LV 15414

(19) **LATVIJAS REPUBLIKAS**
PATENTU VALDE

(11) **LV 15414 B1**

(51) Starpt.pat.kl. **G06N3/04**
G06T1/40

Latvijas patents izgudrojumam
2007g. 15.februāra Latvijas Republikas likums

(12) **Īsziņas**

(21) Pieteikuma numurs:	P-17-87
(22) Pieteikuma datums:	14.12.2017
(43) Pieteikuma publikācijas datums:	20.06.2019
(45) Patenta publikācijas datums:	20.03.2021

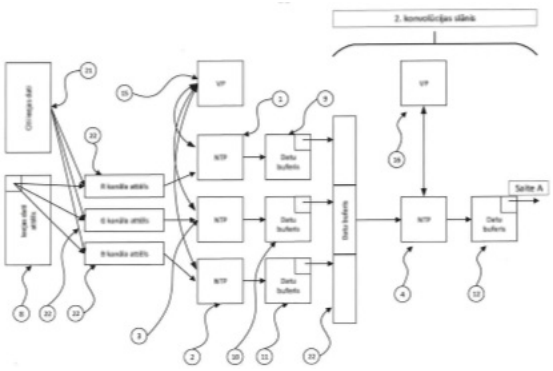
(71) Īpašnieks(i):
RĪGAS TEHNISKĀ UNIVERSITĀTE,Kaļķu iela 1,Rīga,LV

(72) Izgudrotājs(i):
Agris ŅIKITENKO (LV)
Kārlis BERKOLDS (LV)

(54) Izgudrojuma nosaukums: DZIĻĀS APMĀCĪBAS NEIRONU TĪKLA APARATŪRA UN ATTĒLU APSTRĀDES PAŅĒMIENS
HARDWARE ARCHITECTURE FOR DEEP LEARNING NEURON NETWORK AND METHOD FOR IMAGE PROCESSING

(57) Kopsavilkums:

Izgudrojums attiecas uz elektroniku, konkrēti uz elektroniskām sistēmām, kas balstās daudzu specializētu mikroprocesoru izmantošanā vienota uzdevuma risināšanai. Izgudrojums ir dziļās apmācības neironu tīkla aparatūra, kas satur vairākus vispārējas nozīmes programmējamus mikroprocesorus un neironu tīklu mikroprocesorus un ir apvienoti vienotā daudzslāņu dziļās apmācības aparatūrā realizētā neironu tīklā. Katrs no tīkla slāņiem satur vienu vadības mikroprocesoru un vienu neironu tīkla mikroprocesoru, šādi veidojot unificētu konvolūcijas slāņa struktūru, kas ļauj salīdzinoši vienkārši mērogot risinājumu un slāņa vadības programmatūru pielāgot konkrētu uzdevumu veikšanai. Vadības mikroprocesori var tikt sinhronizēti, veidojot datu apstrādes konveijeru, kas ļauj vienlaikus dažādos konvolūcijas slāņos apstrādāt dažādus ieejas datus, būtiski uzlabojot kopējo risinājuma veiktspēju. Piedāvāts paņēmiens attēlu atpazīšanai, izmantojot izgudroto dziļās apmācības neironu tīkla aparatūru.



IZGUDROJUMA APRAKSTS

[001] Izgudrojums attiecas uz elektroniku, konkrēti uz elektroniskām sistēmām un paņēmieniem, kas balstās daudzu specializētu mikroprocesoru izmantošanā vienota uzdevuma risināšanai. Atbilstoši starptautiskajai klasifikācijai, izgudrojuma joma ir G06F 9/28.

Zināmais tehnikas līmenis

[002] Dziļās apmācības neironu tīkli jeb konvolūciju tīkli paredzēti paaugstinātas grūtības uzdevumu risināšanai mašīnapmācības jomā. Parasti dziļās apmācības neironu tīkli ir izmantoti, kā programmatūras konstrukcijas ar mērķi modelēt dabīgo neironu tīklu struktūras, šādi pēc iespējas atdarinot tādas sarežģītas dabīgu neironu struktūras kā smadzenes. Ņemot vērā šādu struktūru sarežģītību, to realizācija aparatūras veidā ir apgrūtināta, lai arī eksistē aparatūras risinājumi vienkāršāku neironu tīklu struktūrām. Piemēram, patentā [1] aprakstītais risinājums nodrošina viena slāņa neironu tīklu ar vairāku aktivizācijas funkciju iespējām [2].

