CRIAÇÃO DE DATASET PARA TREINAMENTO DE REDE NEURAL COM FOCO EM MELHORIA DE QUALIDADE VISUAL DE VÍDEOS COMPRIMIDOS

Murillo Aleixo Mota[1]; Gilberto Kreisler Franco Neto[2]; Guilherme Ribeiro Corrêa[3]

[1]*Universidade Federal de Pelotas – mamota@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – gkfneto@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – gcorrea@inf.ufpel.edu.br*

1. INTRODUÇÃO

Nos dias atuais, temos presenciado um grande avanço tecnológico que impacta principalmente a quantidade de conteúdo de vídeo disponível online. Além disso, vídeos com resoluções cada vez maiores estão se tornando comuns, como o Ultra High Definition (UHD), que precisa de novas tecnologias para que possa haver maior compressão, possibilitando transmitir e armazenar os dados de vídeos.

A compressão de vídeo permite armazenar e transmitir conteúdo de alta qualidade de forma eficiente, reduzindo o tamanho dos arquivos ao eliminar informações redundantes. No entanto, essa compressão pode levar à perda de qualidade proporcional à quantidade de dados removidos, gerando artefatos de compressão.

Com o objetivo de melhorar a qualidade visual de vídeos descomprimidos, o uso de redes neurais profundas (*Deep Neural Networks* — DNN) acabou se tornando uma boa opção para a realização dessa tarefa (BIRMAN; SEGAL; HADAR, 2020). O modelo de DNN envolve o componente *Convolutional Neural Networks* (CNN), que reconhece padrões espaciais em imagens usando filtros convolucionais (MISHRA; GUPTA; DUTTA, 2020). As redes de aprendizado profundo utilizadas para melhorar a qualidade de vídeos normalmente são treinadas utilizando vídeos comprimidos e vídeos originais, sem levar em consideração as informações sobre as perdas de qualidade específicas de cada codec. Então, como diferentes codecs e diferentes ferramentas de codificação levam a diferentes tipos de perda de qualidade, as informações extraídas ao longo do processo de codificação podem ser úteis para auxiliar no treinamento de modelos de melhoria de qualidade.

Com o uso de redes neurais para aprimoramento da qualidade de vídeo, a arquitetura *Spatio-Temporal Deformable Fusion* (STDF), desenvolvida por Deng et al. (2020) é capaz de se adaptar a vários movimentos, posições e tipos de artefatos de compressão presentes no quadro, devolvendo uma imagem com melhor qualidade visual e removendo a ocorrência de artefatos de compressão. Portanto, o STDF possibilita a fusão de informações dos quadros de referência de forma mais precisa e efetiva do que outros métodos anteriores, como o *Multi-Frame Quality Enhancement* (MFQE) proposto por Yang et al. (2018).

Para aprimorar o treinamento de redes neurais para melhoria de qualidade de vídeo, espera-se que a utilização de informações oriundas do processo de decodificação do vídeo possa auxiliar na reconstrução da imagem de baixa qualidade. Assim, este trabalho tem como objetivo extrair informações que possam ser úteis no processo de treinamento da rede neural, como o nível de quantização empregado em cada bloco, os modos de predição utilizados pelo codificador de origem, a quantidade de resíduo gerado pelo processo de predição

e os tipos de transformadas empregadas. Tais informações não foram utilizadas para o treinamento da arquitetura STDF, que baseia-se apenas em aprendizado a partir de pares compostos pela imagem original e pela imagem comprimida. Assim, este trabalho tem por objetivo extrair tais informações para posterior uso na definição e treinamento de novos modelos baseados na arquitetura STDF que potencialmente aumentem os níveis de melhoria de qualidade.

2. METODOLOGIA

Primeiramente, foi realizado um estudo da literatura em busca de artigos relevantes e relacionados com o trabalho proposto. Esses artigos foram necessários para atingir uma compreensão melhor sobre o assunto, possibilitando uma boa base para a realização do trabalho. Os artigos abordaram temas a respeito de aprendizado de máquina, redes neurais profundas, redes neurais convolucionais, codificação de vídeo e melhoria de qualidade de vídeo.

Após, foi realizada a análise das ferramentas a partir das quais poderão ser extraídas as informações para o treinamento da rede neural. Exemplos de possíveis informações a serem extraídas das ferramentas incluem a divisão de cada bloco do quadro (blocos de codificação) em sub-blocos, as informações de parâmetros de quantização empregados em cada bloco, modos de predição intra-quadro utilizados em cada bloco, vetores de movimento obtidos a partir da etapa de estimação de movimento, valores de resíduos de predição gerados pelo codificador, tipos de transformadas empregadas, entre outros.

