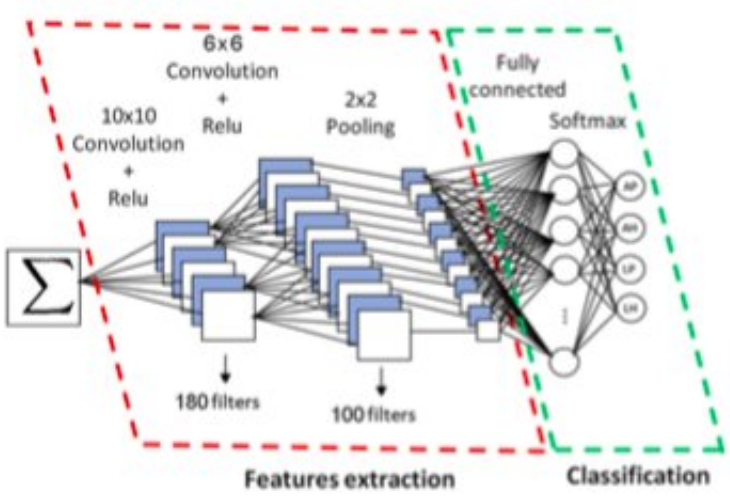
A CNN-based method for math formula script and type identification

March 21 2019, by Ingrid Fadelli



The CNN based system for symbol script and type identification. Credit: Khazri &Echi.

Researchers at the University of Tunis have recently proposed a new system for math formula script and type identification, which is based on convolutional neural networks (CNNs). Their method, presented in a paper published by Springer, can automatically discriminate between printed/handwritten and Arabic/Latin formulas.

In recent years, researchers have tried to develop systems that can identify the forms in which a document is presented, such as the language used and whether the text is machine-printed or handwritten, in

order to select the appropriate recognition system for each document. Most of these approaches focus on identifying different forms of text, while very few are designed to analyze mathematical formulas.

"In this context, we present a new approach dealing with the problem of identification of the script, Arabic or Latin; and the type, hand-written or machine-printed, of math formulas," the researchers at the University of Tunis wrote in their paper. "This work comes as a part of our research on offline recognition of Arabic math formulas."

In their study, the researchers presented a syntax-directed system designed to recognize symbols and analyze their arrangement. To recognize symbols, their approach uses statistical features and a Bayes network classifier.

To analyze the structure of a formula, their system employs a top-down and bottom-up parsing scheme based on operator dominance. In other words, their system carries out a lexical, geometrical and syntactical analysis of a formula, which helps it to identify its script (Latin vs. Arabic) and whether it was handwritten or machine-typed.

"Formula parsing consists in applying, from the dominant operator and its context, the appropriate rule to divide the formulas into sub-formulas, which will be recursively analyzed in the same way," the researchers explained in their paper.

Using a CNN, the approach devised by the researchers first extracts and then classifies connected components of a formula. The researchers trained and evaluated their system using Latin script formulas from the InftyMDB-1 and CROHME databases, as well as Arabic formulas scanned from math books or handwritten by five different writers.

"The proposed recognition system was tested on complex math formulas

containing implicit multiplication, subscripts and superscripts, with satisfactory results," the researchers wrote. "Adding more features, testing other feature selection algorithms and choosing faster classifiers should enhance the performance of the proposed system."

Overall, the evaluations carried out by the researchers yielded highly promising results, with their system achieving a 94.6 percent identification rate. The parser they used to analyze the structure of formulas also appears to be very robust, as it achieved an impressive recognition rate of 97.63 percent. In their future work, the researchers plan to improve the performance of their system by further developing the CNN's filters and architecture.

More information: Kawther Khazri et al. Math Formula Script and Type Identification and Recognition, *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications* (2019). DOI: 10.1007/978-3-030-13469-3_99