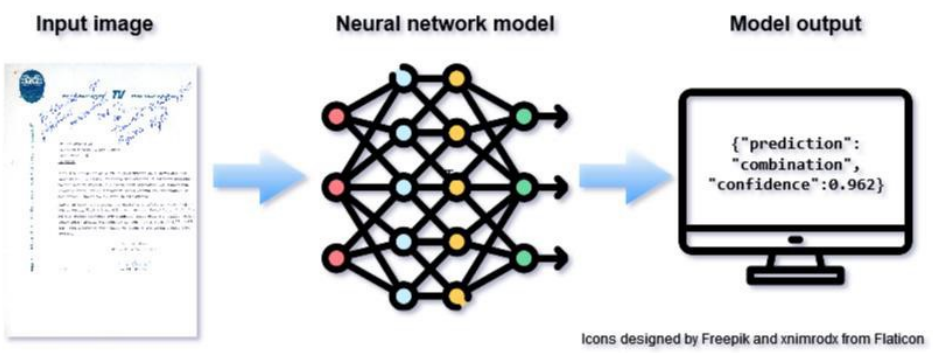
Description of the typeface recognition component

Background, benefits and purpose of use

Digitised documents may contain only typewritten or handwritten text, or both types of writing may be mixed to a varying degree. Common examples of the latter are typed documents that contain a handwritten signature, as well as various manually filled out forms. There may be a need to classify a document based on the various typefaces it contains e.g. when you want to locate those pages of the material that contain handwritten markings. The classification can also be used when you want to forward typed documents to the optical character recognition (OCR) process and handwritten documents to handwritten text recognition (HTR).

What does the component do?



The component processes the image files it receives as input one at a time and forms a prediction of the typefaces contained in the image based on the parameters it has learned during the training. If the component considers that the most likely option is that the image only contains typed text, the image is placed in the typed class, and the other two classes (handwritten, combined) are treated in the same way. The component returns the predicted class to the user.

How is the component used?

The component is available both as part of the Arkkiivi user interface (http://www.arkkiivi.fi/) and as a standalone application offering an application programming interface (API) on the DALAI project's GitHub page (https://github.com/DALAI-project/WritingtypeAPI). More detailed instructions on how to use the component can be found on the attached website.

Component training

Image files in .jpg format aligned to 224 x 224 pixels have been used for component training. Approximately 22,000 image files only containing handwritten texts, approximately 15,000 files only containing typed texts and approximately 19,000 combined files in terms of text type have been used in the training. Material from the late 18th century until the beginning of the 2000s is included in the training data.

Although efforts have been made to compile a variety of example cases for the training material, the amount of the material is limited and errors also occur in component classification. Identifying a combined class poses the most challenges for the component, and error classifications are possible especially if the handwritten text is poorly visible and/or the document only contains a little of it.