RECONHECIMENTO ÓTICO DE CARACTERES PARA TEXTOS CIENTÍFICOS COM NOTAÇÃO MATEMÁTICA

Rafael Rodrigues dos Santos (PIC/Uem), Franklin César Flores (Orientador), e-mail: rafael11rodrigues@hotmail.com.

Universidade Estadual de Maringá / Centro de Tecnologia/Maringá, PR.

Ciências Exatas e da Terra. Ciência da Computação.

Palavras-chave: CNN, OCR matemático, Morfologia Matemática.

Resumo

O problema de reconhecimento ótico de caracteres aplicado a textos científicos com notação matemática é desafiador. Fórmulas matemáticas apresentam características estruturais e semânticas diferentes de um texto em linguagem natural, requerendo assim uma abordagem mais sofisticada. O objetivo deste trabalho é propor um método que resolve parte deste problema, segmentando os símbolos das equações por meio de processamento morfológico e classificando-os individualmente com uma rede neural convolucional. O método foi aplicado em 449 equações simples, sendo que a maior parte dos erros se deu por conta da segmentação, mostrando que a rede neural é um ótimo classificador para símbolos isolados.

Introdução

A Internet se tornou uma fonte inesgotável e imprescindível de informação. Em especial, no mundo acadêmico, onde busca-se cada vez mais disponibilizar textos científicos para que sejam acessados por dispositivos conectados à rede, inclusive materiais impressos mais antigos. No caso desta última categoria de documentos, sua disponibilização é em grande parte possibilitada pela aplicação de sistemas de reconhecimento óptico de caracteres - mais conhecidos pelo acrônimo OCR (*Optical Character Recognition*) – que são capazes de converter documentos escaneados em arquivos digitais – de imagem ou texto.

Embora sejam muito eficientes para a digitalização e interpretação de textos em linguagem natural – como, por exemplo, do nosso cotidiano, os sistemas OCR convencionais apresentam limitações severas ao trabalharem com textos científicos que fazem uso de notações matemáticas, por não reconhecerem adequadamente expressões e símbolos presentes frequentemente nesse tipo de texto (GARAIN, 2009).

Dentre os motivos que tornam o reconhecimento de expressões matemáticas uma tarefa difícil, mesmo considerando que cada símbolo tenha sido reconhecido individualmente, Chang e Yeung (CHAN E YEUNG, 2000) destacam os seguintes:

- Caracteres e símbolos podem ser dispostos em estruturas complexas de duas dimensões e podem apresentar tamanhos diferentes;

- Existem vários tipos de símbolos, cada um com um critério de agrupamento;
- Um mesmo símbolo pode apresentar diferentes significados em diferentes contextos.

Além disso, após o reconhecimento individual de cada símbolo (que também acontece em textos em linguagem natural), expressões matemáticas devem passar por uma análise estrutural para que a relação entre os símbolos (como subescritos, sobrescritos, frações e matrizes) possa ser identificada (MALON, UCHIDA e SUZUKI, 2008).

Essas questões vêm sendo investigadas há mais de quatro décadas e várias técnicas foram propostas ao longo dos anos, cada uma se concentrando em pontos específicos. Por exemplo, há trabalhos que se dedicam a separar expressões matemáticas do restante do texto, outros que buscam reconhecer símbolos individualmente em expressões já isoladas, outros que focam na análise estrutu-ral e outros que objetivam levantar dados estatísticos e categorizar os tipos de situações que podem ser encontradas ao digitalizar documentos científicos.

Os pontos aqui destacados refletem um pouco da complexidade do problema e ressalta a importância de se continuar buscando novos métodos ou combinações e melhorias de métodos já existentes a fim de se obter soluções cada vez mais eficazes para o reconhecimento de expressões matemáticas em documentos escaneados.

Materiais e métodos

Equações utilizadas

O protótipo proposto foi avaliado utilizando 449 equações da base de dados InftyMDB-I. As equações contêm apenas símbolos em uma única linha, podendo ser base, subscrito ou sobrescrito. As imagens usadas são binárias, sendo o fundo branco e os símbolos pretos.

Estratégia de segmentação

Para segmentar os símbolos de uma equação, a mesma foi percorrida da esquerda para a direita com um elemento estruturante equivalente a uma coluna da imagem e aplicando o operador morfológico de erosão. Assim, cada símbolo foi transformado em um retângulo, sendo que a primeira e a última coluna do retângulo limita o símbolo na horizontal. Após extrair cada símbolo, um raciocínio análogo foi usado para encontrar os limites verticais, usando um elemento estruturante no formato de uma linha.

