

SZAKDOLGOZAT

Palkovics Dénes

Debrecen
2019

Debreceni Egyetem
Informatikai Kar

Hibrid fordítás HPC környezetben

Témavezető:

Dr Kovács László
adjunktus

Készítette:

Palkovics Dénes
mérnökinformatikus BSc

Debrecen, 2019

Tartalomjegyzék

Bevezetés	1
1. Neurális hálózatok és a Deep Learning	3
1.1. A neurális hálózatok elmélete	3
1.2. Deep Learning	4
1.3. Függvények, algoritmusok	6
1.3.1. Neuronok aktivációs függvényei	6
1.3.2. Veszteség függvények	7
1.3.3. A hálózat tanító algoritmusai: az optimalizálók	8
Regresszió optimalizálás	10
2. Mély tanuló alkalmazások fejlesztése	12
2.1. TensorFlow	12
2.2. Keras	14
2.2.1. Neurális hálózat definiálása	14
2.2.2. Hálózat betanítása és betanított hálózat futtatása	15
2.2.3. Hatékonyságvizsgálat	16
3. Deep Learning platformok	18
3.1. hybrid deep learning	18
3.2. nGraph	18
3.2.1. Motiváció	19
3.2.2. Gráf szintű optimalizálás	19
3.2.3. Skálázható keretrendszer integráció	20
3.2.4. Növekvő kernel szám	21
3.3. Grafikus processzorok a mély tanulásban	22
3.3.1. Nvidia által fejlesztett gyorsítókártyák	22
3.3.2. Intel grafikus processzorok	24
3.4. Myriad X és az Intel Neural Computer Stick 2	25
3.4.1. Az NCS2 használata OpenVINO keretrendszerben	26

Modell átkonvertálása	26
Inference Engine	28
3.5. Google Edge TPU	30
3.6. Új gyorsítók: Intel Nervana Neural Network Processor	31
3.6.1. NNP-T	31
Hardver tulajdonságai	32
Az NNP-T programozásához használatos szoftver együttes	33
NNP-T Programozási modell	34
3.6.2. NNP-I	34
Összefoglalás	36
Irodalomjegyzék	38
Köszönetnyilvánítás	41

Bevezetés

A Deep learning avagy a mélytanulás a gépi tanulás neurális hálózatokat alkalmazó technikája napjaink egyik legnépszerűbb technológiája, melynek fejlesztését számos kutató intézmény és nagyvállalat végzi. Megjelent a polgári életben is. Felhőalapú alkalmazások háttérében működik, többek között a Google online szolgáltatásaiban és már alkalmazzák a hordozható eszközök, táblagépek és mobiltelefonok biometrikus személyazonosításra.

A mélytanulás használata rengeteg számítási kapacitást igényel —ez bizonyos alkalmazások esetén költséges és nagy méretű számítógépeket jelent— így sok helyen kizorul a használata, illetve teletmetria formájában érhető el csak. Az 5G-nek hála, komolyabb alkalmazásokhoz is felhasználható lesz a felhő technológián működő gépi tanulás. Ennek ellenére igény volna arra, hogy helyben elérhető legyen ez a technika. Ilyen lehet az orvosi alkalmazás, ahol számít a magas rendelkezésre állás vagy az autonóm robotok és önvezető autók, melyeknek bizonyos helyzetekben ott is kell működniük, ahol nincs rádiókapcsolat vagy internet elérés, nem is beszélve a hordozható eszközök olyan funkcióiról melyek használata frekvenciált.

Mikor fellendült a kutatása, legjobb hardverek erre a feladatra a fejlett grafikus kártyák voltak, melyek processzorainak számítási kapacitása és utasításkészlete alkalmassá tette, hogy a neurális hálózatokkal kapcsolatos számításokat hatékonyan végezze. Azonban az iparban megjelentek speciális hardverek kifejezetten neurális hálózatok futtatására optimalizálva, hogy ki tudják elégíteni a megnövekedett számítási igényt, amit a technológia egyre szélesebb körű bevezetése generál. A fenntarthatóság végett azonban nem szabad kihasználatlanul hagyni a már meglévő erőforrásokat. Témavezetőm, Dr. Kovács László projektje, a HuSSar nevet viselő hibrid architektúrájú szuperszámítógép is részben ebből az indíttatásból született. A HuSSar olyan hardverekből tevődik össze, melyek szerverek komponenseként régóta ott van az iparban. Egyedi hibrid architektúrája lehetőséget ad arra, hogy a neurális hálózatokkal kapcsolatos különféle számításokat olyan processzoron futtassuk, melyek azt optimálisan képesek végrehajtani így jelentős teljesítménynövekedés érhető el vele. Ehhez szükséges még egy olyan keretrendszer, mely képes ezeket a számításokat ekképpen optimalizálni.

A Deep learning a szemem láttára fejlődött ki a kezdeti kísérletekből, a mindennapi életben is használt csúcstechnológiává. Úgy érzem leendő szakemberként most van arra alkalmam, hogy közelebbről is megismerkedjek vele, kivehessem részem a fejlesztésében. Észrevettem,

hogy az ipar is nagy erővel fejleszti, ezért úgy vélem, hogy ez a tudás számomra nagyon jövedelmező lehet a munkaerőpiacon is. Témavezetőm fejlesztésével, a HuSSar-al az egyik általa tartott egyetemi kurzus során találkoztam, mikor azt megmutatta nekünk. Beszélt az eszköz felépítéséről és arról, milyen célból kezdte a fejlesztést. Továbbá látom unokaöcsém sikereit, aki ezen a területen kutat. Ezek miatt éreztem úgy, hogy ebben a témában szeretnék dolgozni, ha lehet az egyetemi tanulmányaim után is.

Ebben a szakdolgozatban szeretnék beszámolni, mit sikerült megtudnom a Deep Learning-ről, milyen új megvalósítások születtek az iparban és hogyan boldogultam ezekkel a technológiákkal. Eredeti célkitűzésem az Intel fejlesztés alatt álló *nGraph* nevű környezetének fordítása és telepítése volt a fentebb említett HuSSar-ra. Ez a keretrendszer kifejezetten a neurális hálózatok olyan módú futtatására lett fejlesztve, ahol a hardver több típusú processzort tartalmaz. Ezzel szeretnénk volna, ha sikerül a mélytanulás során alkalmazott neurális hálózatokat az összes processzortípuson elosztottan tanítani és futtatni. Hosszas próbálkozás után sem sikerült ez ügyben eredményt elérni, azonban a munka során megismerkedtem más az Intel által fejlesztett és fejlesztés alatt álló eszközeivel.

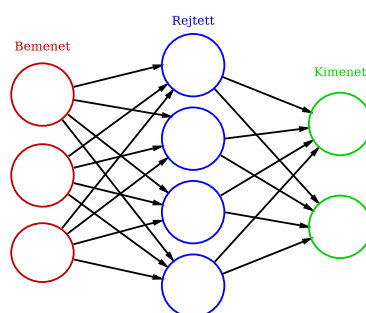
1. fejezet

Neurális hálózatok és a Deep Learning

1.1. A neurális hálózatok elmélete

Olyan számítási modellel, amelynek alapját az idegrendszer hálózata adja először Warren McCulloch és Walter Pitts 1943-ban foglalkozott az „A Logical Calculus of the Ideas Immanent in Nervous Activity” című publikációjukban. Később Donald Hebb tanulással kapcsolatos megfigyeléseivel elindultak a mesterséges neurális hálókkal kapcsolatos kísérletezések.[2][3] Kiderült, hogy ezek a struktúrák kiválóan alkalmasak osztályozási és regresszió számítási problémák megoldására, és a mai gépi tanulással kapcsolatos fejlesztések középpontjába került.

A mesterséges neurális hálózatok egy viszonylag egyszerű modellen alapulnak. Minden neuron a hozzá kapcsolódó neuronok ingereinek összessége alapján ingerli a többi neuront melyekhez ő kapcsolódik, ekképpen az ingerület egy irányba halad a kapcsolatok mentén. Hogy a hálózat áttekinthető legyen, rendezzük a neuronokat rétegekbe úgy, hogy egy réteg neuronjai az ingerületet a közvetlen felső réteg neuronjaitól kapja, és a válasz ingert a közvetlenül alatta lévő réteg neuronjainak továbbítja.



1.1. ábra. neurális hálózat réteges szerkezete forrás: en.wikipedia.org/wiki/Artificial_neural_network

A 1.1 ábrán a csúcsok jelentik a neuronokat és az élek a szinapszisok, melyeken az ingerület vándorol. Egy hálózat 3 nagyobb részre tagolódik: **a)** bemeneti réteg **b)** rejtett rétegek **c)** kimeneti réteg. A bemeneti réteg csúcsai legtöbbször az adatot reprezentáló konstansok jelentik,

tehát az egy egyszerű vektor. Egy neuronban két művelet történik: a bementek összegzése és egy aktiváló függvény kiértékelés. Az összegzést a felsőbb rétegből érkező jelekre elvégezzük:

$$s = \sum_i w_i x_i + b = \vec{w} \cdot \vec{x} + b$$

ahol x_i a felső réteg i -ik neuronjának kimenete, w_i az i -ik neuron szinapszisához tartozó súly, mellyel a szinapszis "erősségét" határozzuk meg. Az "s" összeghez hozzáadunk még egy b büntetőtagot, ami neuron aktiválási küszöbe lesz. Az aktivációs függvény adja a neurális hálózat kimenetét, melynek paramétere s . A neurális hálózatok fejlesztésekor sokféle függvényt találtak alkalmasnak aktivációs függvény gyanánt. Közös jellemzőjük, hogy inflexiós pontjuk $x = 0$ helyen van, illetve 0-ban nem deriválható függvények esetén a töréspont esik ide.