[003] Patenti [3] un [4] apraksta konkrētus algoritmus specializētiem lietojumiem, lai nodrošinātu augstāku objektu atpazīšanas attēlos veiktspēju un sarežģītāku uzdevumu risinājumus, izmantojot neironu tīklus attēlu apstrādes jomās.

[004] Patentā [3] ir aprakstīts paņēmiens ātrai konkrētu apstrādājamā attēlā atpazīto objektu ierāmēšanai, kas tālāk var tikt izmantots, lai atpazītos objektus apstrādātu kādā augstāka līmeņa programmatūras sistēmā. Patentā [3] pieminētais algoritms nodrošina divu pakāpju konvolūcijas tīklu izmantošanu – objektu atpazīšanai un to pozīciju noteikšanai attēlā.

[005] Pretēji minētajam, patentā [4] piedāvātais risinājums koncentrējas uz programmatūras risinājumiem, kur galvenā priekšrocība ir salīdzinoša elastība, lai pielāgotu algoritmus konkrētam pielietojumam. Galvenais trūkums ir salīdzinoši lēna darbība, kas saistīta ar lielu aprēķinu daudzu algoritma darbības laikā.

[006] Lai gan piedāvātie risinājumi programmatūras sistēmām ļauj salīdzinoši labi apstrādāt attēlus un veikt to pozīciju noteikšanu attēlā, galvenā problēma arvien ir saistīta ar to ātrdarbību iegultās sistēmās, kuras raksturīgas ar ierobežotu atmiņas apjomu un skaitļošanas jaudu.

Izgudrojuma mērķis un būtība

[007] Lai minēto problēmu risinātu, dziļās apmācības neironu tīklus nepieciešams realizēt aparatūrā, kas nodrošinās augstu to ātrdarbību, kā arī iespēju tos integrēt neliela izmēra iegultās iekārtās. Izgudrojums apvieno vairākus mākslīgo neironu tīklu mikroprocesorus vienotā iegultā aparatūras sistēmā, kurā katrs no neironu tīkla mikroprocesoriem pilda konvolūcijas slāņa vai ieejas slāņa funkcijas, kas papildināti ar vispārēja lietojuma mikroprocesoriem kopējai sistēmas vadībai un saziņai arī ar ārējām sistēmām. Ieejas slānis

līdzīgi citiem slāņiem sastāv no vadības un specializētā mikroprocesora. Vadības mikroprocesors ir savienots ar trīs ieejas slāņa specializētajiem mikroprocesoriem – vienslāņa neironu tīkla mikroprocesori (NTP), kas ieejas datus sagatavo tālākai apstrādei. Papildus minētajam vadības mikroprocesors ir savienots ar citiem sensoriem, papildinot ieejas datu masīvu. Savienojumi starp ieejas slāni un pārējiem konvolūcijas slāņiem, tiek realizēti gan vadības mikroprocesoru līmenī, gan NTP līmenī, kur vadības mikroprocesoru savienojums sinhronizē slāņus, bet NTP savienojumi ļauj nodot datus no viena slāņa nākošajam. Starp katriem diviem slāņiem ir datu buferis, kas nodrošina datu īslaicīgu uzglabāšanu, pirms tiek nodotie nākošajā slāņa NTP.

[008] Paņēmiens attēlu apstrādei, kuru realizē izgudrojuma aparatūra, ietver šādus soļus: (a) sākotnējā attēla (8) kombinēšana ar citiem ieejas datiem (21) un ieejas datu buferu (22) aizpildīšana; (b) datu nosūtīšana no datu buferiem (22 – R, G, B kanālu attēli) uz ieejas slāņu vienslāņa neironu tīkla mikroprocesoriem (NTP) (1, 2 un 3); (c) pēc vadības mikroprocesora (15) komandas datu apstrādes neironu tīklu mikroprocesoros (1, 2 un 3) un izejas buferu (9, 10 un 11) aizpildīšanas, kas tālāk tiek apkopoti nākošā līmeņa datu buferī (22); (d) pēc kārtas tiek iedarbināti 2., 3. un 4. konvolūcijas slāņi, nosūtot tiem ieejas datus un vadības komandu par datu apstrādi, kurus attiecīgi iedarbina pēc vadības mikroprocesoru (16, 17 un 18) komandas; (e) izejas dati tiek saglabāti datu buferī – izejas dati (20).

