(19) SUOMI - FINLAND
(FI)

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

(10) **FI/EP3761235 T3**

(12) **EUROOPPAPATENTIN KÄÄNNÖS**
ÖVERSÄTTNING AV EUROPEISKT PATENT
TRANSLATION OF EUROPEAN PATENT SPECIFICATION

(45) Käännöksen kuulutuspäivä - Kungörelsedag av översättning - Translation available to the public **23.07.2024**

(97) Eurooppapatentin myöntämispäivä - Meddelandedatum för det europeiska patentet - Date of grant of European patent **01.05.2024**

(51) Kansainvälinen patenttiluokitus - Internationell patentklassificering - International patent classification
G06N 3/063 (2023 . 01)
G06F 7/78 (2006 . 01)
G06F 17/16 (2006 . 01)
G06N 3/045 (2023 . 01)
G06N 3/084 (2023 . 01)

(96) Eurooppapatenttihakemus - Europeisk patentansökan - European patent application EP20174935.5

(22) Tekemispäivä - Ingivningsdag - Filing date **09.03.2018**

(97) Patenttihakemuksen julkiseksitulopäivä - Patentansökans publiceringsdag - Patent application available to the public **06.01.2021**

(30) Etuoikeus - Prioritet - Priority

09.03.2017 US US201715455024

(73) Haltija - Innehavare - Holder

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

(72) Keksijä - Uppfinnare - Inventor

1• Young, Reginald Clifford , Google LLC 1600 Amphitheatre Parkway , Mountain View, CA 94043 , (US)
2• Irving, Geoffrey , Google LLC 1600 Amphitheatre Parkway , Mountain View, CA 94043 , (US)

(74) Asiamies - Ombud - Agent

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

(54) Keksinnön nimitys - Uppfinningens benämning - Title of the invention

NEUROVERKKOMATRIISIEN TRANSPONOINTI LAITTEISTOSSA

TRANSPOSING NEURAL NETWORK MATRICES IN HARDWARE

Patenttivaatimukset

1. Menetelmä (800) neuroverkkomatriisin transponoimiseksi laitteistopiirissä (110), jolloin menetelmä käsittää seuraavat:

suoritetaan iteratiivisesti seuraavat operaatiot laitteistopiirin matriisinlaskentayksikössä (120), jolloin neuroverkkomatriisi on painotusmatriisi ja jolloin laitteistopiiri suorittaa matriisin kertomisoperaatioita vektorien kertomisoperaatioiden sarjana käyttäen matriisinlaskentayksikköä, jolloin matriisinlaskentayksikkö on kaksiulotteinen systolinen matriisi (406), joka sisältää useita soluja (404), ja jolloin jokaisessa iteraatiossa jokainen solu prosessoi tietyn painotussyötteen soluun ja tietyn aktivointisyötteen soluun muodostaen tulon ja lisää kyseisen tulon ensimmäisen viereisen solun kertyneeseen tulokseen muodostaen kertyneen tuloksen ja välittää muodostetun kertyneen tuloksen toiseen viereiseen soluun:

jaetaan välivaiheen neuroverkkomatriisi iteraationa useiksi alimatriiseiksi nykyiseen alajakoon,

luodaan (802) useita vektoreita, joista jokainen vastaa välivaiheen neuroverkkomatriisin kutakin riviä ja sisältää välivaiheen neuroverkkomatriisin kunkin rivin arvot,

hankitaan (804) kyseisille useille vektoreille useita osittaisia tunnistusmatriiseja, jotka on konfiguroitu poimimaan tietty osa arvoista kustakin useista vektoreista ja samalla mitätöimään arvojen loppuosa, kun kyseiset useat osittaiset tunnistusmatriisit kerrotaan kyseisillä useilla vektoreilla,

kerrotaan (806) kukin useista vektoreista yhdellä tai useammalla kyseisistä useista osittaisista tunnistusmatriiseista, jolloin muodostuu päivitetyn neuroverkkomatriisin rivi, jolloin päivitetty neuroverkkomatriisi sisältää neuroverkkomatriisin elementtejä, mutta siinä nykyisen alajaon useiden alimatriisien ylempi oikea ja alempi vasen neljännes on vaihdettu, ja

yhdistetään (810) muodostetut rivit välivaiheen neuroverkkomatriisin päivittämiseksi ja

muodostetaan kaikkien suoritettujen operaatioiden iteraatioiden perusteella välivaiheen neuroverkkomatriisi transponoituna neuroverkkomatriisina.

2. Patenttivaatimuksen 1 mukainen menetelmä (800), joka käsittää lisäksi sen, että vastaanotetaan pyyntö suorittaa laskutoimituksia neuroverkolle laitteistopiirissä, jolloin pyyntö määrittää neuroverkkomatriisille suoritettavan transponointitoiminnon.

