra técnica comum é a "sumarização abstrativa", que é capaz de criar um resumo usando palavras e frases que não estão necessariamente presentes no texto original. Isso é feito treinando o modelo para reconhecer padrões na estrutura do texto e, em seguida, gerar um resumo usando linguagem natural que capture o significado geral do texto original. Dentre as referências bibliográficas importantes sobre o assunto, podemos citar (NENKOVA; MCKEOWN, 2012)), que apresenta uma revisão das técnicas de sumarização de texto e suas aplicações. Além disso, o artigo (ALLAHYARI et al., 2017) também apresenta uma revisão das técnicas de sumarização de texto, com foco na sumarização automática de notícias.

A combinação do Whisper e do LED no sistema SmartSummy permite que ele seja capaz de receber um vídeo em inglês, transcrevê-lo automaticamente usando o Whisper, gerar um resumo do texto usando o LED e, em seguida, traduzi-lo para o português.

3 SmartSummy

A solução SmartSummy foi desenvolvida em Python, utilizando modelos pré-treinados para processamento de áudio e texto. O processo é dividido em quatro fases. Na primeira fase, o áudio do vídeo é extraído utilizando a biblioteca MoviePy. Em seguida, o áudio é processado pelo Whisper, O texto gerado pelo Whisper é então repassado para o LED (Longformer-Encoder-Decoder). Por fim, o texto resumido é traduzido para o português utilizando a biblioteca GoogleTrans

O modelo Whisper treinado em dados de reconhecimento de voz está disponível no repositório: https://github.com/openai/whisper. O modelo LED foi treinado com o dataset BookSum (KRYŚCIŃSKI et al., 2021) que garante uma boa performance para textos longos. O LED treinado está disponível através da biblioteca Pytorch em https://huggingface.co/pszemraj/led-base-book-summary. A implementação do SmartSummy pode ser acessado em: repositório.

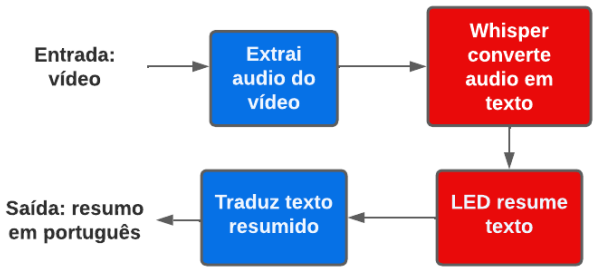

Figura 1. Fluxo funcionamento SmartSummy
Fonte: De autoria própria

4 Resultados

O SmartSummy é uma solução que permite extrair um resumo do conteúdo de um vídeo, auxiliando na obtenção de informações importantes de maneira rápida e eficiente. Ao realizar testes com alguns vídeos, foi possível verificar que o processo é bastante eficiente e produz resultados precisos em um tempo razoável. A exemplo, ao utilizar o SmartSummy para resumir o vídeo da palestra diposnivel na plataforma TED, "How Modern Audiences Can Talk about Aging Art | Margaret Hall | TED" foi possível obter resultados satisfatórios. Com duração de 10 minutos. Utilizando um computador com processador Intel Core i7-10610U e 16Gb de ram utilizando Windows, o SmartSummy levou 7 minutos e 17 segundos para gerar o resumo do vídeo, que contou com 184 palavras e resumiu bem as ideias centrais da palestra. O texto gerado apresentou a especialista em história do teatro musical "Midsouth" que busca mudar a maneira como as pessoas pensam na arte na era da "arte envelhecida". Informações sobre os resultados do SmartSummy podem ser vistos em A implementação do SmartSummy pode ser acessado em: repositório.

5 Conclusão

Em conclusão, o SmartSummy é uma solução muito promissora para gerar resumos de vídeos em inglês, tornando o processo de obtenção de informações mais rápido e eficiente. Utilizando modelos pré-treinados como o Whisper e o LED, o SmartSummy foi capaz de produzir resumos de alta qualidade para vídeos em inglês, com resultados satisfatórios em testes realizados com dois vídeos diferentes. Uma possível evolução para o SmartSummy seria disponibilizá-lo como um serviço online, permitindo que um número maior de pessoas possa se beneficiar da solução. Além disso, seria interessante explorar a possibilidade de treinar o LED com outros datasets, a fim de aumentar sua precisão e tornar o resumo ainda mais efetivo. Seu potencial é imenso, e a solução tem a possibilidade de ajudar estudantes, pesquisadores e profissionais em diversas áreas, tornando o processo de pesquisa e estudo mais ágil e eficiente.

Referências

ALLAHYARI, Mehdi et al. *Text Summarization Techniques: A Brief Survey*. [S.l.: s.n.], 2017. arXiv: 1707.02268 [cs.CL].

BELTAGY, Iz; PETERS, Matthew E.; COHAN, Arman. Longformer: The Long-Document Transformer. *arXiv:2004.05150*, 2020.

CAFFERATA, Pepe. *ChatGPT, a inteligência artificial como você nunca viu, é a próxima revolução | McKinsey*. Edição: Donald Neumann e Georgina Jabbour. [S.l.: s.n.], fev. 2023. Disponível em: https://www.mckinsey.com.br/our-insights/all-insights/chatgpt-e-a-revolucao-da-inteligencia-artificial. Acesso em: 2 mai. 2023.

KRYŚCIŃSKI, Wojciech et al. BookSum: A Collection of Datasets for Long-form Narrative Summarization. *arXiv:2105.08209 [cs]*, mai. 2021. Disponível em: https://arxiv.org/abs/2105.08209.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015.

NENKOVA, Ani; MCKEOWN, Kathleen. A Survey of Text Summarization Techniques. In: *Mining Text Data*. Edição: Charu C. Aggarwal e ChengXiang Zhai. Boston, MA: Springer US, 2012. p. 43–76. ISBN 978-1-4614-3223-4. DOI: 10.1007/978-1-4614-3223-4_3. Disponível em: https://doi.org/10.1007/978-1-4614-3223-4_3.

RADFORD, Alec et al. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv:2212.04356 [cs, eess]*, dez. 2022. Disponível em: https://arxiv.org/abs/2212.04356.