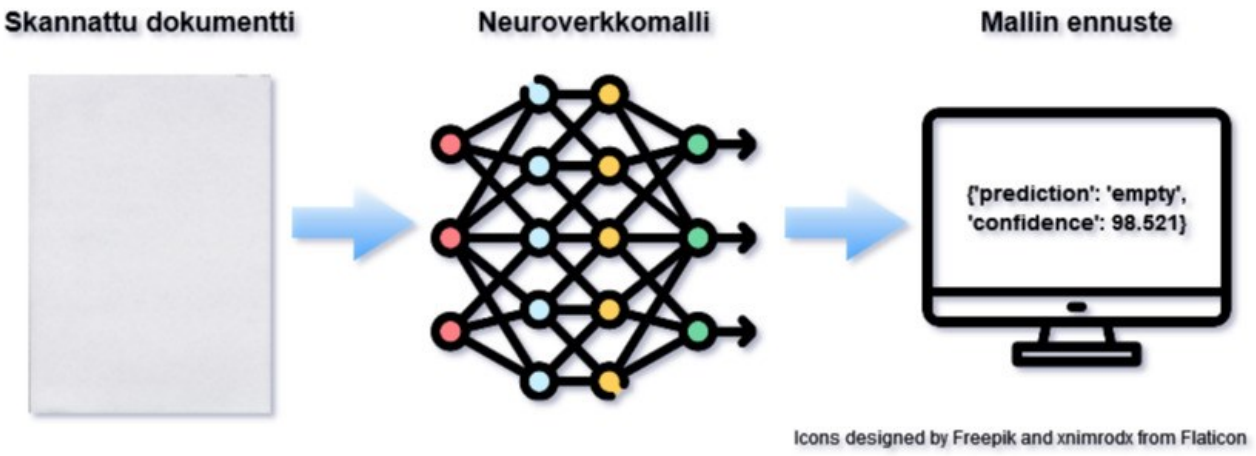
Tyhjien tunnistus -komponentin kuvaus

Tausta, hyöty ja käyttötarkoitus

Asiakirjojen digitointiprosessissa käytetään usein kaksipuoleista skannausta. Jos asiakirjat ovat yksipuoleisia, syntyy digitoinnissa paljon tyhjiä kuvia. Koneellinen luokittelu tyhjiin ja sisällöllisiin kuviin helpottaa digitoijan työtä, ja tunnistuksen laadun ollessa riittävän hyvällä tasolla se mahdollistaa myös tyhjien kuvien piilottamisen tai jopa poistamisen säilytyksestä.

Mikäli asiakirjat ovat asiakkaan katsottavissa palvelussa, voi mahdollisuus piilottaa tyhjät kuvat näkymästä auttaa parantamaan asiakaskokemusta. Mikäli tyhjät kuvat tunnistetaan ennen myöhempiä dokumentin prosessoinnin vaiheita (mm. virheiden tunnistus, tekstin tunnistus, metatietojen tunnistus), voidaan myös nopeuttaa dokumenttien prosessoinnin koko ketjua.

Mitä komponentti tekee?



Komponentti käsittelee syötteenä saamansa kuvatiedostot yksi kerrallaan, ja muodostaa koulutuksen aikana oppimiensa parametrien pohjalta ennustuksen kuvan sisällöstä. Mikäli komponentti tunnistaa kuvasta sisältöä yli 50 % todennäköisyydellä, kuva luokitellaan sisällölliseksi, muutoin tyhjäksi. 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/EmptyAPI). Tarkemmat ohjeet komponentin käyttöön löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

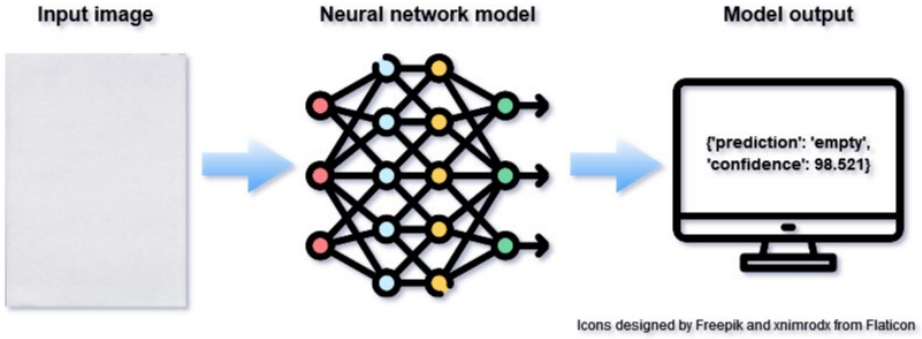
Komponentin koulutukseen on käytetty .jpg-muotoisia kuvatiedostoja, joiden koko on yhdenmukaistettu 224 x 224 pikseliin. Komponenttia on koulutettu useaan kertaan, eikä alkuperäisen koulutusaineiston määrästä ole tarkkaa tietoa. Lisäkoulutuksissa käytetty aineisto sisältää noin 100 000 tyhjää ja 130 000 sisällöllistä kuvaa.

