

(19)



SUOMI - FINLAND
(FI)

PATENTTI- JA REKISTERIHALLITUS
PATENT- OCH REGISTERSTYRELSEN
FINNISH PATENT AND REGISTRATION OFFICE

(10)

FI/EP3539059 T3

(12)

**EUROOPPAPATENTIN KÄÄNNÖS
ÖVERSÄTTNING AV EUROPEISKT PATENT
TRANSLATION OF EUROPEAN PATENT SPECIFICATION**

(45)

Käännöksen kuulutuspäivä - Kungörelsedag av översättning -
Translation available to the public

24.05.2024

(97)

Eurooppapatentin myöntämispäivä - Meddelandedatum för
det europeiska patentet - Date of grant of European patent

28.02.2024

(51)

Kansainvälinen patenttiluokitus - Internationell patentklassificering -
International patent classification
G06N 3/063 (2023 . 01)
G06N 3/0464 (2023 . 01)

(96)

Eurooppapatenttihakemus - Europeisk patentansökan -
European patent application

EP17761170.4

(22)

Tekemispäivä - Ingivningsdag - Filing date

23.08.2017

(97)

Patenttihakemuksen julkiseksitilopäivä - Patentansökans
publiceringsdag - Patent application available to the public

18.09.2019

(86)

Kansainvälinen hakemus - Internationell
ansökan - International application

23.08.2017 PCT/US2017048123

(30)

Etuoikeus - Prioritet - Priority

10.11.2016 US US201615348199

23.03.2017 US US201715467382

(73)

Haltija - Innehavare - Holder

1• Google LLC, 1600 Amphitheatre Parkway, Mountain View, CA 94043, (US)

(72)

Keksijä - Uppfinnare - Inventor

1• YOUNG, Reginald Clifford, 1600 Amphitheatre Parkway, Mountain View, California 94043, (US)

2• GULLAND, William John, 1600 Amphitheatre Parkway, Mountain View, California 94043, (US)

(74)

Asiamies - Ombud - Agent

Papula Oy, P.O.Box 981, 00101 Helsinki, (FI)

(54)

Keksinnön nimitys - Uppfinningens benämning - Title of the invention

KERNEL-ASKELLUKSEN SUORITTAMINEN LAITTEISTOSSA

PERFORMING KERNEL STRIDING IN HARDWARE

PATENTTIVAATIMUKSET

1. Tietokoneella toteutettu menetelmä, joka sisältää:

5 vastaanotetaan pyyntö toteuttaa konvoluutioneuroverkko erikoiskäyttöisellä laitteistopiirillä (110), ja vastaanotetaan ja käsitellään neuroverkko-syötteitä käyttämällä neuroverkkoa aiheuttamalla laitteistopiirin (110) suorittaa käskyjä, joka neuroverkko
10 käsittää ensimmäisen konvoluutioneuroverkkokerroksen, jonka askel on suurempi kuin yksi, joka laitteistopiiri (110) on integroitu piiri neuroverkkolaskutoimitusten suorittamiseen ja sisältää matriisilaskentayksikön (120; 312), joka on sovitettu suorittamaan vektori-mat-
15 riisikertolaskuja laitteistossa ja vektorilaskentayksikön (140; 314; 602), joka sisältää pooling-piiristön (608), joka on sovitettu suorittamaan poolingia matriisilaskentayksikön (120; 312) lähdöille laitteistossa; ja

20 vasteena, generoidaan käskyt, jotka laitteistopiirin (110) suorittamana aiheuttavat laitteistopiirin (110), neuroverkon tulotensorin käsittelyn aikana, generoida kerroslähtötensori, joka vastaa ensimmäisen konvoluutioneuroverkkokerroksen ulostuloa suorittamalla
25 operaatioita, jotka sisältävät:

 matriisilaskentayksikkö (120; 312) käsittelee tulotensorin ensimmäiseen konvoluutioneuroverkkokerrokseen käyttämällä toista konvoluutioneuroverkkokerrosta, jonka askel on yhtä suuri kuin yksi, mutta
30 joka muuten vastaa ensimmäistä konvoluutioneuroverkkokerrosta, ensimmäisen tensorin muodostamiseksi;

 vektorilaskentayksikkö (140; 314; 602) nolaa ensimmäisen tensorin elementit, joita ei olisi generoitu, jos toisella konvoluutioneuroverkkokerroksella olisi ensimmäisen konvoluutioneuroverkkokerroksen askel, toisen tensorin muodostamiseksi; ja