A szemléletesség kedvéért tekintsünk meg az egyrétegű perceptront, vagyis egy egyetlen rétegből álló neurális hálózatot k darab neuronnal. A bemenet legyen az $\vec{x} = (x_1, \dots, x_n)$ vektor. A szinapszisok súlyait a $W = \{w_{ij} : i = 1 \dots n, j = 1 \dots k\}$ mátrix (\vec{w}_i az i . bemeneti adatból kiinduló szinapszisokhoz tartozó súlyok vektora lesz), a neuronok büntetőtagját a $\vec{b} = (b_1, \dots, b_k)$ tartalmazza. Az aktivációs függvény f . A hálózat kimenetét, vagyis a $\vec{y} = (y_1, \dots, y_k)$ elemeit megkapjuk a következőképpen:

$$y_j = f(\vec{w}_j \cdot \vec{x} + b_j)$$

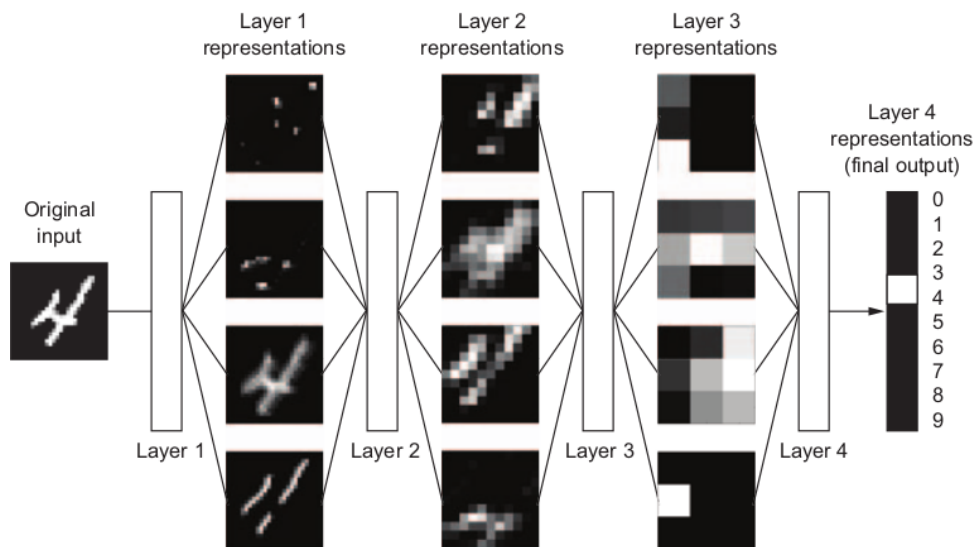
A fentiekből látszik, hogy a hálózat tervezésénél annak négy tulajdonságát kell meghatároznunk: 1. a rétegek és azok neuronjainak számát 2. a neuronok aktivációs függvényét (rétegenként egy típusú függvény az összes neuronra) 3. a szinapszisok súlyát ($W = (\vec{w}_1, \dots, \vec{w}_n)$) 4. a neuronok aktiválási küszöbét (\vec{b}). Minden réteghez külön W mátrixot és \vec{b} vektort kell meghatározni. **Az egyszerűség kedvéért néha az egy réteghez tartozó W -t és \vec{b} -t együttesen nevezik a réteg súlyainak.** Ez nagyon sok külön meghatározandó változót jelent, tehát csak az 1. és 2. tulajdonság meghatározása elvárható. Kell egy algoritmus, mellyel az egész hálózathoz tartozó paraméterek sokasága – vagyis minden réteg súlyának paraméterei – meghatározható.

1.2. Deep Learning

Chollet könyvében részletesen kifejti mit is értünk *Deep Learning* alatt, hogyan kapcsolódik a gépi tanulás fogalmához.[1]

A gépi tanulás egy teljesen más programozási paradigmát jelent, ugyanis a klasszikus programozás során a feldolgozandó adatokhoz a programozó adja az adat feldolgozásának szabályait, amit végig követve a gép kiszámítja a kívánt eredményt. Ezzel szemben a gépi tanulás során a programozó az adathoz a kívánt eredményt adja meg, amiből a gép felállítja a megoldáshoz vezető szabályokat.

A Deep Learning más néven a mély tanulás a gépi tanulás egy fajtája. Chollet szerint a név arra utal, hogy a kezdeti adaton több transzformációt végrehajtva egymás után egyre közelebb kerülünk egy olyan reprezentációhoz, ami megfelel a kívánalmainknak. Ezzel kontrasztban beszélhetünk sekély tanulásról, amikor kevés, egy vagy két transzformáció után kapjuk meg az adat megfelelő reprezentációját. A neurális hálózatok rétegeltsége adja a *mélységet* a gépi tanulásban. Eredeti elgondolás szerint minden egyes neuron-réteg egyre összetettebb tulajdonságokat ismer fel a bemeneti adatból. Valójában a rétegenkénti transzformációk egyre kisebb összetettségű hipotézis térbe visznek át, a reprezentáció egyre kevesebb – a felhasználó számára fölösleges – információt tartalmaz. Minél több réteg van a hálózatban, annál *mélyebb* a modell.



1.2. ábra. Írott szám hozzárendelése az ábrázolt számértékhez[1]

Itt kapcsolódik össze a neurális hálózat és a mély tanulás. Az 1.1 alfejezetben kifejtettem, hogy a neurális hálózat szinapszisainak paraméterezéseért felelős W mátrixok és a neuronok küszöbszintjének állítására szolgáló \vec{b} vektorok összes koordinátájának száma hatalmas lehet, – alkalmazástól függően több százezer, akár millió, egymástól független változóról beszélünk – tehát beállításukhoz valamilyen algoritmusra van szükség. Ezért a neurális hálózatok másik komponense, egy tanulási algoritmus, mely beállítja ezen paramétereket. Négy megközelítés létezik, amikor gépi tanulásról van szó.

Ellenőrzött tanulás során a neurális hálózatnak felcímkézett adatokat adunk meg, tehát olyan y értéket rendelünk az x mintákhoz, amelyet szeretnénk, hogy a hálózat produkáljon. Ezen $z = (x, y)$ összerendezések halmazát *tanítókészletnek* hívjuk. A hálózat leképezi az adatot a meghatározott reprezentációvá. A tanuló algoritmus ebből és a címkéből egy *veszteség függvény* kiszámításával meghatározza, hogy mekkora az eltérés, a valamilyen értelemben vett távolság a kapott és az elvárt eredmény között. Ez alapján frissíti a W mátrixot és \vec{b} vektorokat.

Ellenőrizetlen tanulás, mely során az adatokat nem címkézzük fel, hanem arra vagyunk kíváncsiak, hogy miféle összefüggések állnak fenn közöttük. Ezt a módszert adatbányászat során alkalmazzák.

Az *Önellenőrzött tanulás* hasonló az ellenőrzöttökhöz, azonban az adatok felcímkézését nem emberi erővel végezzük, hanem az adatokból állítjuk elő valamilyen heurisztikát felhasználva. Egyik alkalmazási területe az autóenkóderek tanítása.

A *Megerősítéssel tanulás* egy újfajta megközelítése a neurális hálózatok alkalmazásának. Ennél a metodikánál a hálózatot egy ágens alkalmazza, így a hálózat bemenete az ágens által megfigyelt környezet a kimenete pedig valamilyen cselekedet, beavatkozás és tanítás során az ágens igyekszik valamilyen környezetbeli értéket maximalizálni. Gyakori alkalmazás valamilyen játékot játszó ágens, ahol azt tanulja, adott helyzetekre milyen reakcióval tudja maximalizálni játékbeli pontszámát. Vizsgálódásomat az *ellenőrzött tanulásra* korlátoztam, így a továbbiakban ennek tükrében folytatom dolgozatomat.

1.3. Függvények, algoritmusok

Az alábbiakban szeretném megfogalmazni a neurális hálózatokban alkalmazott tipikus függvényeket és algoritmusokat.

1.3.1. Neuronok aktivációs függvényei

Mint korábban kifejtettem minden neuron kimenete egy függvény kiértékelése, melynek paramétere a bemenetek súlyozott összege. Ezt a függvényt hívjuk aktivációs függvénynek.

A szigmoid függvény Az utolsó, kimeneti neuronok rétegének aktivációs függvényeként alkalmazzuk, ahol a várt eredmény egyetlen valószínűségi érték. Ez bináris osztályozási problémák esetén alkalmazandó, tehát a program célja, hogy egy bemeneti adatról eldöntse, hogy az egy bizonyos kategóriába esik-e vagy sem, illetve erről mekkora "magabiztossággal" döntött.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1.1)$$

A softmax függvény A kimeneti réteg aktivációs függvénye. D dimenziójú x vektorok koordinátáit normalizálja, másként fogalmazva egy tetszőleges D elemű szám n -est azon D elemű n -esek halmazába képezi, melyek elemeinek összege 1. Így tehát a x koordinátái egy diszkrét

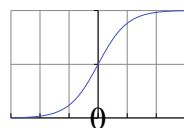
valószínűségi eloszlás értékkészlete.[28]

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{j=1}^D e^{x_j}}, \quad i = 1, \dots, D \quad (1.2)$$

Éppen ezért többosztályos problémáknál használatos. A hálózat egy mintára adott válasza annak a diszkrét valószínűségi változónak az eloszlása, mely a minta egy adott kategóriába tartozásának a valószínűségét adja meg.