[009] Izgudrojums ir paskaidrots ar šādiem zīmējumiem:

1. zīm. Izgudrojuma vienslāņa neironu tīkla mikroprocesori (NTP) (1)–(7) atbilstoši patentam [1]; sākotnējais attēls (8), kas izteikts kā punktu matrica; citi ieejas dati (21); dažādu konvolūcijas slāņu īpašību attēli/matricas (9)–(14); dažāda apjoma datu buferi (22); izejas dati (20); galvenais vadības mikroprocesors (15), kas veic ieejas datu apkopošanu un sagatavošanu pirmā slāņa neironu tīklam, kombinējot sākotnējā attēla daļas – R, G un B kanāla attēlus (22) un citus ieejas datus (21).

[010] 2. zīm. Dažādu konvolūcijas slāņu vadības mikroprocesori (16)–(19), kas var tikt apvienoti vienotā, lielākas skaitļošanas jaudas mikroprocesorā, atkarībā no konkrēta pielietojuma. 1. un 2. zīmējums kopā paskaidro konkrētas variācijas slēguma konceptuālo shēmu. Shēmā 1. zīmējums attiecas uz ieejas slāni un pirmo konvolūcijas slāni, bet 2. zīmējums vairākus sekojošos konvolūcijas slāņus un izejas slāni.

Izgudrojuma realizācijas piemērs

[011] Dziļās apmācības neironu tīkla aparatūra (1. un 2. zīm.) sastāv no datu ieejas slāņa, konvolūciju slāņiem un apstrādāto datu izejas slāņa (1. un 2. zīm.), kas kopā veido aparatūrā realizētu daudzslāņu neironu tīkla struktūru. Katrs no tīkla slāņiem ieejā saņem apstrādājamo datu buferi (22), veic tā sagatavošanu neironu tīkla apstrādei ar vadības mikrokontrolieru

starpniecību (15–19), kas nodrošina arī neironu tīkla mikroprocesoru (1–7) vadību, veic datu apstrādi ar neironu tīkla mikroprocesoru palīdzību (1–7) un apstrādātos datus datu bufera (22) veidā nosūta nākošajam datu konvolūcijas slānim vai izvades slānim. Konkrētā realizācija paredz izmantot *CogniMem® CM1K* procesorus, bet par vadības procesoriem var izmantot 32 bitu *ARM®* tipa procesorus. Vadības komandu un datu nosūtīšanai izmanto virknes tipa pieslēgumu, bet datu buferim (22) vadības procesoru (15–19) atmiņu.

[012] Izgudrojuma risinājums tādejādi sastāv no trīs tehniski atšķirīgiem slāņiem: datu ieejas slāņa, vairākiem konvolūcijas slāņiem un izejas slāņa. Katra slāņa faktiskā darbība ir atkarīga no attiecīgā slāņa vadības mikroprocesoriem (15–19), kas nodrošina slāņa specifisko funkciju veikšanu. Tādejādi konvolūciju slāņu skaits var tikt mainīts atkarībā no konkrētā uzdevuma veikšanas, nodrošinot augstu mērogojamību, kā arī risinājuma kopējā darbība pielāgota konkrētam uzdevumam, pateicoties slānu vadības mikroprocesoru programmatūras pielāgošanas iespējām.

[013] Papildus minētajam slāņu savstarpējā darbība ir neatkarīga vienam no otras, kas ļauj vienlaikus apstrādāt vairākus ieejas datu masīvus, t.i. kamēr viena ieejas masīva dati tiek apstrādāti 2. vai 3. konvolūciju slānī, datu ieejas slānis jau var uzsākt nākamā datu masīva apstrādi.

[014] Šeit aprakstītais izgudrojuma variants tiek uzskatīts par labāko, jo nodrošina minimālo nepieciešamo konvolūcijas soļu skaitu vizuālu objektu atpazīšanai. Neskatoties uz to, tas var tikt modificēts, izmantojot lielāku skaitu konvolūcijas slāņu, sinhronizētus vadības mikroprocesorus (15–19) ātrdarbības paaugstināšanai sistēmā kopumā, kā arī mainīt konvolūcijas slāņu datu bufera (22) izmērus, kas var tikt pielāgots konkrēta uzdevuma veikšanai.

