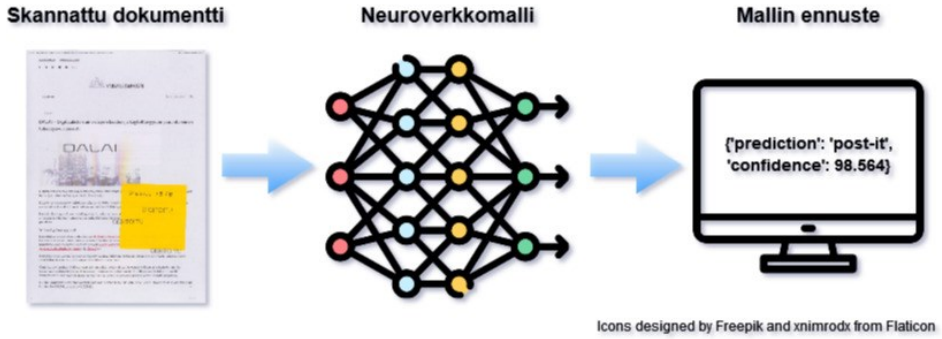
Post-itien tunnistus –komponentti

Komponentin tausta, hyöty ja käyttötarkoitus

Asiakirjojen digitointiprosessissa skannatuksi saattaa tulla dokumentteja, joiden päälle on kiinnitetty yksi tai useampi post-it-lappu. Mikäli post-it-lapun sijainti vaikuttaa asiakirjan sisällön luettavuuteen, on hyvä, jos tällaiset tapaukset voidaan tunnistaa ja tarvittaessa skannata uudelleen. Post-it-lappujen koneellinen tunnistus auttaa näin parantamaan digitaalisen aineiston laatua ja vähentää tarvetta manuaaliseen laaduntarkistukseen.

Mitä komponentti tekee?



Komponentti käsittelee syötteenä saamansa kuvatiedostot yksi kerrallaan ja muodostaa koulutuksen aikana oppimiensa parametrien pohjalta ennustuksen kuvan sisällöstä. Mikäli komponentti tunnistaa kuvasta post-it-lapun yli 50 % todennäköisyydellä, kuva luokitellaan virheelliseksi. 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/PostitAPI). Tarkemmat ohjeet komponentin käyttöön eri ympäristöissä löytyvät oheisilta verkkosivuilta.

Komponentin koulutus

Komponentin koulutukseen on käytetty .jpg-muotoisia kuvatiedostoja, joiden koko on yhdenmukaistettu 224 x 224 pikseliin. Koulutuksessa käytettyjä dokumentteja on yhteensä n. 55 000 kappaletta, joista post-it-lappuja sisältäviä dokumentteja on n. 4 000. Post-it-lappuja sisältäviä kuvia on sekä tehty itse että poimittu Kansallisarkiston massadigitointiprosessin yhteydessä manuaalisesti virheelliseksi luokitelluista kuvista. Dokumentit sisältävät vaihtelevan määrän eri kokoisia ja eri värisiä, eri puolille dokumenttia sijoitettuja post-it-lappuja, joissa voi olla myös tekstiä.

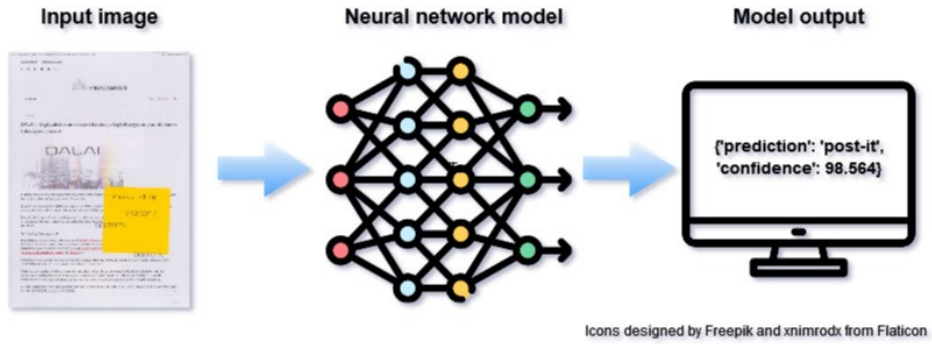
Vaikka koulutusaineistoon on pyritty kokoamaan monipuolisesti erilaisia esimerkkitapauksia, aineiston määrä on rajallinen ja komponentin luokittelussa tapahtuu myös virheitä. Esimerkiksi post-it-lappuja muistuttavat sisältöelementit, kuten värilliset neliskulmaiset tekstikentät, saattavat johtaa virheelliseen luokitukseen.