Nível do símbolo na linha

Para determinar se um símbolo está no nível de subscrito, sobrescrito ou base, foram consideradas a distância da primeira linha da imagem até a primeira linha do símbolo, a distância da última linha da imagem até a última linha do símbolo e a altura do símbolo. Se a diferença entre as distâncias mencionadas for menor ou igual a 20% da altura da imagem ou se a altura do símbolo é 75% ou mais da altura

da imagem, então o símbolo é base. Caso contrário, se ele estiver mais perto do topo da imagem é sobrescrito, se estiver mais perto do fundo da imagem é subscrito.

Classificador

O modelo de classificação usado foi uma rede neural convolucional com 3 camadas de convolução, com kernels de tamanho 3 x 3 e cada uma com 32, 64 e 128 mapas de características, respectivamente. Cada camada é seguida por uma operação de *max-pooling* com uma janela 2 x 2, além de *dropouts* de 0,25. Tal estrutura é conectada a uma camada densa de 128 neurônios e estes são conectados à camada de saída com 265 neurônios que aplicam a função *softmax*. Todos os outros neurônios aplicam a função *Leaky ReLU*. O modelo foi treinado e avaliado com símbolos isolados da base InftyCDB-3 (PROJECT, 2006), com cerca de 260.000 símbolos, após serem redimensionados para 50 x 50 com a ferramenta ImageMagick (LLC, 2019). Foram usados 70% dos símbolos para treinamento e 30% para teste. Após treinar o modelo por 192 épocas, ele foi usado para classificar cada símbolo segmentado de cada equação.

Implementação

O protótipo foi implementado em Python utilizando as bibliotecas OpenCV e NumPy para manipular as imagens e o framework Tensorflow para implementar o classificador.

Resultados e Discussão

Ao segmentar as 449 equações, foram obtidos 7481 símbolos, entretanto esse número não corresponde à quantidade real de símbolos porque alguns deles foram aglutinados devido à forma de segmentação empregada. Por exemplo, “(p” muitas vezes foi segmentado como um único símbolo, o que também interfere na classificação (neste exemplo, “(p” foi classificado como “ϕ” ou “φ”).
A tabela 1 apresenta os tipos de erro e as respectivas quantidades encontradas.

Tabela 1 – Tipos de erros encontrados

Tipo de erro	Quantidade de símbolos errados
Nível	719
Segmentação	280
Classificação	207

A tabela mostra que a maior dificuldade encontrada foi em determinar o nível do símbolo. Isso ocorre devido à grande variedade de fontes e tamanhos dos símbolos, além de que equações com símbolos mais altos interferem nas posições das linhas limitantes dos símbolos com relação à altura da imagem.
Outro problema encontrado foi com relação à segmentação. Como mencionado anteriormente, vários símbolos foram aglutinados. Isso frequente ocorre porque a

ponta de um símbolo mais alto avança na direção do símbolo vizinho e, no processo de erosão empregado, ambos os símbolos se fundem em um único retângulo.
Por fim, vale ressaltar que boa parte dos símbolos classificados erroneamente são símbolos com detalhes que não estão presentes na base de treinamento da CNN (como barras verticais, “~” e “^”, por serem símbolos muito parecidos (como “0”, “O” e “o”) ou pela presença de ruídos.

Conclusões

Este trabalho buscou apresentar um protótipo de OCR matemático usando operadores morfológicos para segmentar os símbolos das equações e uma rede neural convolucional para classificar cada símbolo. Foi possível observar que a maior dificuldade está na segmentação correta dos símbolos e na determinação do nível em que cada símbolo se encontra, visto que apenas cerca de 3% dos símbolos segmentados corretamente foram classificados erroneamente pela CNN.
Embora as equações utilizadas sejam bastante simples, acredita-se a técnica empregada pode ser aprimorada para contemplar equações com estruturas mais complexas (como frações, raízes e matrizes). Para tal, é necessário investigar outras formas de segmentar os símbolos, além de aumentar o número de símbolos usados no treinamento do classificador.

Agradecimentos

Agradeço ao professor Franklin pelo incentivo e pela confiança depositada em mim e ao Programa de Educação Tutorial (MEC/SESu) pelo fomento à pesquisa, ensino e extensão.

Referências

CHAN, K.-F.; YEUNG, D.-Y. Mathematical expression recognition: a survey. International Journal on Document Analysis and Recognition, Springer, v. 3, n. 1, p. 3–15, 2000.

GARAIN, U. Identification of mathematical expressions in document images. In: IEEE. Document Analysis and Recognition, 2009. ICDAR’09. 10th International Conference on. [S.l.], 2009. p. 1340–1344.

LLC, I. S. ImageMagick. 2019. Disponível em: <https://imagemagick.org/index.php>. Acesso em 20 abr. 2019.

MALON, C.; UCHIDA, S.; SUZUKI, M. Mathematical symbol recognition with support vector machines. Pattern Recognition Letters, Elsevier, v. 29, n. 9, p. 1326–1332, 2008.

PROJECT, I. Infty Project - Databases. 2006. Disponível em: <http://www.inftyproject.org/en/database>. Acesso em 03 mar. 2019.