ReLU Teljes néven *rectified linear unit* függvény vagy közismerten rámpafüggvény a rejtett neuron rétegek aktivációs értéke szokott lenni. Először 2011-ben mutatták be, hogy hatékonyabban taníthatók a neurális hálózatok, mintha csak szigmoid függvényt használnánk neuronok aktivációs függvényeként.[32]

$$f(x) = \max\{0, x\} \quad (1.3)$$



(a) Sigmoid függvény



(b) ReLu függvény

1.3. ábra. Aktivációs függvények grafikonja

1.3.2. Veszteség függvények

A neurális hálózat tanításához szükséges meghatározunk egy $c(\vec{y}, \vec{y}')$ veszteségfüggvényt, mely megadja a hálózat kimenetének eltérését az elvárt eredményhez képest adott súlyok mellett. Ezt az eltérést egy skalárértékként képezi le. Tanítás során ezt a függvényt kell minimalizálni.

Átlagos négyzetes hiba Többosztályos problémánál a neurális hálózat egy valószínűségi változó eloszlása az összes osztályon. \vec{y} vektor a hálózat válasza a \vec{x} bemenetre, \vec{y}' pedig a kívánt kimenet vektora – egy egységvektor, melynek 1 értékű koordinátája reprezentálja a megfelelő osztályt. Erre az esetre olyan veszteségfüggvényt alkalmazhatunk mely ekvivalens a $(\vec{x}, \vec{y}) \in Z$ tanítási készletből számított MSE átlagos négyzetes hibáinak átlagával.[3]

$$MSE(\vec{y}, \vec{y}') = \frac{1}{n} \sum_{i=1}^n (y_i - y'_i)^2$$

$$C(w, b) \equiv \frac{1}{2m} \sum_{j=1}^m MSE(\vec{y}_j, \vec{y}'_j)$$

A fenti egyenletekben n az kimeneti vektor dimenziója, m a tanító minták száma. Chollet szerint osztályozási problémánál érdemes kereszt-entrópiát veszteségfüggvényként alkalmazni az átlagos négyzetes hiba helyett.[1]

Kereszt-entrópia Osztályozási feladatoknál olyan módszert használhatunk a veszteség vagy hiba érték számításához, melynél az \vec{y} és \vec{y}' kereszt entrópiáját határozzuk meg. Ilyen jellegű feladatoknál az említett vektorok valószínűségi eloszlások.

$$H(y, y') = - \sum_{x \in \mathcal{X}} y(x) \log y'(x)$$

A mély tanuláshoz használt keretrendszereknél gyakran másként implementálják a kereszt-entrópiát kétosztályos és többosztályos esetre. Ezekre az implementációkra *bináris kereszt-entrópia* és *kategorikus kereszt-entrópia* néven hivatkoznak.

1.3.3. A hálózat tanító algoritmusai: az optimalizálók

Korábban említettem, hogy egy neurális hálózat tanításán azt az eljárást értjük, amely során a hálózat paramétereit változtatjuk. Neurális hálózatoknál gradienscsökkentésen alapuló technikákat alkalmazunk. A módszer megértéséhez, most tekintsünk úgy a költségre, mint a hálózat konfigurációjának jóságát értékelő függvényre.

$$C(\vec{w}_1, b_1, \dots, \vec{w}_L, b_L)$$

Ezen függvény képe egy úgymond *hibafelület*, melynek megkeressük minimum helyeit tanítás során. Gyakorlatban nem kivitelezhető a globális minimum meghatározása, és lokális minimumok keresését is iteratív módszerekkel célszerű végezni. Többváltozós függvények esetén a gradiens vektor ellentettje $(-\nabla f)$ segítségével meghatározható, hogy adott pontban – azaz megadott függvényparaméterek esetén – a paramétereket milyen mértékben változtassuk, hogy a függvény kimeneti értékét a legdrasztikusabb mértékben csökkentsük. Belátható, hogy C ekvivalens a c -k átlagával.

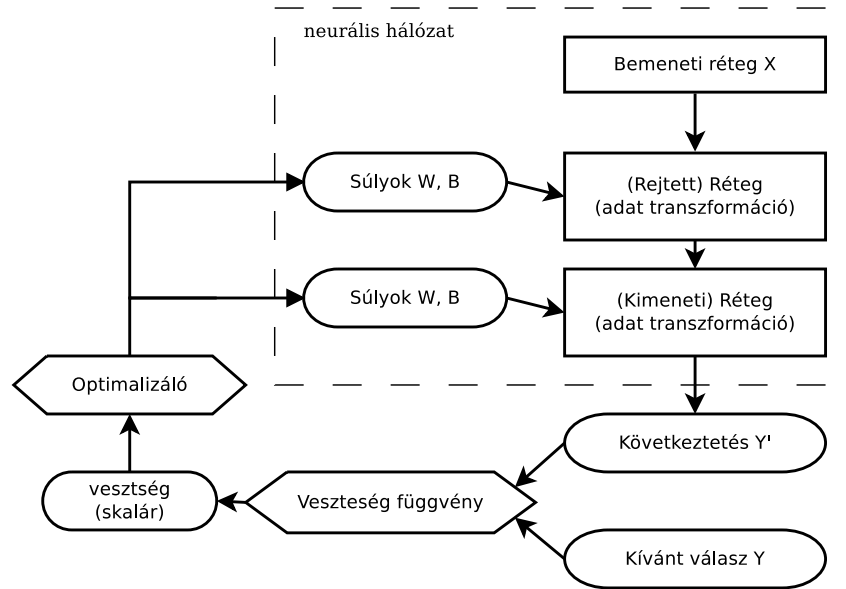
$$\frac{1}{n} \sum_{i=1}^n c(y_i, y'_i) \equiv C(\vec{w}_1, b_1, \dots, \vec{w}_L, b_L)$$

A gradienścökkentés naiv megközelítése pedig a következő:

$$\vec{W}_i = (\vec{w}_{i1}, b_{i1}, \dots, \vec{w}_{iL}, b_{iL})$$

$$\vec{W}_{i+1} = \vec{W}_i - \nabla C(\vec{W}_i)$$

Ahol \vec{W}_i a tanítás i -edik ciklusában az L rétegű hálózat összes paraméterének vektora.



1.4. ábra. A neurális hálózat tanításának folyamata

A tanítási folyamat iterációit *eposzoknak* (angolul: epoch) nevezi a szakirodalom. Minden eposzban az optimalizáló hatására a kimeneti hiba csökken, a hálózat *következtetése* egyre közelebb kerül az elvárthoz.

Sztochasztikus gradienścökkentés A naiv megközelítést alapul véve meghatározunk egy η tanítás sebességet (angolul: learning rate), mellyel azt befolyásoljuk, hogy egy eposzban a hálózat paramétereit mekkora mértékben változtatjuk. A tanítási sebesség az optimalizálók egy másik paramétere, tehát más eljárás esetén is alkalmazzuk valamilyen formában.

$$\vec{W}_{i+1} = \vec{W}_i - \eta \nabla C(\vec{W}_i)$$

Ezen paraméter változtatásánál fontolóra kell vennünk két tényezőt. Nagy tanítási sebesség gyorsabb gradiens csökkenést eredményez, ugyanakkor az elmozdulás a hibafelületen nagy kitérésekkel történik a minimumpont irányába. Túl kicsi tanítási sebességgel – azon túl, hogy megnövekszik az eposzok szám – ugyanakkor az optimalizálás "beragadhat" egy kedvezőtlen lokális minimum helyen.

négyzetes közép visszaterjesztése Konstans tanulási rátát nehezen lehet jól megválasztani, ezért olyan eljárásokat dolgoztak ki, ahol ezen érték adaptálódik a tanítási folyamat során, hogy kiküszöbölje az előző bekezdésben említett nagy kitéréssel történő konvergenciát és a "beragadást". Az egyik ilyen módszer az, hogy a tanulási rátát elosztjuk az előző ciklusokban számított gradiensek nagyságának négyzetes közepével.

$$v(w, y) = \gamma v(w, y - 1) + (1 - \gamma)(\nabla C(W_i))^2$$

Ahol γ az ún. felejtési tényező amivel szabályozhatjuk, hogy mekkora befolyása legyen az újabb gradienseknek.

A hálózat paramétereit a következőképpen frissítjük:

$$W_{i+1} = W_i - \frac{\eta}{\sqrt{v(w, y)}} \nabla C(W_i)$$

Tanítás során a veszteségfüggvény sztochasztikusan konvergál a nullához és ezzel együtt a hálózat egyre pontosabban következtet a tanító készlet adataiból. A tapasztalat azonban az, hogy ez nem feltétlenül igaz más, a tanítás során nem használt adatokra, sőt elérkezhet a hálózat egy olyan állapotba, amikor a használat közben rosszabbul teljesít, mint a tanítás végén. Ilyenkor a hálózat specializálódik a tanító adathalmazra. Ezt a jelenséget, amikor a hálózat jobban következtet a tanítás során használt adathalmazon *túlilleszkedésnek* nevezzük és *alul-illeszkedésnek* azt, mikor a hálózat nem elég pontos, más szóval túl kevés tanítási cikluson esett még át. Hogy a hálózat a hipotézis térből a megfelelő függvények – nem törekszünk az adatok egy tökéletes leképezésére – valamelyikét modellezze, szükséges a tanítást felügyelni, hogy az ne illeszkedjen se túl, se alul. A gyakorlatban ezért felügyeljük a tanítási folyamatot, és az optimalizálás minden ciklusában a hálózatnak egy külön validálásra szánt tanítókészletet adunk, és ezekre nem optimalizáljuk a hálózatot. Ezután össze tudjuk hasonlítani, hogy változik a hálózat hibája a tanító és validáló adatok mellett. Ennek tükrében abban a ciklusban ér el a hálózat a jól illesztett állapotot, amelyben a validációs veszteség érték eléri a minimumot.

