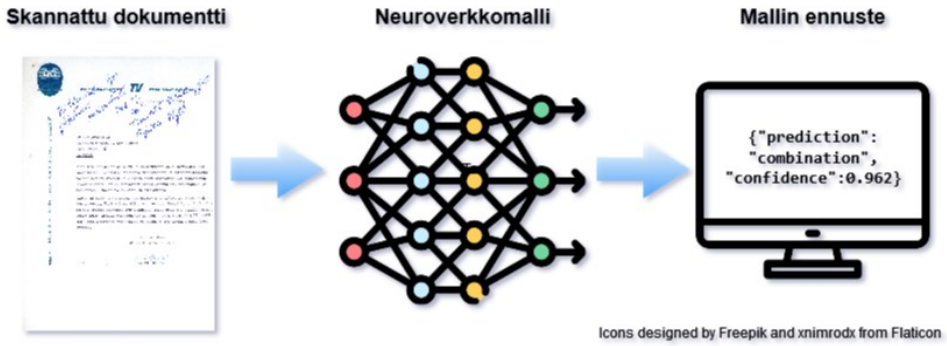
Kirjoitustyypin tunnistus -komponentin kuvaus

Tausta, hyöty ja käyttötarkoitus

Digitoidut asiakirjat voivat sisältää pelkästään kone- tai käsinkirjoitettua tekstiä, tai vaihtelevassa määrin molempia kirjoitustyyppejä sekaisin. Yleisiä esimerkkejä jälkimmäisistä ovat konekirjoitetut dokumentit, jotka sisältävät käsin kirjoitetun allekirjoituksen, sekä erilaiset käsin täytetyt lomakkeet.

Dokumentin luokittelulle sen sisältämien eri kirjoitustyyppien perusteella voi olla tarvetta esimerkiksi, kun halutaan paikantaa aineistosta ne sivut, jotka sisältävät käsin kirjoitettuja merkintöjä. Luokittelua voidaan myös hyödyntää silloin, kun halutaan kohdennetusti ohjata konekirjoitetut dokumentit konekirjoitetun tekstin tunnistusprosessiin (optical character recognition, OCR) ja käsin kirjoitetut dokumentit vastaavasti käsin kirjoitetun tekstin tunnistukseen (handwritten text recognition, HTR).

Mitä komponentti tekee?



Komponentti käsittelee syötteenä saamansa kuvatiedostot yksi kerrallaan, ja muodostaa koulutuksen aikana oppimiensa parametrien pohjalta ennustuksen kuvan sisältämistä tekstityypeistä. Mikäli komponentti pitää todennäköisimpänä vaihtoehtona sitä, että kuva sisältää vain konekirjoitettua tekstiä, sijoitetaan se konekirjoitettujen luokkaan, ja vastaavasti toimitaan kahden muun luokan (käsin kirjoitettu, sekamuoto) tapauksessa. Komponentti palauttaa käyttäjälle ennustetun luokan.

Miten komponenttia käytetään?

Komponentti on saatavilla sekä osana Arkkiivi-käyttöliittymää (http://www.arkkiivi.fi/) että itsenäisenä, ohjelmointirajapinnan (API) tarjoavana sovelluksena DALAI-hankkeen GitHub-sivulla (https://github.com/DALAI-project/WritingtypeAPI). Tarkemmat ohjeet komponentin käyttöön löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

Komponentin koulutukseen on käytetty .jpg-muotoisia kuvatiedostoja, joiden koko on yhdenmukaistettu 224 x 224 pikseliin. Pelkästään käsin kirjoitettuja tekstejä sisältäviä kuvatiedostoja on koulutuksessa käytetty n. 22 000 kappaletta, pelkästään konekirjoitettua tekstejä sisältäviä tiedostoja n. 15 000 kappaletta ja tekstityypin suhteen sekamuotoisia tiedostoja n. 19 000. Ajallisesti mukana on aineistoa 1700-luvun lopulta 2000-luvun alkuun asti.

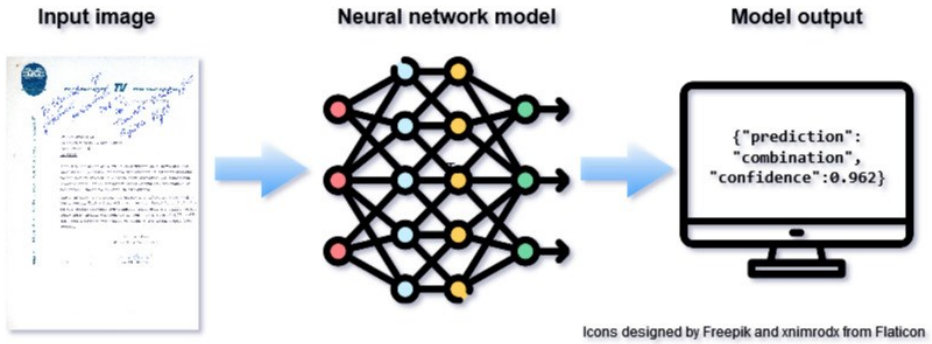
Vaikka koulutusaineistoon on pyritty kokoamaan monipuolisesti erilaisia esimerkkitapauksia, aineiston määrä on rajallinen ja komponentin luokittelussa tapahtuu myös virheitä. Yhdistelmäluokan tunnistaminen tuottaa komponentille eniten haasteita, ja virheluokitukset ovat mahdollisia erityisesti, jos käsin kirjoitettu teksti on heikosti havaittavaa ja/tai sitä esiintyy dokumentissa vähän.

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.

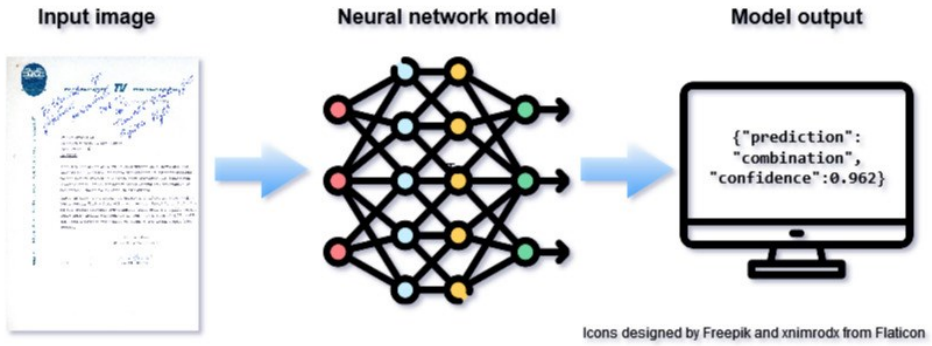
Although efforts have been made to compile a variety of example cases for the training material, the amount of the material is limited and errors also occur in component classification. Identifying a combined class poses the most challenges for the component, and error classifications are possible especially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

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.