3. Patenttivaatimuksen 1 mukainen menetelmä (800), joka käsittää lisäksi seuraavat:

määritetään, että neuroverkkomatriisi ei ole *i* x *i* -matriisi, jossa i on vektorin pituusarvo laitteistopiirissä;

tämän perusteella päivitetään neuroverkkomatriisi muodostamalla *i* x *i* -matriisi lisäämällä nollia aiemman neuroverkkomatriisin reunoille ja

muunnetaan transponoitu neuroverkkomatriisi sen transponointia edeltävään tilaan poistamalla transponoinnin aikana lisätyt nollat.

4. Patenttivaatimuksen 1 mukainen menetelmä (800), joka käsittää lisäksi seuraavat:

hankitaan data, joka osoittaa, että yksi tai useampi neuroverkkomatriisin arvo on nolla-arvo; ja

estetään laitteistopiiriä suorittamasta mitään operaatiota arvojoukolle, joka sisältää vähintään yhden neuroverkkomatriisin yhdestä tai useammasta arvosta, jotka ovat nolla-arvoja.

5. Järjestelmä (100), joka käsittää laitteistopiirin (110) ja yhden tai useamman tietokoneen ja yhden tai useamman komentoja tallentavan tallennuslaitteen, jolloin laitteistopiiri (110) käsittää matriisinlaskentayksikön (120), joka on kaksiulotteinen systolinen matriisi (406), ja jolloin kaksiulotteinen systolinen matriisi (406) sisältää useita soluja (404) ja jolloin komennot kykenevät, kun kyseinen yksi tai useampi tietokone suorittaa ne, saamaan järjestelmän suorittamaan operaatioita neuroverkkomatriisin transponoimiseksi, mikä käsittää seuraavat:

suoritetaan iteratiivisesti seuraavat operaatiot matriisinlaskentayksikölle (120), jolloin neuroverkkomatriisi on painotusmatriisi ja jolloin laitteistopiiri suorittaa matriisin kertomisoperaatioita vektorien kertomisoperaatioiden sarjana käyttäen matriisinlaskentayksikköä, ja jolloin jokaisessa iteraatiossa jokainen solu prosessoi tietyn painotussyötteen soluun ja tietyn aktivointisyötteen soluun muodostaen

tulon ja lisää kyseisen tulon ensimmäisen viereisen solun kertyneeseen tulokseen muodostaen kertyneen tuloksen ja välittää muodostetun kertyneen tuloksen toiseen viereiseen soluun:

jaetaan välivaiheen neuroverkkomatriisi iteraationa useiksi alimatriiseiksi nykyiseen alajakoon,

luodaan useita vektoreita, joista jokainen vastaa välivaiheen neuroverkkomatriisin kutakin riviä ja sisältää välivaiheen neuroverkkomatriisin kunkin rivin arvot,

hankitaan kyseisille useille vektoreille useita osittaisia tunnistusmatriiseja, jotka on konfiguroitu poimimaan tietty osa arvoista kustakin useista vektoreista ja samalla mitätöimään arvojen loppuosa, kun kyseiset useat osittaiset tunnistusmatriisit kerrotaan kyseisillä useilla vektoreilla,

kerrotaan kukin useista vektoreista yhdellä tai useammalla kyseisistä useista osittaisista tunnistusmatriiseista, jolloin muodostuu päivitetyn neuroverkkomatriisin rivi, jolloin päivitetty neuroverkkomatriisi sisältää neuroverkkomatriisin elementtejä, mutta siinä nykyisen alajaon useiden alimatriisien ylempi oikea ja alempi vasen neljännes on vaihdettu, ja

yhdistetään muodostetut rivit välivaiheen neuroverkkomatriisin päivittämiseksi ja

muodostetaan kaikkien suoritettujen operaatioiden iteraatioiden perusteella välivaiheen neuroverkkomatriisi transponoituna neuroverkkomatriisina.

6. Patenttivaatimuksen 5 mukainen järjestelmä (100), jossa kyseinen yksi tai useampi tietokone suorittaa operaatioita, jotka käsittävät lisäksi sen, että

vastaanotetaan pyyntö suorittaa laskutoimituksia neuroverkolle laitteistopiirissä (110), joka pyyntö määrittää neuroverkkomatriisille suoritettavan transponointitoiminnon.

7. Patenttivaatimuksen 5 mukainen järjestelmä (100), jossa kyseinen yksi tai useampi tietokone suorittaa operaatioita, jotka käsittävät lisäksi seuraavat:

määritetään, että neuroverkkomatriisi ei ole *i* x *i* -matriisi, jossa i on vektorin pituusarvo laitteistopiirissä (110);

tämän perusteella päivitetään neuroverkkomatriisi muodostamalla *i* x *i* -matriisi lisäämällä nollia aiemman neuroverkkomatriisin reunoille ja

muunnetaan transponoitu neuroverkkomatriisi sen transponointia edeltävään tilaan poistamalla transponoinnin aikana lisätyt nollat.