Regresszió optimalizálás

Osztályozási problémákon túl a mesterséges neurális hálózatok felhasználhatóak regresszióanalízisre. Ilyen feladatra tervezett hálózat kimeneti rétegének neuronjai aktivációs függvény nélküliek, tehát lineárisan képezik le a bemenetüket. Erre a kézenfekvő magyarázat, hogy a várt eredmény egy széles skálán mozoghat és bármínemű aktivációs függvény alkalmazása lekorlátozza az kimeneti adatok halmazát. Természetesen a többi réteg esetében alkalmazunk aktivációs függvényeket. Tanítás során az átlagos négyzetes hibafüggvényt – amiről a 1.3.2 bekezdésben már kitértem – alkalmazzuk veszteségfüggvény gyanánt. A hálózat pontosságának

mérésére érdemes kiszámolni a hálózat következtetésének és az elvárt eredménynek átlagos abszolút eltérését, mint metrikát. A gyakorlatban a regressziószámítás sok változó skálán mozgó értékkel történik. A széles tartományon mozgó adatokon a hálózat nehezebben tanul, ezért a gyakorlatban érdemes a bemeneti adatokat normalizálni:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

ahol x a bemenő adathalmaz egy jellemzőjének egyetlen mintapontja, n az egy jellemzőhöz tartozó minták száma, \bar{x} pedig a minták átlaga. Regressziószámítás a statisztikai alkalmazáson túl objektum detektálási feladatnál is felmerül: az objektumhatároló keretekhez tartozó vektorokat is többváltozós regresszióanalízissel határozhatjuk meg.

2. fejezet

Mély tanuló alkalmazások fejlesztése

A mély tanulás, konkrétan a mesterséges neurális hálózatok gyors fejlődését és részben népszerűségét a hozzá készített programozási keretrendszereknek köszönheti, melyek java szabad hozzáférésű. Ezek a keretrendszerek arra hivatottak, hogy támogassák a neurális hálózatok fejlesztését új programozási módszertant adva. Már létező programozási nyelvekre épülnek, leginkább python-ra. Ezekben a keretrendszerekben egyszerűen implementálhatunk neurális hálózatokat úgy, hogy egyfajta nyelvi eszközkészletet adnak neurális hálózatok definiálására.

2.1. TensorFlow

A Google mély tanulással kapcsolatos projektjeihez saját eszközöket fejleszt ki. Ilyen a TensorFlow (továbbiakban: TF) keretrendszer is, mely elsősorban mély tanuláshoz lett készítve, de elég flexibilis ahhoz, hogy egyéb numerikus számítást programozzunk általa. Több hardver típust is támogat, és képes elosztott számítás végrehajtására is – ezalatt értem a többprocesszoros rendszereken és a számítógép-füzéreken történő futtatást. A TF az *adatfolyam programozási* paradigmát valósítja meg. Ez a programozási modell jól illeszkedik a párhuzamos rendszerekhez. Eszközei *adatfolyam gráfok* implementálását teszik lehetővé, amik a TF-ben távolról hasonlítanak a neurális hálózatok gráfjaira. Ezen felül leegyszerűsíti többdimenziós tömbök (*tenzorok*) kezelését. Több programozási nyelvhez is biztosított API, így eszközei hívhatóak Python, JavaScript, C++ és Java kódokból, bár a dokumentáció felhívja a figyelmet, hogy a Python API szolgáltatja a legteljesebb eszközkészletet és tekinthető a legstabilabbnak.[12] Jelenleg két, egymással részben kompatibilis stabil verziója van a TF-nak. A Google részletes dokumentációt biztosít a keretrendszerhez,[13] illetve első megjelenésekor kiadott fehér könyv tervezési oldalról mutatja be a TensorFlow-t.[11]

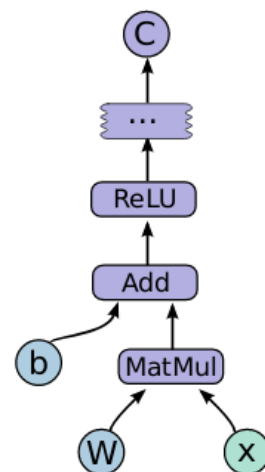
TensorFlowban a műveletek végrehajtásához műveleti csomópontokat kell definiálni, melyek egy-egy műveletet – melyek a `Operation` osztály példányai – valósítanak meg a számítási gráfban. Egy ilyen csomópontnak több bemenő, illetve kijövő éle lehet, melyek az át-

```

1  import tensorflow as tf
2  b = tf.Variable(tf.zeros([100]))
3  W = ↵
      tf.Variable(tf.random_uniform([784,100], -1,1))
4  x = tf.placeholder(name="x")
5  relu = tf.nn.relu(tf.matmul(W, x) + b)
6  C = [...]
7  s = tf.Session()
8  for step in xrange(0, 10):
9  input = ...construct 100-D input array...
10 result = s.run(C, feed_dict={x: input})

```

(a) TensorFlow kódrészlet minta



(b) A kódban definiált számítási gráf

2.1. ábra. Számítási gráfok TensorFlowban

adott részszeres számításokat szimbolizálják – a változók általánosan véve tenzorok a keretrendszeren belül. A *Session* egységbe zárja a számításokat – másként fogalmazva az *Operation* és a *Tensor* osztályok példányait –, így a műveleteket nem a gráfban implementációjában hajtódnak végre. Tehát a vezérlés külön van választva a program állapotától és a felhasználó az előbbi képes programozni. A TensorFlow új verziójában ez a megközelítés el lett rejtve, így például a `outputs = session.run(f(placeholder), feed_dict={placeholder: input})` kifejezés helyett a sokkal rövidebb `outputs = f(input)` használatos.

Több processzoros rendszeren az egyes számításokat úgy oszthatjuk ki az eszközök között, hogy a megfelelő csomópontokat a célhardverhez társítjuk. A kliens – az alkalmazás, melyben a számítási-gráf definiálva lett – megosztásokat tehet a kiszolgálónak (dokumentációban: *master*), hogy adott csomópontot – vagy egy komplex munkafolyamatot – az eszközök mely csoportján lehet futtatni. Ezeket a megkötéseket a `tf.device()` függvény paramétereként adhatjuk meg, és a megfelelő erőforrás-kezelővel tér vissza. A megkötések megadásának általános formája:

`/job:<JOB_NAME>/task:<TASK_INDEX>/device:<DEVICE_TYPE>:<DEVICE_INDEX>`
 ahol,

JOB_NAME egy betűvel kezdődő alfanumerikus karaktersorozat,

DEVICE_TYPE az eszköz típusa (például GPU vagy CPU),

TASK_INDEX egy nemnegatív egész szám, ami a `<JOB_NAME>` egy feladatának sorszáma,

DEVICE_INDEX egy nemnegatív szám, ami több hasonló típusú eszköz esetén a megfelelő eszköz sorszáma.

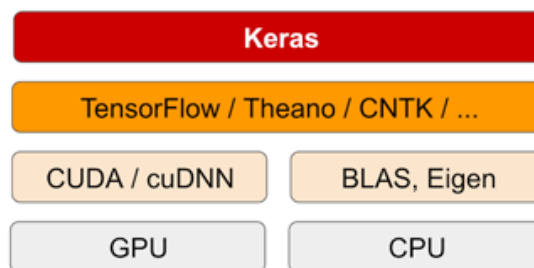
Példaképpen a képbetöltést és mátrix szorzat számítást az alábbi módon tudjuk szétosztani az alaplap és a grafikus processzor között:[26]

```
1 with tf.device("/device:CPU:0"):
2     img = tf.decode_jpeg(tf.read_file("img.jpg"))
3
4 with tf.device("/device:GPU:0"):
5     result = tf.matmul(weights, img)
```

A TensorFlow tehát egy nagyon rugalmas és jól skálázható keretrendszer, azonban a programozóra hárítja a neurális hálózat rétegeinek pontos implementációját, továbbá párhuzamos rendszerek esetén a számítási feladatok szétosztását. Születtek ezért olyan eszközök, melyek a TF „előtétjeként” igyekeznek megszabadítani a felhasználót a hosszadalmas és aprólékos fejlesztéstől.