A extração de tais informações pode auxiliar no treinamento de modelos preditivos que possam compreender artefatos de compressão gerados por diferentes padrões de codificação de vídeo utilizados. Desse modo, espera-se que o modelo possa aprender e associar as perdas de qualidade com informações como subdivisões dos blocos, tipos de predição empregados, entre outros. Com o uso dessas informações, pretende-se aumentar a diversidade dos dados de treinamento, o que pode auxiliar o treinamento da rede neural para a melhoria de qualidade de vídeo. O *codec* que foi avaliado e utilizado para esta extração de dados foi o HEVC, pois este codificador foi usado para gerar os dados de treinamento do modelo STDF original (DENG et al., 2020), a partir do qual as informações serão extraídas por meio da execução do seu decodificador.

Em seguida, foi realizada a modificação no código-fonte de referência do *codec* HEVC, nas funções do decodificador, para extração dos dados e escrita em arquivos externos durante o processo de decodificação de vídeos. Foram utilizados 126 vídeos não comprimidos (108 para treinamento e 18 para testes) em diferentes resoluções disponíveis no dataset MFQE. Todos os vídeos foram codificados com parâmetro de quantização (QP) em valor 37, que é o mesmo utilizado pelos autores do modelo STDF original (DENG et al., 2020). Após realizada a codificação, os vídeos foram codificados pelo decodificador com o código-fonte modificado com as rotinas de extração de dados. Na primeira implementação, reportada neste resumo, os dados extraídos referem-se exclusivamente às decisões de particionamentos dos blocos em sub-blocos menores, realizadas pelo codificador HEVC e utilizadas pelo decodificador para reconstruir cada imagem.

Para visualização dos dados extraídos, foi desenvolvido um algoritmo em Python, utilizando a biblioteca OpenCV, para criar um dataset de imagens. Essas imagens representam a subdivisão dos blocos, onde o menor bloco de tamanho 8x8 (pixels) é representado pela cor branca, o próximo tamanho de bloco (16x16),

determinado por um cinza claro, o tamanho de 32x32 representado pelo cinza escuro e, por último, o bloco 64x64, representado pela cor preta.

Além de permitir a visualização fácil das informações extraídas, esse conjunto de dados auxiliará, no futuro, o processo de treinamento de modelos com base na arquitetura STDF para melhoria de qualidade visual de múltiplos padrões de codificação sob diferentes configurações de *codec*. Por fim, será avaliada a compatibilidade e a utilidade das informações extraídas para o treinamento e teste da rede neural baseada na arquitetura STDF.

3. RESULTADOS E DISCUSSÃO

Até o momento, os resultados obtidos incluem a criação de um conjunto de dados de imagens, conforme mencionado anteriormente. O resultado da extração de informações do decodificador foi um arquivo de texto para cada vídeo. Esses arquivos contêm as coordenadas de cada bloco, bem como a profundidade utilizada para a subdivisão de cada bloco no quadro. Em seguida, esses resultados foram usados como entrada para o algoritmo desenvolvido em Python, permitindo a montagem das imagens a partir dessas informações. Isso possibilitou a reconstrução das imagens de cada quadro, identificando a profundidade utilizada na decodificação dos blocos.

A Figura 1 e a Figura 2 apresentam dois exemplos de imagens (à esquerda) geradas a partir das informações extraídas no processo de decodificação do primeiro e do segundo quadro do vídeo *Akiyo* (à direita). Ao analisar a Figura 1, é possível perceber que há uma grande quantidade de subdivisões de blocos. Isso ocorre pois este é o primeiro quadro do vídeo, o que implica que a predição utilizada na codificação (e, consequentemente, na decodificação) foi a intra-quadro, onde apenas informações referentes ao quadro atual são utilizadas.

Ao analisar a Figura 2, é possível notar que não houve muitas subdivisões como na primeira figura. Isso ocorre porque o segundo quadro do vídeo utiliza o quadro anterior como referência através da predição inter-quadros, e como a imagem não teve mudanças significativas, o nível de subdivisão se mantém quase o mesmo em todo o quadro.

Os passos seguintes, listados na metodologia, serão realizados assim que outros tipos de informações sejam extraídas do decodificador. Após a extração de todas as informações, um processo semelhante ao de geração de imagens de subdivisão de blocos será realizado, criando várias camadas de imagens que poderão alimentar o treinamento da rede neural juntamente com cada par de quadro original e comprimido.