Post-it note identification component

Background, benefits and purpose of use of the component

Documents on which one or more post-it notes have been attached may be scanned in the document digitisation process. If the location of the post-it note affects the readability of the document's content, it is useful if such cases can be identified and, if necessary, scanned again. Machine identification of post-it notes thus helps improve the quality of digital material and reduces the need for manual quality control.

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 a post-it note in the image with a probability of more than 50%, the image is classified as incorrect. 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/PostitAPI). More detailed instructions on how to use the component in different environments 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. A total of 55,000 documents have been used in the training, of which approximately 4,000 are documents containing post-it notes. Images containing post-it notes have been both self-made and extracted manually from images classified as incorrect in connection with the mass digitisation process of the National Archives. The documents contain a variable number of post-it notes of different size and colour, placed in different parts of the document, and the notes may also contain text.

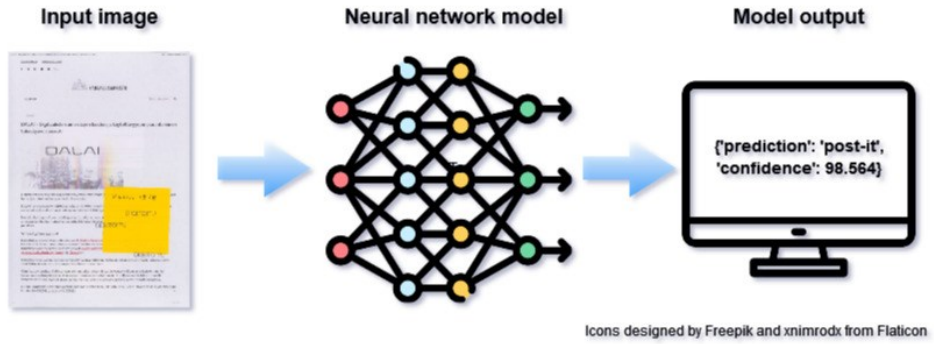
Although efforts have been made to compile a variety of example cases for the training material, the amount of the material is limited and errors also occur in the classification of the component. Among others, content elements that resemble post-it notes, such as coloured rectangular text fields, may lead to incorrect classification.

Igenkänning av post-it-lappar

Bakgrund, fördelar och ändamål med komponenten

Vid digitaliseringen av dokument kan det hända att dokument med en eller flera post-it-lappar har överlämnats för skanning. Om post-it-lappens placering påverkar läsbarheten av innehållet, är det bra om sådana fall kan kännas igen och vid behov skannas på nytt. En maskinell igenkänning av post-it-lappar bidrar därmed till att förbättra kvaliteten på digitalt material och minskar behovet av en manuell kvalitetskontroll.

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 en post-it-lapp på bilden med mer än 50 procents sannolikhet, klassificeras bilden som felaktig. 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/PostitAPI). Närmare anvisningar om användningen av komponenten inom olika miljöer 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. Det totala antalet dokument som använts i utbildningen uppgår till cirka 55 000 exemplar, av vilka cirka 4 000 är dokument som innehåller post-it-lappar. Bilder som innehåller post-it-lappar har både gjorts själva och inhämtats från bilder som i samband med Riksarkivets massdigitaliseringsprocess manuellt klassificerats som felaktiga. Dokumenten innehåller ett varierande antal post-it-lappar i olika storlekar och färger, som är placerade på olika sidor av dokumentet och som även kan innehålla text.

Trots att man har försökt samla många olika exempel till utbildningsmaterialet, är materialmängden begränsad och det uppstår också fel i komponentens klassificering. Till exempel kan innehållselement som liknar post-it-lappar, t.ex. fyrkantiga textfält i färg, leda till en felaktig klassificering.