8. Patenttivaatimuksen 5 mukainen järjestelmä (100), jossa kyseinen yksi tai useampi tietokone suorittaa operaatioita, jotka käsittävät lisäksi seuraavat:

hankitaan data, joka osoittaa, että yksi tai useampi neuroverkkomatriisin arvo on nolla-arvo; ja

estetään laitteistopiiriä (110) suorittamasta mitään operaatiota arvojoukolle, joka sisältää vähintään yhden neuroverkkomatriisin yhdestä tai useammasta arvosta, jotka ovat nolla-arvoja.

9. Patenttivaatimuksen 5 mukainen järjestelmä (100), jossa kyseinen yksi tai useampi tietokone suorittaa operaatioita, jotka käsittävät lisäksi kyseisen pyynnön lähettämisen laitteistopiirille.

10. Tietokoneen tallennusväline, johon on koodattu komentoja, jotka järjestelmän (100), joka käsittää laitteistopiirin (110) ja yhden tai useamman tietokoneen, jolloin laitteistopiiri (110) käsittää matriisilaskentayksikön (120), joka on kaksiulotteinen systolinen matriisi (406), joka sisältää useita soluja (404), suorittamina saavat järjestelmän suorittamaan operaatioita neuroverkkomatriisin transponoimiseksi, mikä käsittää seuraavat:

suoritetaan iteratiivisesti seuraavat operaatiot matriisinlaskentayksikölle (120), jolloin neuroverkkomatriisi on painotusmatriisi ja jolloin laitteistopiiri suorittaa matriisin kertomisoperaatioita vektorien kertomisoperaatioiden sarjana käyttäen matriisinlaskentayksikköä, ja jolloin jokaisessa iteraatiossa jokainen solu prosessoi tietyn painotussyötteen soluun ja tietyn aktivointisyötteen soluun muodostaen tulon ja lisää kyseisen tulon ensimmäisen viereisen solun kertyneeseen tulokseen muodostaen kertyneen tuloksen ja välittää muodostetun kertyneen tuloksen toiseen viereiseen soluun:

jaetaan välivaiheen neuroverkkomatriisi iteraationa useiksi alimatriiseiksi nykyiseen alajakoon,

luodaan useita vektoreita, joista jokainen vastaa välivaiheen neuroverkkomatriisin kutakin riviä ja sisältää välivaiheen neuroverkkomatriisin kunkin rivin arvot, kun kyseiset useat osittaiset tunnistusmatriisit kerrotaan kyseisillä useilla vektoreilla,

hankitaan kyseisille useille vektoreille useita osittaisia tunnistusmatriiseja, jotka on konfiguroitu poimimaan tietty osa arvoista kustakin useista vektoreista ja samalla mitätöimään arvojen loppuosa,

kerrotaan kukin useista vektoreista yhdellä tai useammalla kyseisistä useista osittaisista tunnistusmatriiseista, jolloin muodostuu päivitetyn neuroverkkomatriisin rivi, jolloin päivitetty neuroverkkomatriisi sisältää neuroverkkomatriisin elementtejä, mutta siinä nykyisen alajaon useiden alimatriisien ylempi oikea ja alempi vasen neljännes on vaihdettu, ja

yhdistetään muodostetut rivit välivaiheen neuroverkkomatriisin päivittämiseksi ja muodostetaan kaikkien suoritettujen operaatioiden iteraatioiden perusteella välivaiheen neuroverkkomatriisi transponoituna neuroverkkomatriisina.

11. Patenttivaatimuksen 10 mukainen tietokoneen tallennusväline, jolloin kyseiset operaatiot käsittävät lisäksi sen, että

vastaanotetaan pyyntö suorittaa laskutoimituksia neuroverkolle laitteistopiirissä (110), joka pyyntö määrittää neuroverkkomatriisille suoritettavan transponointitoiminnon.

12. Patenttivaatimuksen 10 mukainen tietokoneen tallennusväline, jolloin kyseiset operaatiot käsittävät lisäksi seuraavat:

määritetään, että neuroverkkomatriisi ei ole *i* x *i* -matriisi, jossa i on vektorin pituusarvo laitteistopiirissä (110);

tämän perusteella päivitetään neuroverkkomatriisi muodostamalla *i* x *i* -matriisi lisäämällä nollia aiemman neuroverkkomatriisin reunoille ja

muunnetaan transponoitu neuroverkkomatriisi sen transponointia edeltävään tilaan poistamalla transponoinnin aikana lisätyt nollat.