2.2. Keras

A Keras egy python nyelven írt programkönyvtár, vagy ahogy önmagát hívja ”magas szintű neurális hálózat API”[14]. Fejlesztésébe François Chollet kezdett bele. Érdekessége, hogy más olyan keretrendszerekkel együttesen használható, amelyek a Keras-hoz hasonlóan magas absztrakciós szinten biztosítják a hálózatok implementálását, mint például a TensorFlow. Chollet könyve részletesen bemutatja a keretrendszer használatát, továbbá nyíltan hozzáférhető annak dokumentációja is.[1][14]



2.2. ábra. Keras szoftver rendszer¹

2.2.1. Neurális hálózat definiálása

Keras-ban egy neurális hálózatot *model*-nek hívunk (gyakran máshol is így neveznek egy konkrét neurális hálózatot). Egy *model* létrehozásához a `Keras.models` modulban definiált metódusokkal lehetséges. A keretrendszerben az adatokat tenzorokként kell reprezentálnunk, ezért

¹Forrás:[1]

érdeemes a *numpy*² nevű python csomaggal együtt használni. A keretrendszer tetszőleges alakú tenzorokat képes kezelni, tehát nincs megkötés arra vonatkozóan, hogy egy réteg bemenete vektor, mátrix vagy kiterjedtebb struktúrában – magasabb dimenziójú tenzorban – szerepeljen. A következő kódokban szeretném szemléltetni a Keras használatának módját.

A 2.2.1 kód egy két rétegből álló neurális hálózat definiálását mutatja be.

2.1. kód. Neurális hálózat rétegeinek definiálása

```
1 from keras import models
2 from keras import layers
3
4 network = models.Sequential()
5 network.add(layers.Dense(15, activation='relu', input_shape=(784,)))
6 network.add(layers.Dense(10, activation='softmax'))
```

A kód 4. sorában inicializáljuk a hálózatot, az utána következő sorokban új neuron réteget adunk a hálózathoz. Kerasban különböző előre definiált rétegekapsolatok vannak. Itt a `layers.Dense` osztály egy a szomszédos rétegekkel teljesen összekapcsolt réteget implementál – azaz egy neuron az előző réteg összes neuronjához kapcsolódik. A réteg bemenetként $28 \times 28 = 784$ elemű vektorokból álló mátrixot tud fogadni, másik dimenziója nem meghatározott. Ez a kötegetelt adatfeldolgozás szempontjából fontos, tehát a bemeneti vektorok folyamát egyik dimenziójában nem meghatározott méretű tenzorral definiáljuk. A réteg kimenete egy 15 dimenziós vektor. A következő réteg pedig 10 elemű tömböt állít el. Megfigyelhető, hogy nem adtuk meg a bemenet méretét, ugyanis a keretrendszerben ez implicit módon rendelődik a réteghez.

Mindkét réteghez tartozik egy *activation* paraméter, mely string típus kell legyen, és a réteg neuronjaihoz tartozó aktivációs függvényt adja meg. A példában a 'relu' és 'softmax' string a 1.3 fejezet (1.3) és (1.2) függvényére utal, azaz ezen bemeneti paraméterek esetén olyan `Dense` objektum jön létre, mely ilyen aktivációs függvényekkel rendelkező neuron réteget valósít meg.

A `Keras.layers` modulban a `Dense` osztályon kívül implementálva vannak konvolúciós, összefésülő, zaj, stb. rétegek, melyek a fenti módon tetszés szerint egymásra szervezhetőek a keretrendszerben, így szinte tetszőleges hálózat alakítható ki.

2.2.2. Hálózat betanítása és betanított hálózat futtatása

A megalkotott hálózatot a `network` objektum definiálja. Hogy tanítható legyen el kell látni egy optimalizálóval és egy veszteségfüggvénnyel, melyek együtt adják a hálózatot betanító algoritmust. A `compile()` metódus „összerakja” a neurális hálózatot a betanítóval, a `fit()` pedig elvégzi a tanítást, és *epoch*-ok számának megfelelő mé-

²lásd:<https://numpy.org/>

retű tömböket tartalmazó objektum referenciájával tér vissza, mely tömbökben össze vannak gyűjtve a veszteségfüggvény értékei és a felhasználói metrikák eposzonként.

2.2. kód. Hálózat betanítása

```
1 network.compile(optimizer = 'sgd',
2 loss = 'mean_squared_error',
3 metrics = ['acc'])
4
5 history = network.fit(train_datas,
6                       train_labels,
7                       epochs=20,
8                       batch_size=512)
```

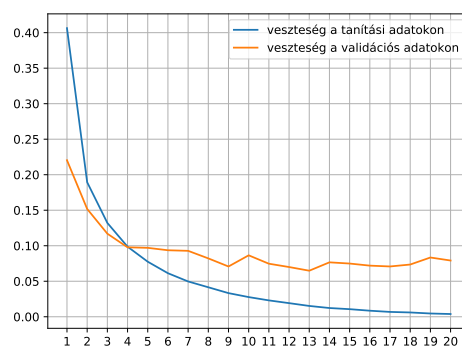
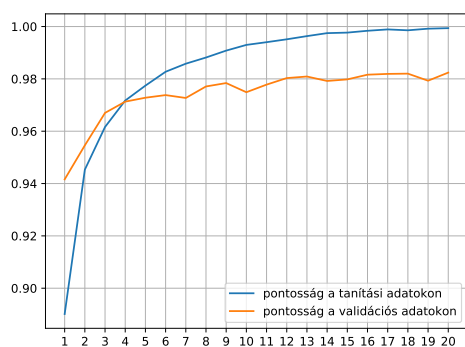
A `train_datas` és `train_labels` a tanítókészletet alkotó minták és azok címkéi, a várt kimeneti értékek (ezek elemszáma kötelezően meg kell egyezzen). A `batch_size` paraméter a *mini batch-ok* méretét adja, ami az egy *epoch*-ban egyszerre feldolgozandó adatokat jelenti. A betanítás végén a `network` egy *betanított*, a célfeladat megoldására felhasználható neurális hálózat lesz. Alkalmazásához a `predict()` metódus használatos, mely a paraméterként megadott bemeneti adatokhoz tartozó predikciókkal tér vissza.

2.3. kód. Az eszközre töltött hálózat futtatása az adatokon

```
1 result = network.predict(datas)
```

2.2.3. Hatékonyságvizsgálat

A tanítás felügyeletére a tanításhoz rendelkezésre álló adatok egy részét külön kell választanunk a tanítókészletől. Ez a készlet lesz a validálási készletünk, mellyel minden tanítási ciklus végén leellenőrizzük, hogy hogyan teljesít a hálózat olyan adaton, amit még „nem látott”. A validálást Keras-ban megtehetjük a tanítással egy lépésben, a `fit()` algoritmussal, ha a `validation_data` opcionális paraméterének megadunk egy `(data, label)` tuple típusú változót. Ekkor a metódus olyan referenciával tér vissza (2.2.2 kódrészletben ez a `history`), melyen keresztül a tanítási és validálás statisztikák is beolvashatóak. A 2.2.2 szakaszban említett metrikák a validálási folyamathoz is rögzítve lesznek.



2.3. ábra. Példa kimutatás a Keras által gyűjtött statisztikákról

3. fejezet

Deep Learning platformok

3.1. hybrid deep learning

A mai nagy teljesítményű párhuzamos számítási rendszerek igyekeznek kiaknázni az eltérő architektúrájú processzorok együttes alkalmazásában rejlő lehetőséget. Az ilyen heterogén rendszerek nemcsak nagyobb számítási kapacitással rendelkeznek, – összehasonlítva az homogén rendszerek egy processzormagra eső számítási teljesítményét – de jelentősen kisebb az energiafelhasználásuk. Rengeteg területen alkalmaznak ilyen rendszereket, többek között elektrontomográfiás képrekonstrukciónál[4] párhuzamos számítást igénylő tudományos alkalmazásokhoz[5] és az ún. *smart manufacturing* gyártástechnológia esetén.[6]

A párhuzamos adatfeldolgozás miatt ezek a rendszerek elsősorban grafikus processzor architektúrákat – konkrétan *általános célú grafikus processzorokat* (GPGPU) – alkalmaznak hagyományos CPU-k mellett.¹ Sok olyan rendszer épült korábban, ahol pár általános célú processzor mellett több GPGPU működött, azonban a legtöbb ilyen rendszer esetében a CPU-k pusztán az operációs rendszer futtatásában és számítási feladatok szétosztásában vett részt. Mély tanulás esetén előfordulhatnak olyan feladatok, amiket több CPU magra lehetne bízni.

A következő alfejezetekben szeretnénk bemutatni olyan szoftvereket és hardvereket, melyek lehetséges építőkövei egy ilyen mély tanulásra használható hibrid számítási rendszernek.

3.2. nGraph

Az 3.1 alfejezetben említett hibrid számítási rendszereken a neurális hálózatok működtetése nehézkes feladat. Szükséges a neurális hálózatból származtatható számítási gráf optimalizálása a különböző processzor architektúrákra, hogy képesek legyünk kihasználni a több architektúra adta sebességnövekedést. Többek között erre a célra az Intel belekezdett az nGraph projekt

¹Hibrid számítási rendszereknél előfordul még az FPGA-k és a DSP-k használata is.

fejlesztésébe. Az nGraph hivatalos dokumentációja részletesen leírja, hogy a fejlesztők milyen problémák megoldására tervezték ezt a szoftvert.[15].