vektorilaskentayksikön (140; 314; 602) pooling-piiristö (608) suorittaa maksimipoolingin toiselle tensorille kerroslähtötensorin muodostamiseksi, missä ensimmäisen tensorin elementtejä nollaa vektorilaskentayksikkö (140; 314; 602) käsittää: vektorilaskentayksikkö (140; 314; 602) suorittaa peitetensorin ja ensimmäisen tensorin elementtikohdaisen kertomisen toisen tensorin muodostamiseksi, missä peitetensori käsittää (i) nollia jokaisessa peitetensorin elementtikohdassa, joka vastaa ensimmäisen tensorin elementtiä, jota ei olisi generoitu, jos toisella konvoluutioneuroverkkokerroksella olisi ensimmäisen konvoluutioneuroverkkokerroksen askel, ja (ii) ykkösiä peitetensorin kunkin muun elementin kohdalla.

15

2. Tietokoneella toteutettu menetelmä, joka sisältää:

vastaanotetaan pyyntö toteuttaa konvoluutioneuroverkko erikoiskäyttöisellä laitteistopiirillä (110), ja vastaanotetaan ja käsitellään neuroverkko-syötteitä käyttämällä neuroverkkoa aiheuttamalla laitteistopiirin (110) suorittaa käskyjä, joka neuroverkko käsittää ensimmäisen konvoluutioneuroverkkokerroksen, jonka askel on suurempi kuin yksi, joka laitteistopiiri (110) on integroitu piiri neuroverkkolaskutoimitusten suorittamiseen ja sisältää matriisilaskentayksikön (120; 312), joka on sovitettu suorittamaan vektori-matriisikertolaskuja laitteistossa ja vektorilaskentayksikön (140; 314; 602), joka sisältää pooling-piiristön (608), joka on sovitettu suorittamaan poolingin matriisilaskentayksikön (120; 312) lähdöille laitteistossa; ja

vasteena, generoidaan käskyt, jotka laitteistopiirin (110) suorittamana aiheuttavat laitteistopiirin (110), neuroverkon tulotensorin käsittelyn aikana, generoida kerroslähtötensori, joka vastaa ensimmäisen

35

konvoluutioneuroverkkokerroksen ulostuloa suorittamalla operaatioita, jotka sisältävät:

matriisilaskentayksikkö (120; 312) käsittelee tulotensorin ensimmäiseen konvoluutioneuroverkkokerrokseen käyttämällä toista konvoluutioneuroverkkokerrosta, jonka askel on yhtä suuri kuin yksi, mutta joka muuten vastaa ensimmäistä konvoluutioneuroverkkokerrosta, ensimmäisen tensorin muodostamiseksi;

vektorilaskentayksikkö (140; 314; 602) nollaa ensimmäisen tensorin elementit, joita ei olisi generoitu, jos toisella konvoluutioneuroverkkokerroksella olisi ensimmäisen konvoluutioneuroverkkokerroksen askel toisen tensorin muodostamiseksi; ja

vektorilaskentayksikön (140; 314; 602) pooling-piiristö (608) suorittaa maksimipoolingin toiselle tensorille kerroslähtötensorin muodostamiseksi,

missä ensimmäisen tensorin elementtejä nollaa vektorilaskentayksikkö (140; 314; 602) käsittää:

vektorilaskentayksikkö (140; 314; 602) suorittaa ensimmäisen peitetensorin ja ensimmäisen tensorin elementtikohtaisen kertomisen muunnetun ensimmäisen tensorin muodostamiseksi, missä ensimmäinen peitetensori käsittää (i) nollia peitetensorin jokaisessa elementtikohdassa, joka vastaa ensimmäisen tensorin elementtiä, jota ei olisi generoitu, jos toisella konvoluutioneuroverkkokerroksessa olisi ensimmäisen konvoluutioneuroverkkokerroksen askel, ja (ii) vastaavan nollasta poikkeava arvon peitetensorin jokaisessa elementtikohdassa, joka vastaa ensimmäisen tensorin elementtiä, joka olisi generoitu, jos toisella konvoluutioneuroverkkokerroksella olisi ensimmäisen konvoluutioneuroverkkokerroksen askel; ja

vektorilaskentayksikkö (140; 314; 602) suorittaa toisen peitetensorin ja muunnetun ensimmäisen tensorin elementtikohtaisen kertomisen, missä toinen peitetensori käsittää kussakin elementtikohdassa, joka

vastaa ensimmäisen tensorin elementtiä, joka generoitaisiin, jos toisella konvoluutioneuroverkkokerroksella olisi ensimmäisen konvoluutioneuroverkkokerroksen askel, ensimmäisen peitetensorin vastaavan nollasta poikkeavan arvon käänteisarvon.