IZMANTOTIE INFORMĀCIJAS AVOTI

1. IN2H2. Hardware enhancements to radial basis function with restricted coulomb energy learning and/or K-nearest neighbor based neural network classifiers. USA patent US9269041 B2. 2016-02-23.
2. *Bishop C.M. Neural networks for pattern recognition. Oxford: Clarendon Press, 1995. 499 pp. ISBN 0198538499.*
3. ADOBE SYSTEMS INCORPORATED. Image assessment using deep convolutional neural networks. US patent application US20160035078 A1. 2016.02.04.
4. BAIDU USA LLC. Systems and methods for end-to-end object detection. US patent application US20170147905 A1. 2017.05.25.

PRETENZIJAS

1. Dziļās apmācības neironu tīkla aparatūra, kas satur vairākus vadības mikroprocesorus (15, 16, 17 un 18) un neironu tīkla mikroprocesorus (1, 2, 3, 4, 5 un 6), datu buferus (9, 10, 11, 12, 13 un 22), kurus vada neironu tīklu mikroprocesori un izejas datus (20), kas atšķiras ar to, ka katram neironu tīkla mikroprocesoriem (4, 5 un 6) atbilst vismaz viens vadības mikroprocesors (16, 17 un 18), kas kopā veido konvolūcijas slāni.

2. Paņēmiens attēlu atpazīšanai, kuru realizē aparatūra saskaņā ar 1. pretenziju, kas ietver šādus secīgus soļus:

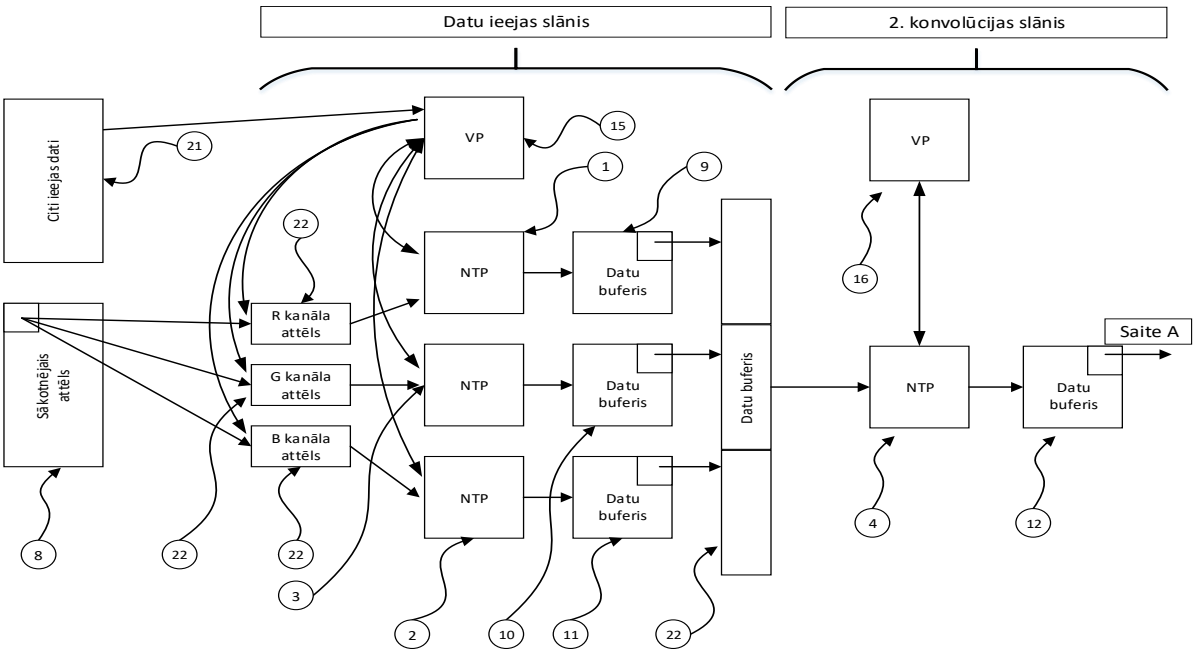
a. sākotnējā attēla (8) datu kombinēšana ar citiem ieejas datiem (21) un datu buferu (22) aizpildīšana;

b. datu nosūtīšana no buferiem (22) uz ieejas slāņa neironu tīklu mikroprocesoriem (1, 2 un 3);