Az nGraph az Intel által fejlesztett programkönyvtár és egy futtatási környezet/fordító készlet melyet mély tanuló projektekhez terveztek. Legszembetűnőbb tulajdonsága, hogy képes többféle hardver architektúrán futtatni és beépíthető számos keretrendszerbe. Ezek azok a jellemzők, amiket kerestünk témavezetőmmel, hogy mély tanulást tudjunk végeztetni hatékonyan a *HuSSar-on*. Ezen túlmenően a jövőbeli fejlesztések az iparban egyre jelentősebben igényelik, hogy a modern MI rendszerek skálázhatóak legyenek, mert egyrészt a neurális hálók egyre komplexebbé válnak, másrészt az általuk feldolgozott adatok mennyisége is rohamosan növekszik. Jelenleg két bevett gyakorlat van a mély tanulás felgyorsítására:

1. **Dedikált hardver tervezése a mély tanulással kapcsolatos számításokhoz** – Sok vállalkozás tervez *Alkalmazásspecifikus integrált áramköröket* (ASIC) neurális hálózatok betanítására és futtatására.
2. **Szoftver optimalizáció** – Programkönyvtárakat tartalmazó fejlesztési keretrendszerek fejlesztése, melyek képesek a hálózatokkal kapcsolatos számításokat több szálon, optimalizáltan futtatni. Az nGraph, mint fordító is egy ilyen megoldás.

3.2.1. Motiváció

A mai legmodernebb szoftveres megoldás mély tanulásra az, ha integrálunk kernel könyvtárakat² mély tanulási keretrendszerekbe. Ilyen integráció lehet például ha a Tensorflow keretrendszer alatt az Nvidia CuDNN könyvtárát használjuk. A kernel könyvtárak egy adott célarchitektúrára optimalizált kernelekből és egyéb műveleti szintű optimalizálásokból állnak, ezekkel érve el teljesítménynövekedést. Azonban a kernel könyvtáraknak három fő problémája van:

1. Nem biztosítanak gráf szintű optimalizálást
2. A keretrendszerek integrációja a kernel könyvtárakkal nem skálázható
3. A szükséges lefordított kernel könyvtárak száma növekszik, ahogy új processzorok jelennek meg

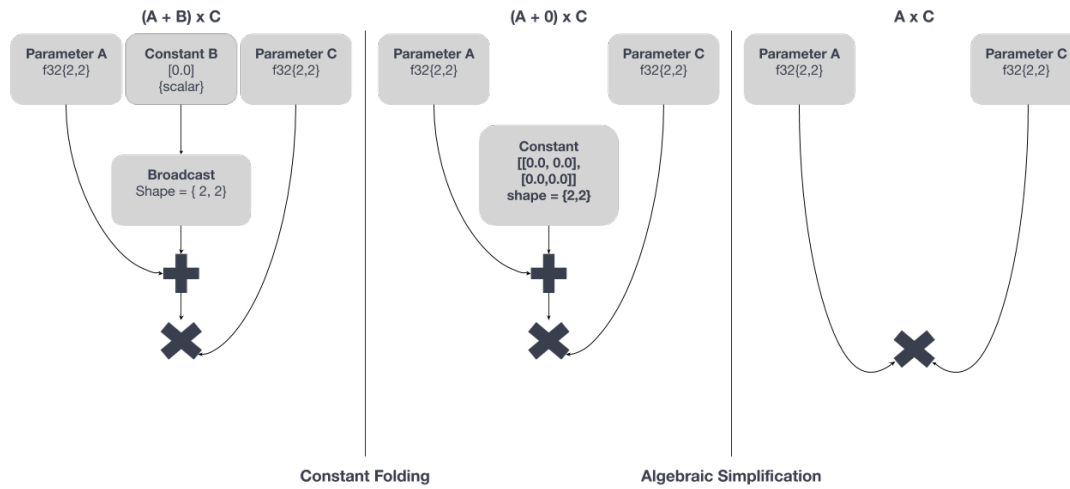
Az nGraph fordító megoldás az első két problémára, és a *PlaidML*-el ötvözve kiküszöbölhető a harmadik probléma.

3.2.2. Gráf szintű optimalizálás

Az nGraph dokumentációjában áll egy példa arra vonatkozóan, hogyan lehet, hogy egy mély tanulási keretrendszer integrálva egy kernel könyvtárral a számításokat optimálisan végzi, mégis a

²Programkönyvtár, mely neurális hálókkal kapcsolatos elemi, *mag* függvényeket tartalmaz

számítási gráf a csúcsait alkotó műveletek szempontjából mégsem optimális. A 3.1 ábra bemu-



3.1. ábra. két optimalizálás módszer: konstans összehajtás és algebrai egyszerűsítés a gráfon. forrás: [15]

tatja, hogyan egyszerűsíthetünk az A, B és C tenzorokat feldolgozó, $(A + B) * C$ tenzorműveletet végrehajtó számítási gráfon. Fordítási időben megállapítható, hogy B egy skálár konstans, így a *konstans összehajtásnak*[29] nevezett optimalizálás elvégezhető, és a 2 dimenziós vektorra való kiterjesztés művelete elhagyható (helyette inkább közvetlenül létrehozunk egy 2×2 tenzort). Ebben a példában $B = 0$ skálár volt, így a belőle létrejött tenzor egy nullmátrix, így az semleges a kifejezés kiértékelésének szempontjából. *Algebrai egyszerűsítést* végezve az $(A + 0) * C$ leegyszerűsíthető az $A * C$ kifejezésre, így összesen két csúccsal csökkentettük a gráfunkat. Ez az optimalizáció tehát a számítási gráf szintjén lett elvégezve. Így belátható, hogy a mély tanulási keretrendszerbe integrált kernel programkönyvtárak nem optimális futást végeznek, hiába a műveletek szintjén elért optimalizálás.

3.2.3. Skálázható keretrendszer integráció

Ahogy gyarapodik a mély tanuláshoz használható gyorsítókártya architektúrák és keretrendszerek száma, a meglévő mély tanulást alkalmazó fejlesztési platformok bővítése egyre több munkát igényel és egyre nő a hibák megjelenésének a valószínűsége. Az integráció kapható készen, szakértő fejlesztőcsapatoknak kell implementálnia. Minden új keretrendszert manuálisan kell integrálni a meglévő hardverek kernel könyvtárával és minden újonnan megjelenő hardvercsalád meghajtó programkönyvtárát be kell integrálni egyesével a meglévő keretrendszerekbe. Ez a munka önmagában is hatalmasra tud nőni, de egy sok eszközből álló összeállítás nagyon törekeny és költséges a fenntartása. Az nGraph úgy oldja meg ezt a problémát, hogy ún. *hida-*

kat alkalmaz, amikkel integrálható valamelyik mély tanulási keretrendszerbe. A híd megkapja a keretrendszerben megalkotott számítási gráfot vagy ahhoz hasonló struktúrát és átalakítja egy ún. *közbenső reprezentációvá*³. Ezzel kaptunk egy egységes, platformfüggetlen számítási gráfot, így nem kell egy új programkönyvtárat beintegrálni minden egyes meglévő keretrendszer alá, elegendő csak az, hogy az nGraph-ban, mint programkönyvtárban implementált *primitív műveleteket* támogassa az új programkönyvtár.

3.2.4. Növekvő kernel szám

Egy kernel könyvtár integrálása egyszerre több mély tanulási keretrendszerrel nehéz feladat és egyre komplexebbé válik, ahogy növekszik az optimális teljesítményhez szükséges kernelek száma. Régen a mély tanulással kapcsolatos kutatások egy kis számú *primitív* számítást használtak, mint a konvolúció, általános mátrixszorzás, stb. Az MI kutatás előrehaladtával és az ipari mély tanuló alkalmazások továbbfejlesztésével, a szükséges kernelek száma (k) exponenciálisan nő. Ez a szám a processzor architektúrák számán (h), adattípusokon (t), műveleteken (p) és az egyes paraméterek számosságán (m) alapul ($k = h \times t \times m \times p$).

Hardver	Művelet	Adattípus	Paraméterek
CPU	konvolúció	16 bites lebegőpontos	NCHW vagy NHWC
GPU	MatMul	32 bites lebegőpontos	2D, 3D és 4D tenzorok
FPGA	Normalizálás	8 bites egész	...
...

3.1. táblázat. Néhány példa, tényezőnként hányféle esetre kell külön fordítani kernelt könyvtárt

Ezen probléma megoldásához jön képbe a PlaidML. Ez egy *tenzor fordító*⁴, mely azt célozza, hogy képes legyen neurális hálózatokat tanítani és futtatni bármilyen típusú hardveren. Más szavakkal segíti a magas szintű keretrendszerek (Keras, ONNX, nGraph) integrálni olyan eszközökkel, melyekhez nincs meg a szükséges támogatás vagy a meglévő szoftverkészlet hozzájuk szigorúan lincszelt.[25][16]

Az nGraph tehát integrálható a PlaidML-el. Elsősorban az nGraph a platform független IR-rel igyekszik orvolsoni a skálázható backend-el kapcsolatos kihívást. A PlaidML ezt támogatja azzal, hogy képes az IR-ből származó gráfokból LLVM, OpenCL, OpenGL, CUDA és Metal kódot generálni melyek a megfelelő hardveren futtathatóak. Így egy magas szintű keretrendszerben írt neurális háló lefordul Intel és AMD processzorokon valamint grafikus processzorokon, az nVidia processzorain, továbbá az Apple cég által feljlesztett eszközökön. Az nGraph gráf szintű optimalizációját ráadásul kiegészíti automatikusan a PlaidML alacsonyabb szinten, ezzel teljesítmény növekedést érve el.

³IR: Intermediate Representation

⁴Olyan fordító, melynek nyelve arra lett fejlesztve, hogy főleg tenzorműveleteket igénylő algoritmusokat tudjunk hatékonyan programozni[31]

Összegzésül tehát az nGraph feldarabolja a neurális hálózathoz tartozó számítási gráfot processzor architektúrájának megfelelően, majd ezen gráfokat a PlaidML lefordítja a megfelelő kódokra, melyeket aztán a célprocesszorokra lefordítunk és futtatunk.

