Gustavo Lacerda – Agosto 2021

MOURA, Camila Stéffane Fernandes Teixeira de et al. Detecção de Deepfakes a partir de técnicas de visão computacional e aprendizado de máquina. 2021.

O texto introduz que a informação em vídeo não é algo mais incontestável. Com o avanço da internet e da descentralização da informação passamos de espectadores para provedores de notícias e isso gerou um aumento substancial nas fake news. Essas notícias falsas têm impacto em várias áreas diferentes e com o passar do tempo se tornou difícil verificar a fonte para confiar na veracidade de certas informações, por causa da velocidade e alcance que as notícias falsas possuem.

O trabalho comenta que focará nas deepfakes, que é uma tecnologia que utiliza aprendizagem de máquina para criar imagens e vídeos falsos. Essa tecnologia surgiu em um fórum do reddit em 2017 e ganhou popularidade por ser utilizada na geração de vídeos de figuras públicas e políticos. Assim, os deepfakes avançaram em qualidade e técnicas atuando em três principais áreas: política, entretenimento e conteúdo adulto.

Parte da popularização das deepfakes também se deve ao fato da tecnologia estar cada vez mais acessível e a melhoria e otimização das técnicas utilizando GANs. Essa popularização em conjunto ao impacto que as deepfakes possuem aumentaram a demanda por métodos de detectá-las. Como essa tecnologia ainda é bastante recente, tanto ela quanto as técnicas de detecção têm evoluído de forma exponencial chamando atenção de vários pesquisadores e empresas que estão preocupadas com os problemas geradas por esses conteúdos falsos. Um exemplo de empresa que está ativamente atuando nessa área é o facebook, que criou o DeepFake Detection Challenge (DFDC), para incentivar a produção de técnicas para detecção de deepfakes em 2019.

O autor fala que, na busca de uma solução para o problema de classificar conteúdo falso e real, foram abordadas as seguintes técnicas: procurar características físicas, detecção de artefatos e como os dados são apresentados para a rede. O trabalho procura, portanto, desenvolver uma técnica nova mais generalista que seja eficaz em reconhecer deepfakes em cenários conhecidos e desconhecidos através de artefatos de imagem. Foi, então, coletado informações sobre as principais técnicas utilizadas para a geração de deepfakes, além da busca por conjuntos de dados específicos continuamente, para sempre acompanhar a evolução da técnica. Nesses dados foram aplicados aumentações a partir da mudança de características das imagens, fornecendo mais variedade sem necessariamente ter dados completamente novos. Ademais, foi decidido que o novo método detectará deepfakes feitas com rostos e foram utilizadas algumas regiões chaves da face: olhos, nariz e boca a fim de auxiliar os resultados obtidos.

A manipulação de faces é cada vez mais comum e acessível e envolve a modificação de atributos faciais, troca ou transformações de duas faces, geração de rostos sintéticos ou re-encenação de expressões faciais em imagens ou vídeos. Para isso é utilizado Redes Neurais Profundas (*Deep Neural Networks*, DDNs). Dentre essas manipulações, o FaceSwap é uma categoria famosa de trocas de rostos. Nesta abordagem, se transfere uma região de um rosto para outro.

Ademais, o autor fala sobre as deepfakes, no qual esse termo está relacionado à substituição de faces em imagens ou vídeos que surgiu em 2017 em um fórum do reddit. Inicialmente esta técnica utilizava auto encoders para reconstruir imagens de treinamento do ator e do alvo. A técnica utilizava uma entrada com o ator no qual eram extraídas suas características faciais. Depois, é feito um alinhamento para ajustar as características, a mesclagem das características do ator no alvo, o realinhamento e por fim, a colagem para finalização da troca. Esse método foi aprimorado a partir da utilização de máscaras com o formato do rosto do ator e do alvo, facilitando assim, a troca. Com a evolução das Redes Adversárias Generativas (GANs), elas se tornaram a principal ferramenta utilizada na troca de faces.

O texto também fala sobre a re-encenação, no qual há uma técnica chamada Face2Face que permite a transferência de expressões faciais em tempo real.