Description of the blank image detection component

Background, benefits and purpose of use of the component

Double-sided scanning is often used in the process of digitising documents. If the documents are one-sided, many blank images are created during digitisation. Machine classification into blank and content images facilitates the digitiser's work, and if the quality of identification is good enough, it also makes it possible to hide or even remove blank images from storage. If the documents are viewable by the customer in the service, the possibility to hide blank images from the view can help improve the customer experience. If blank images are identified before the subsequent stages of document processing (e.g. error detection, optical character recognition, metadata detection), the whole document processing chain can also be accelerated.

What does the component do?



The component processes the image files received as input one at a time and forms a prediction of the content of the image based on the parameters it has learned during the training. If the component identifies content in the image with a probability of more than 50%, the image is classified as having content, and otherwise blank. 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/EmptyAPI). More detailed instructions on how to use the component can be found on the attached website.

Component training

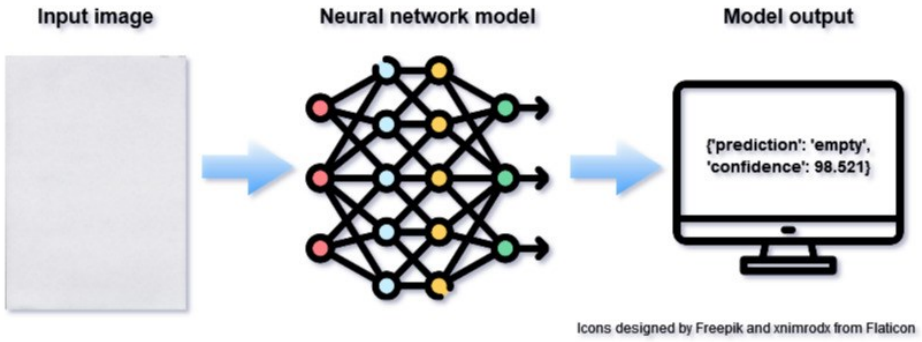
Image files in .jpg format aligned to 224 x 224 pixels have been used for component training. The component has been trained several times, and the exact amount of original training material is not known. The material used in the additional training includes approximately 100,000 blank images and 130,000 images with content.

Igenkänning av tomma sidor

Bakgrund, fördelar och ändamål med komponenten

I processen för digitalisering av dokument används ofta dubbelsidig skanning. Om dokumenten är ensidiga skapas många tomma bilder vid digitaliseringen. En maskinell klassificering av tomma bilder och innehållsliga bilder underlättar digitaliserarens arbete. När igenkänningens kvalitet är tillräckligt hög, är det också möjligt att dölja tomma bilder eller till och med ta bort tomma bilder från förvaringen. Om dokumenten finns tillgängliga för kunden i tjänsten, kan möjligheten att dölja tomma bilder från vyn bidra till att förbättra kundupplevelsen. Om tomma bilder känns igen före senare faser i processen för behandling av dokument (bl.a. feligenkänning, textigenkänning, metadataigenkänning), kan hela kedjan för behandling av dokument också påskyndas.

Vad gör komponenten?



Komponenten behandlar de bildfiler som den fått som indata en efter en och tar fram en prognos för bildens innehåll på grundval av de parametrar som den har lärt sig under utbildningen. Om komponenten känner igen innehåll i bilden med mer än 50 procent procents sannolikhet, klassificeras bilden som innehållslig och i annat fall som tom. 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/EmptyAPI). 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. Komponenten har utbildats flera gånger och det finns ingen exakt information om det ursprungliga utbildningsmaterialets omfattning. Materialet som använts vid fortbildningarna innehåller cirka 100 000 tomma bilder och 130 000 innehållsliga bilder.