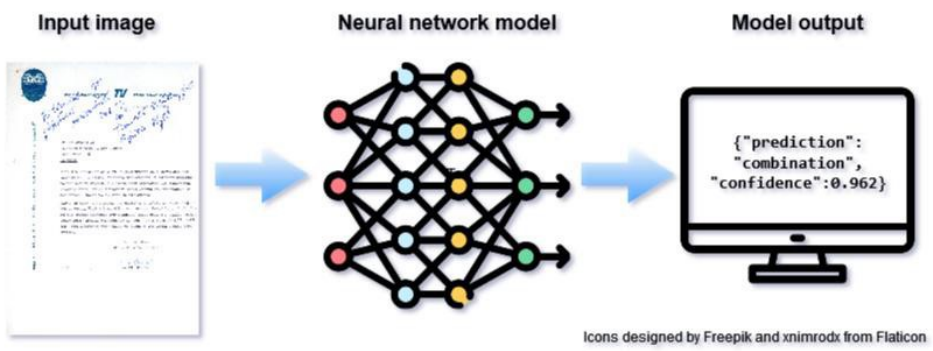
Igenkänning av skrivtyp

Bakgrund, fördelar och ändamål

Digitaliserade dokument kan innehålla enbart maskinskriven text eller handskriven text eller i varierande grad båda skrivtyperna. Vanliga exempel på sistnämnda är maskinskrivna dokument med handskrivna underskrifter och olika blanketter som har fyllts i för hand. Klassificering av ett dokument utifrån de olika skrivtyperna i dokumentet kan vara nödvändig, om man till exempel vill hitta de sidor i materialet som innehåller handskrivna anteckningar. Klassificeringen kan också användas, om man genom riktning vill dirigera maskinskrivna dokument till processen för igenkänning av maskinskriven text (optical character recognition, OCR) och handskrivna dokument i sin tur för igenkänning av handskriven text (handwritten text recognition, HTR).

Vad gör komponenten?



Komponenten behandlar de bildfiler som den fått som indata en efter en och tar fram en prognos för bildens texttyper på grundval av de parametrar som den har lärt sig under utbildningen. Om komponenten anser att det mest sannolika alternativet är att bilden endast innehåller maskinskriven text, placeras bilden i kategorin maskinskriven text. På motsvarande sätt görs med bilder i de två andra kategorierna (handskriven, blandad form). Komponenten återställer den förväntade klassen till användaren.

Hur används komponenten?

Komponenten finns tillgänglig både som en del av Arkkiivi-gränssnittet (http://www.arkkiivi.fi/) och som en fristående app som tillhandahåller ett programmeringsgränssnitt (API) på DALAI-projektets GitHub-webbplats (https://github.com/DALAI-project/WritingtypeAPI). Närmare anvisningar om användningen av komponenten finns på den webbplatsen.

Utbildning av komponenten

Bildfiler i .jpg-format har använts för att utbilda komponenten och deras storlek har harmoniserats till 224 x 224 pixlar. I utbildningen användes cirka 22 000 bildfiler som enbart innehöll handskriven text, cirka 15 000 filer som enbart innehöll maskinskriven text och cirka 19 000 filer som innehöll varierande texttyper. Tidsmässigt sett användes material från slutet av 1700-talet till början av 2000-talet. Trots att man har försökt samla många olika exempel till utbildningsmaterialet, är materialmängden begränsad och det uppstår också fel i komponentens klassificering. Igenkänningen av kategorin kombination innebär de största utmaningarna för komponenten, och felklassificeringar är möjliga, särskilt om handskriven text är svår att upptäcka och/eller det finns lite handskriven text i dokumentet.