3.3. Grafikus processzorok a mély tanulásban

Változatos megoldások születnek a mély tanulás gyorsítására a hardvergyártók részéről. Kezdetben a nagy teljesítményű általános célú grafikus processzorokat (GPGPU) alkalmazták szinte kizárólag a neurális hálózatokkal kapcsolatos számításokhoz, mert ezek az architektúrák kifejezetten hatékonyan tudnak vektor és mátrixműveleteket végezni – tulajdonképpen pont erre tervezték őket. Azonban ahogy elkezdett terjedni a mély tanulás olyan területeken is, ahol számít az alacsony hűtésigény, kiegyensúlyozott energiafogyasztás vagy a hordozhatóság, új architektúrák is kezdtek megjelenni. A következőkben bemutatok pár – a mély tanulásban használt – hardvert és két szoftveres környezetet, melyek jól körvonalazzák, milyen irányba fog fejlődni a gépi tanulás technológiája.

3.3.1. Nvidia által fejlesztett gyorsítókártyák

Abban az időben mikor a mély tanulás bekerült a köztudatba a grafikus processzor gyártóknak olyan megoldásaik voltak, amikkel általános számítási feladatok programozását tette lehetővé a termékeiken. Ilyen az Nvidia CUDA (*Compute Unified Device Architecture*) is, ami egy párhuzamos számítási platform és alkalmazásprogramozási interfész. Maga a platform egy olyan szoftverréteg, mely közvetlen hozzáférést enged a GPU utasításkészletéhez.[34] Az API pedig olyan programozási eszközöket nyújt, amivel hatékonyan lehet CUDA képes GPU-n futó programot írni. A cég által közzé tett dokumentációk alapján nagy vonalakban bemutatnám a neurális hálózatok fejlesztéshez használható eszközeit.[17] Ahhoz, hogy neurális hálózatokat futtassunk CUDA platform szükséges a CuDNN (*CUDA Deep Neural Network Library*) könyvtárcsomag, mely a nélkülözhetetlen elemi függvényeket implementálja:

1. Konvolúció
2. Adatok összevonása (pooling)
3. neuron aktivációs függvények
4. tenzor transzformáció

A CuDNN használatakor szükséges egy kezelőt (handler) inicializálni a `cudnnCreate()` metódus hívásával. A metódus hívásával inicializálódik a programkönyvtár is, elvégzi a gazdarendszer és a gyorsítókártya erőforrásainak allokációját. Továbbá a visszaadott eszközközkezelő mutat

az inicializált gyorsítóra, így használható a későbbi függvényhívásoknál, melyeket az adott eszközön akarunk végrehajtani.

A vállalat GPU-jainak magjai közül CUDA magoknak hívja azokat a speciális számítási egységeket, melyek egyedileg programozhatóak a platformon (tehát CUDA alatt nem utilizálható az összes mag). A mai legmodernebb gyorsítókártyákat az NVIDIA Turingnak nevezett architektúrájú processzorokkal szereli fel. Több fejlesztési irányt követnek a gyártók: Standard asztali és mobil grafikus processzorok, High-end asztali GPU (Nvidiánál: Quadro és TitanRTX[7]), szerver oldali GPU (Tesla T4 [8]), valamint kifejezetten HPC-re szánt GPU (Tesla v100[9])

TITAN RTX Igen nagy teljesítményű grafikus kártya, mely a cég *turing* mikroarchitektúrájú TU102 processzorral van felszerelve, ami 4608 CUDA processzormagot, valamint 576 tenzor magot tartalmaz és alapjáraton 1350 MHz órajelen üzemel. A kártyát felszerelték 24 GB GDDR6 memóriával, melynek sávszélessége 672 GB/s⁵. Elméleti számítási kapacitása 16.31 TFLOP/s (billió lebegőpontos művelet másodpercenként) egyszeres pontosságú lebegőpontos műveletek esetén. Ez az érték félszeres pontosság esetén – amit a mai neurális hálózatoknál gyakran használnak – dupla ennyi. Ez egy nagy energiafogyasztású kártya, a gyártó 600 W teljesítmény leadására képes tápegységet ajánl az eszköz mellé.

Tesla T4 Ezt a grafikus kártyát kifejezetten adatközpontok komponenseként szánták, ezért nem is szolgáltat videó jelet. Számítási teljesítményes is bizonyos szempontból szerényebb a TITAN RTX-nél. Szintén turing architektúrájú TU104 processzora 2560 CUDA magot tartalmaz. A szerényebb felszereltségnek ugyancsak az adatközponti felhasználása miatt van jelentősége, hogy így kisebb áramfelvétel és hőleadás mellett tudjon üzemelni – 350 W ajánlott leadási teljesítmény és összehasonlításképp a TITAN RTX *Thermal Design Pointja* 280 W, míg ennek a kártyának csupán 70 W. Egyszeres pontosságú lebegőpontos műveleteket alkalmazó számítások elméleti sebessége 8,141 TFLOP/s, viszont a félszeres pontosságnál nyolcszor gyorsabb, azaz ekkor 65,13 TFLOP/s az elméleti sebessége. Ebből látszik, hogy ez a kártya kifejezetten mély tanuláshoz és olyan alkalmazásokhoz lett optimalizálva, ahol félszeres pontosságú lebegőpontos számításokat kell végezni túlnyomó részt. A teljesség kedvéért megemlítem, hogy a kártya 16 GB szintén GDDR6 típusú memóriával lett felszerelve, melynek sávszélessége 320 GB/s (256 bites buszszélesség).

Tesla V100 A Tesla V100 a T4 elődje. Ennél a kártyánál a minél nagyobb teljesítmény elérésére törekedtek. HBM2 memóriamodulokkal szerelték 16 és 32 GB kapacitással. A memória sávszélessége 900 GB/s 4096 bites adatbuszon. 5120 CUDA magot tartalmaz a Volta archi-

⁵a memóriaegységek 384 bit széles adatbuszon továbbítanak adatot 14 Gb/s-os bitrátával

tektúrájú GPU, melyet a kártyára szereltek. A processzor maximálisan 1455 MHz órajel mellett üzemel. Kétféle kártyakialakítással árulják: PCI express interfésszel ellátott kártya az adatárházak számára (32 GB/s sávszélesség a kártya és az alaplap között), és a cég saját fejlesztésű *NVLink* csatolófelülettel szerelt kártya HPC-hez. Utóbbi interfész sávszélessége 300 GB/s. A két kártya között jelentős eltérés a számítási kapacitás tekintetében, ugyanis míg a PCIe kártya 14 TFLOP/s számítási kapacitást nyújt, az NVLink változat teljesítménye 31,33 TFLOP/s félszeres pontosságú lebegőpontos számításoknál. Ez noha kevesebb a T4 teljesítményénél, tenzorműveletek számításánál 125 TFLOP/s teljesítményt képes produkálni.[10]

3.3.2. Intel grafikus processzorok

Az Intel legtöbb CPU-ját integrált grafikus processzorral szereli, szinte csak a munkaállomásokra szánt Xeon család processzorai nem tartalmazznak GPU-t (kivétel a Xeon E3 széria).[30] Az Intel – akárcsak a többi nagy hardvergyártó – ki akarja venni a részét a deep learning technológiák fejlesztésében és olyan szoftvereket fejlesztett ki – új architektúrák mellett –, melyekkel mély tanulásra hasznosítható a szinte mindenütt jelenlévő GPU-ik számítási kapacitása. Ezek közül a legfontosabb az OpenCL alapú *c/DNN* programkönyvtár, mely lehetőség ad, hogy az cég processzoraiba integrált grafikus processzorokat mély tanulásra használhassuk. A másik a korábban bemutatott nGraph, mely több másik architektúra mellett Intel GPU-kat is képes kezelni.

Intel Xe grafikus processzorarchitektúra Nem olyan rég az Intel bejelentette, hogy készül egy új grafikus processzor családot, az Xe szériát piacra dobni. Eltérően az eddig általuk készített GPU szériáktól az Xe lapkák diszkrét – tehát nem CPU-ba integrált – grafikus processzorok lesznek. Előzetes bejelentés alapján 2020-ra kerül forgalomba, de jelenleg kevés információ ad közre a gyártó a processzorról és az azt tartalmazó videokártyáról.[35] Mindenesetre az már biztos, hogy elérkezett a tesztelés fázisához az első grafikus kártyája a cégnek, melyet DG1-nek neveztek el.[36] Ezekkel a chippekkel a vállalat a nagyvállalati felhasználókat célozza meg. Adatközpontok és szuperszámítógépek alkotórészének szánják, hangsúlyt fektetve a mesterséges neurális hálózatokban való felhasználásra.

Dr. Cutress anandtech.com oldalon megjelent cikkében jobban részletezi, mi várható az Xe szériától.[37] Egyrészt az Intel az új architektúrával az összes olyan technológiai ágazatot le akarja fedni, melyek grafikus processzorokra alapoznak. Ezt úgy kivitelezzi, hogy az Xe – mely egy architektúrát takar – több mikro- és szub-architektúra alkotja. Hogy szuperszámítógépes környezetben is megállja a helyét ezek a chippek – vagy a széria néhány darabja – kezelni tudja a dupla pontosságú lebegőpontos számításokat. A kifejezetten HPC rendszerekre szánt GPU kódneve *Ponte Vecchio*.