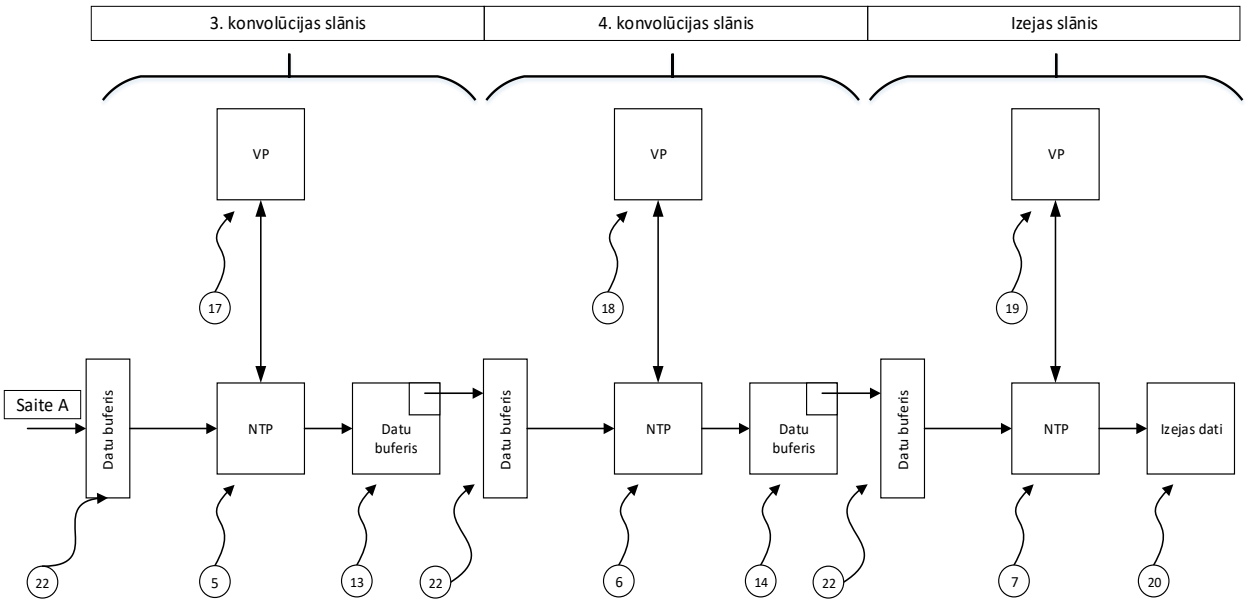
c. pēc vadības mikroprocesora (15) komandas datu apstrādes neironu tīklu mikroprocesoros (1, 2 un 3) un izejas buferu (9, 10 un 11) aizpildīšanas, kas tālāk tiek apkopoti nākošā līmeņa datu buferī (22);

d. pēc kārtas tiek iedarbināti 2., 3. un 4. konvolūcijas slāņi, nosūtot tiem ieejas datus no datu buferiem (22) un padodot vadības komandu par datu apstrādi, kurus attiecīgi iedarbina pēc vadības mikroprocesoru (16, 17 un 18) komandas;

e. izejas dati (20) tiek saglabāti atbilstošā datu buferī.


Datu ieejas slānis
2. konvolūcijas slānis
Citi ieejas dati
Sākotnējais attēls
R kanāla attēls
G kanāla attēls
B kanāla attēls
VP
NTP
NTP
NTP
Datu buferis
Datu buferis
Datu buferis
Datu buferis
VP
NTP
Datu buferis
Saite A
21
15
1
9
22
16
8
22
22
3
2
10
11
22
4
12

1.zīm.


3. konvolūcijas slānis
4. konvolūcijas slānis
Izejas slānis
VP
VP
VP
17
18
19
Saite A
Datu buferis
NTP
Datu buferis
Datu buferis
NTP
Datu buferis
Datu buferis
NTP
Izejas dati
22
5
13
22
6
14
22
7
20

2.zīm.