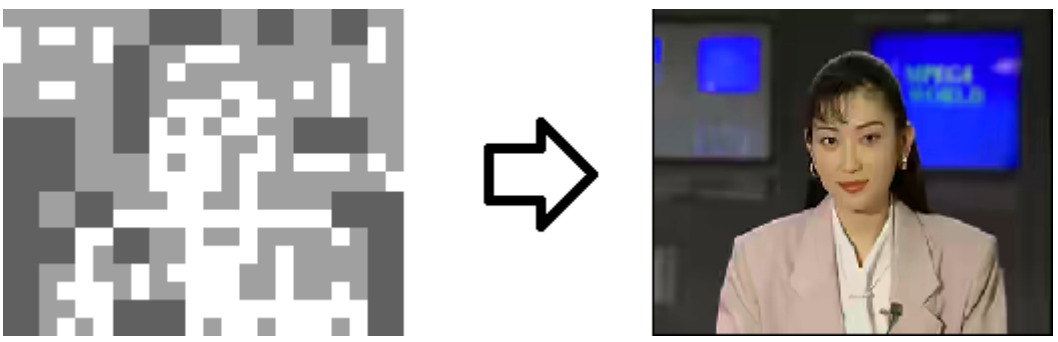

Figura 1 - Demonstração da reconstrução de imagem do primeiro quadro do vídeo Akiyo. Em branco, temos blocos de tamanho 8x8; em cinza claro, blocos de tamanho 16x16; em cinza escuro, blocos de tamanho 32x32. Fonte: Autor.

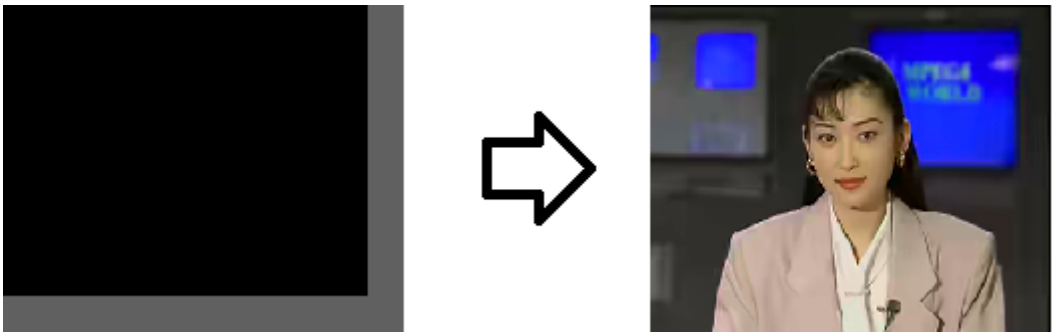

Figura 2 - Demonstração da reconstrução de imagem do segundo quadro do vídeo Akiyo. Em cinza escuro, blocos de tamanho 32x32; em preto, blocos de tamanho 64x64. Fonte: Autor.

4. CONCLUSÕES

Este trabalho tem como objetivo a extração de dados do codec HEVC para a criação de um conjunto de dados para o treinamento de uma rede neural baseada na arquitetura STDF, com foco na melhoria da qualidade visual de vídeos comprimidos. Até o momento, pode-se observar que, nas imagens geradas, o primeiro quadro apresenta um alto nível de subdivisão de blocos, pois utiliza a predição intra-quadro. Já o segundo quadro utiliza a predição inter-quadros, dependendo de quadros anteriores como referência. Além da subdivisão de blocos, os próximos passos incluem a extração de mais informações do decodificador para auxiliar na formação do conjunto de dados de treinamento para a rede neural, incluindo tipo de predição utilizado, níveis de quantização, tipos de transformadas, vetores de movimento, entre outras. Espera-se que o uso de todas essas informações sejam úteis ao serem empregadas como novas camadas de imagens, que poderão alimentar o treinamento da rede neural.

5. REFERÊNCIAS BIBLIOGRÁFICAS

BIRMAN, R.; SEGAL, Y.; HADAR, O. Overview of Research in the field of Video Compression using Deep Neural Networks.Multimedia Tools and Applications, [S.l.],v.79, p.11699–11722, 2020.

.

MISHRA, R.; GUPTA, H. P.; DUTTA, T. Overview of Research in the field of Video Compression using Deep Neural Networks.arXiv preprint arXiv:2010.03954, [S.l.],2020.

DENG, J.; WANG, L.; PU, S.; ZHUO, C. Spatio-temporal deformable convolution for
compressed video quality enhancement. In: AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2020. Proceedings. . . [S.l.: s.n.], 2020. v.34, n.07, p.10696–10703.

YANG, R.; XU, M.; WANG, Z.; LI, T. Multi-frame quality enhancement for compressed video. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION, 2018. Proceedings. . . [S.l.: s.n.], 2018. p.6664–6673.