Para criar o método de detecção de deepfakes, o autor primeiramente fala sobre a detecção facial, sendo essa uma parte importante pois as técnicas abordadas no trabalho serão focadas em faces de indivíduos. Além disso, o autor traz um breve briefing sobre redes neurais e redes neurais convulsionais que serão ferramentas utilizadas no método.

Como metodologia, o trabalho consiste em analisar bases existentes e técnicas convolucionais que sejam mais eficientes na classificação de imagens reais e falsas. No modelo proposto há um conjunto de faces que é dividido em conjuntos menores que podem ter que passar por tratamento prévio ou serão aplicados diretamente na rede neural convolucional. Em seguida a rede é usada para extrair características que serão ajustadas por camadas conectadas fornecendo o resultado para a classificação intra ou inter da base de dados.

Os pesquisadores concluem que, usando apenas o conjunto FullTIMIT os dados se mostraram bastante limitados para a pesquisa do novo método. Por essa razão eles construíram e rotularam seu próprio conjunto de dados. Este banco utiliza a base de dados VGG_Faces2 como classe real e vídeos coletados como classe de vídeos sintéticos/falsos.

A base de dados da pesquisa é de autoria dos pesquisadores e é denominada Sensitive. Ela foi montada a partir de um tratamento dos vídeos de várias bases de dados. Essas imagens foram padronizadas e houve a extração de regiões chaves do rosto: olhos, sobrancelhas, boca e nariz. Além disso houve a aumentação dos dados utilizando a base criada a parir de mudanças como a iluminação, angulação e rotação das imagens existentes. Assim foi criado mais dados com a base já existente. As etapas para a criação desta base foram as seguintes utilizando vídeos presentes em sites da internet.
Levantamento: Inicialmente, realizamos um levantamento de onde podem ser encontrados vídeos ou imagens com DeepFakes. Isso foi realizado através do monitoramento de fóruns que continham referência a esse tipo de conteúdo em redes sociais como o Twitter e o Reddit.
Coleta: Em seguida, realizamos a coleta desses vídeos para compor a nossa base.
Processamento: Para esta etapa, fizemos o processamento dos dados coletados a fim de verificar a qualidade dos vídeos. Nesse momento, removemos os vídeos que apresentaram características indesejadas, tais como, múltiplas faces.

Extração de imagens: Após o processamento, fizemos a extração de quadros dos vídeos selecionados, de modo a gerar um conjunto de imagens.
Composição do conjunto final: Compomos a nossa base final Sensitive associando as imagens reais e as imagens sintéticas/falsas coletadas por nós.
Extração de Faces: Para cada imagem da base de dados (real e sintética/falsa) utilizamos algoritmos de detecção facial para selecionar apenas as faces.
Rotulação: Por fim, a rotulação consistiu na separação das imagens em duas classes: real, contendo imagens do VGG_Faces2 e sintética/falsa.
Além disso foi utilizado os bancos Notre Dame Synthetic Face – NDSF, FullTIMIT, DeepFake Detection Challenge e DeepFake Detection.

Foi então, a partir dos dados obtidos, feito um certo tratamento, no qual era detectado em cada imagem a face e os patches (Partes chaves do rosto: olhos, sobrancelhas, nariz e boca). Esses dados tratados foram divididos entre treino e validação na proporção 80% para treino e 20% para validação.

Os testes foram aplicados nas diferentes bases de dados com diferentes arquiteturas. Os modelos de rede que mais se adequaram ao que os pesquisadores propunham foram o a InceptionResNetV2, Xception e a MobileNet. Depois o texto fala sobre encontrar o melhor otimizador e a melhor taxa de aprendizado, que pode impactar significativamente no desempenho da classificação. Dados os experimentos, se concluiu que o Nadam com taxa de aprendizagem 0. 00001 demonstrou os melhores resultados. Além disso foi escolhido a taxa de congelamento de 1.0.

Utilizando as bases de dados e as especificações apresentadas o trabalho conclui que os métodos utilizados foram razoáveis na classificação de deepfakes. Os testes realizados conseguiram demonstrar que é possível construir um modelo capaz de realizar a classificação correta de quase 100% das imagens.