3. Patenttivaatimuksen 1 tai 2 mukainen menetelmä, missä patenttivaatimuksen 1 tai 2 mukaiset peitetensorit on tallennettu muistiin, johon laitteistopiiri (110) pääsee käsiksi.

4. Patenttivaatimuksen 3 mukainen menetelmä, missä useita peitetensoreja tallennetaan muistiin vastaamaan useita askelia, jotka ovat suurempia kuin yksi, vastaavasti, ja

menetelmä edelleen käsittää sen, että valitaan peitetensori mainittujen useiden peitetensorien joukosta, jotka vastaavat ensimmäisen konvoluutioneuroverkkokerroksen askelta.

5. Jonkin patenttivaatimuksen 1-4 mukainen menetelmä, missä vektorilaskentayksikön (140; 314; 602) pooling-piiristö (608), joka suorittaa maksimipoolingin, käsittää sen, että hankitaan kullekin yhdelle tai useammalle toisen tensorin ikkunalle, jotka määritetään ensimmäisen konvoluutioneuroverkkokerroksen askeleella, ikkunassa olevien elementtien maksimiarvoelementti.

6. Patenttivaatimuksen 5 mukainen menetelmä, missä kukin toisen tensorin mainitusta yhdestä tai useammasta ikkunasta on suorakaiteen muotoinen ikkuna, jonka mitat vastaavat konvoluutioneuroverkkokerroksen askelta ja sisältää toisen tensorin eri elementtejä.

7. Jonkin patenttivaatimuksen 1-6 mukainen menetelmä, missä maksimipoolingin suorittava vektorilas-

kentayksikön (140; 314; 602) pooling-piiristö (608) käsittää sen, että hankitaan toisen tensorin kullekin yhdelle tai useammalle osajoukolle osajoukon maksimiarvoelementti.

5

8. Jonkin patenttivaatimuksen 1-7 mukainen menetelmä, missä konvoluutioneuroverkkokerros on ensimmäinen neuroverkkokerros konvoluutioneuroverkossa, ja missä tulotensori on kuvaus digitaalisesta kuvasta, joka käsittää elementtejä, jotka vastaavat digitaalisen kuvan pikseleitä.

9. Jonkin patenttivaatimuksen 1-8 mukainen menetelmä, missä tulotensori tallennetaan laitteistopiiriin (110) yhtenäiseen puskuriin ja toisen konvoluutioneuroverkkokerroksen painot tallennetaan laitteistopiiriin (110) dynaamiseen muistiin (110), ja missä tulotensorin käsittely ensimmäiseen konvoluutioneuroverkkokerrokseen käyttämällä toista konvoluutioneuroverkkokerrosta käsittää:

lähetetään tulotensori yhdistetystä puskurista matriisilaskentayksikköön (120; 312);

lähetetään toisen konvoluutioneuroverkkokerroksen painot dynaamisesta muistista matriisilaskentayksikköön (120; 312); ja

käsitellään matriisilaskentayksiköllä (120; 312) tulotensori käyttämällä toisen konvoluutioneuroverkkokerroksen painoja ensimmäisen tensorin muodostamiseksi.

30

10. Järjestelmä, joka sisältää:

erikoiskäyttöisen laitteistopiiriin (110), joka on integroitu piiri neuroverkkolaskutoimitusten suorittamiseen ja joka sisältää matriisilaskentayksikön (120; 312), joka on sovitettu suorittamaan vektori-matriisikertolaskuja laitteistossa ja vektorilaskentayksikön (140; 314; 602), joka sisältää pooling-piiriin (608),

joka on sovitettu suorittamaan poolingia matriisilaskentayksikön (120; 312) lähdöille laitteistossa; ja
yhden tai useamman tallennuslaitteen, joka
tallentaa käskyjä, jotka suoritettuna laitteistopiirillä (110) aiheuttavat laitteistopiirin (110) suorittaa jonkin patenttivaatimuksen 1 - 9 mukainen menetelmä.

11. Tietokoneohjelma, joka käsittää käskyjä, jotka, jos ne suoritetaan patenttivaatimuksen 10 mukaisella erikoiskäyttöisellä laitteistopiirillä (110), aiheuttavat erikoiskäyttöisen laitteistopiirin (110) suorittaa jonkin patenttivaatimuksen 1 - 9 mukainen menetelmä.