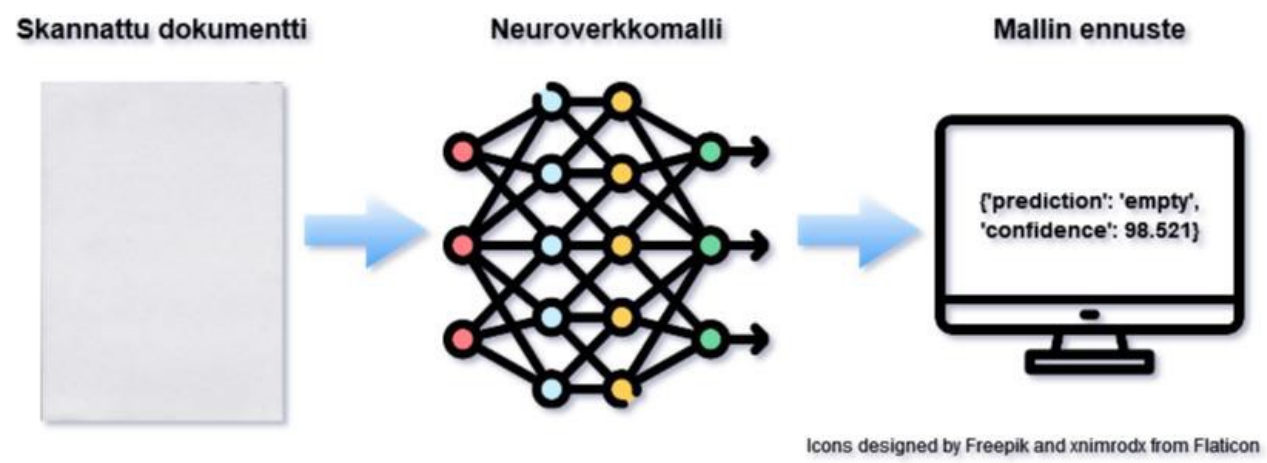
Tyhjien tunnistus -komponentin kuvaus

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.