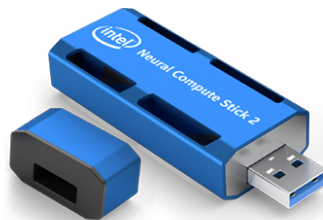
Ezzel az új termékcsaláddal a vállalat komoly versenytársa lehet majd az Nvidia-nak, mert emellett kifejleszt a CUDA-hoz hasonló *OneAPI* nevű programozási modellt, mellyel egységesen lehet számításokat végezni Intel CPU-in, GPU-in és FPGA-in.[38]

3.4. Myriad X és az Intel Neural Computer Stick 2

Ma az iparban mély tanulással működő modern alkalmazások nagy méretű és teljesítményű számítógépeket igényelnek. Erre az igényre válaszul felhő szolgáltatást nyújtó vállalatok külön platformot építettek ki. Az ilyen kliens-szerver architektúrájú gépi tanulás hatékony lehet, azonban a valós-idejű alkalmazásoknál és hordozható eszközökben történő felhasználásnál komoly hátrányt jelenthet a hálózati késleltetés. Hogy ilyen területeken is alkalmazható legyen a mély tanulás, hardvergyártók olyan kis energiafogyasztású *alkalmazás-specifikus integrált áramkörök* fejlesztésére törekedtek, melyek képesek neurális hálózatokat hatékonyan futtatni. Ilyen az Intel által kifejlesztett Myriad X lapka is, melyet főleg *gépi látáshoz* terveztek.

Maga a Myriad egy Videójel feldolgozó processzor széria, melynek 3. generációja a Myriad X architektúrája kiegészült egy úgynevezett *Neural Compute Engine-el*, mely a betanított hálózatok futtatását hivatott gyorsítani. Ezzel együtt a lapka 16 darab 128-bit VLIW processzormagok, ún. SHAVE magot és 2,5MB 400GB/s sávszélességű memóriát tartalmaz. Számításokat 8 bites egészek és 16 bites lebegőpontos változókon tud végezni.⁶

A Myriad X kifejlesztésén túlmenően a vállalat tervezett egy gyorsítókártyát is ehhez a processzorhoz, melyet Neural Compute Stick (NCS) névvel forgalmaz. Az Intel kiadta a második verzióját az eszköznek Neural Compute Stick 2 (NCS2) névvel és az első verziót nem forgalmaz már. Ezért a továbbiakban a leírtakat az NCS2-re kell érteni. Az eszköz USB 3.0 -ás interfésszel rendelkezik, így sok gazdagéppel használható, többek között Raspberry Pi-vel is. Meghajtására az Intel OpenVINO eszközkészlete használható, melyet gépi látáson alapuló alkalmazások fejlesztésére készített a vállalat. Lehetőség van egy platformon több NCS2 modul együttes használatára, így egyszerűen növelhető a rendszer teljesítménye.



forrás: software.intel.com/en-us/neural-compute-stick

3.2. ábra. Neural Compute Stick 2 gyorsító

⁶forrás: <https://www.movidius.com/myriadx>

3.4.1. Az NCS2 használata OpenVINO keretrendszerben

Az OpenVINO eszközkészletéről részletes és nyílt hozzáférésű dokumentációt ad az Intel. [18] A felépített neurális hálózatokra a modell szó használata általános, így a dokumentáció is így nevezi azokat.

Az OpenVINO eszközkészlet alkalmas arra, hogy gyorsan készítsünk olyan alkalmazásokat, melyek kamera szenzor adatainak valós idejű feldolgozását végzik. Képesek vagyunk segítségével az NCS2-n kívül más Intel gyártmányú hardverrel dolgozni, például Intel CPU-n és integrált GPU-n. Továbbá népszerű keretrendszereket támogat, mint a TensorFlow, Caffe, MXNet vagy az ONNX. Az eszközkészlet komponensei közül az alábbiakat emelném ki:

Inference Engine C++ és Python API, aminek segítségével különböző gyorsítókártyák megszólíthatók egységes módon.

Model Optimizer parancssoros alkalmazás modelleket definiáló állományok átkonvertálására.

Ahhoz, hogy a különböző keretrendszerekben megalkotott modelleket egységesen tudja kezelni az Inference Engine, szükség van azok átkonvertálására egy platformfüggetlen struktúrába.

OpenCV az OpenCV közösségi fejlesztésű verziójának Intel hardverekre fordított változata.

Az OpenVINO-val való munka folyamatának három fő lépése van: Betanított Neurális hálózat beszerzése vagy implementálása; a hálózat átkonvertálása a Model Optimizer-rel; Inference Engine-re épülő program implementálása, mely beolvassa az átkonvertált állományokat és átadja a neurális hálózat modelljét a célhardvernek;

Modell átkonvertálása

Mint már korábban említettem a platform különböző neurális hálózatok implementálására kifejlesztett keretrendszereket támogat. Ezek a keretrendszerek képesek a definiált és betanított hálózatokat valamilyen állomány formátumban elmenteni. Hogy ezeket az állományokat minden gyorsító eszközön egységesen legyen képes kezelni az *Inference Engine* szükséges az állományokat átkonvertálni a *Model Optimizer* segítségével. Ez egy Python-ban írt kereszt-platformos parancssoros alkalmazás, mely a beolvasott modell állományokban tárolt hálózatot elmenteni egy *közbenső reprezentációvá*⁷, melyre a továbbiakban *IR-ként* fogok hivatkozni. Az átkonvertálás során a Model Optimizer analizálja a modellt⁸ és olyan korrekciókat hajt végre rajta, hogy optimális módon hajtsódjon végre az egyes gyorsítókon.

Egy TensorFlow-ban implementált neurális hálózat átkonvertáláshoz az alábbi lépéseket kell végrehajtani:

⁷Intermediate Representation (IR)

⁸betanított, kvázi használatra kész neurális hálózat

Model Optimizer konfigurációja Az alkalmazás előre elkészített szkriptekkel települ a különböző keretrendszerekhez történő konfigurációhoz. A szkriptek a <maindirectory>/deployment_tools/model_optimizer/install_prerequisites könyvtárban találhatóak. TensorFlow-val való munkához az install_prerequisites_tf.sh – Windows operációs rendszeren ugyanezen fájlneven .bat kiterjesztéssel – konfigurációs szkriptet kell lefuttatni.

TensorFlow modell befagyasztása Ez egy keretrendszer specifikus lépés. Amikor Pythonban definiáljuk a hálózatot, rendszerint elmentjük azt egy *inference graph* fájlban. Ez általában olyan módon történik, hogy a hálózat tanítható marad, tehát annak paraméterei, mint változók tárolva vannak az állományban. A Model Optimizer csak úgy tud dolgozni ezen az állományon, ha az úgymond be van *fagyasztva*. A keretrendszerben befagyasztott modell kimentésének műveletét a következő kódrészlet mutatja:

```
1 import tensorflow as tf
2 from tensorflow.python.framework import graph_io
3
4 frozen = tf.graph_util.convert_variables_to_constants(sess, ↵
    sess.graph_def, ["output_node"])
5 graph_io.write_graph(frozen, './', 'model.pb', as_text=False)
```

ahol:

`sess` egy TensorFlow Session példánynak a referenciaváltozója

`["output_node"]` a kimeneti csomópontok nevének listája. Egy befagyasztott hálózat csak azokat a csomópontokat fogja tartalmazni a az eredeti `sess.graph_def`;

`'model.pb'` a létrehozandó modell fájl neve;

`as_text` boolean típusú változóval megadható, hogy a `write_graph` metódus a bináris vagy karakteres formátumban írja ki a gráfot.

Modell átkonvertálása Ennél a pontnál használjuk a Model Optimizer-t, azaz ekkor történik meg a konverzió. Az alkalmazás a <főskönyvtár>/deployment_tools/model_optimizer könyvtárba települ. Hogy a `model.pb` állományunkat átkonvertáljuk a következő parancsot kell kiadni:

```
python3 mo_tf.py --input_model model.pb
```

A kimenet maga az IR, két állomány: egy xml kiterjesztésű fájl, amiben a hálózat topológiája van definiálva és egy bináris fájl, ami a hálózat paramétereit – súlyait és küszöbértékeit – tartalmazza. A program további argumentumaival specializálhatjuk a konvertálás folyamatát.

Inference Engine

Az OpenVINO eszközkészlet tartalmaz egy futtatókörnyezetet az *Inference Engine*-t, egy programkönyvtár csomagot melynek központi része a `libinference_engine.so`⁹ könyvtár. A könyvtár csomag C++ nyelven íródott, melyhez létezik C++ és python alapú API. Az utóbbihoz készült API szerényebb funkcionalitással bír. Feladatai az IR-ben tárolt modell beolvasása, az következtetés (inference) optimalizálása a célhardveren történő végrehajtáshoz és annak futtatása. A hardvercsoportokkal történő kommunikációs eljárások külön programkönyvtárakba – ahogy a dokumentáció hivatkozik rájuk *pluginokba* – lettek szervezve.

GPU plugin (libclDNNPlugin.so)	Intel HD Graphics és Intel Iris Graphics
CPU plugin (libMKLDNNPlugin.so)	Intel Xeon AVX2 és AVX512 utasításkészlettel, Intel Core AVX2 utasításkészlettel, Intel Atom SSE utasítással
FPGA plugin (libdliaPlugin.so)	Intel Vision Accelerator eszköz Intel Arria 10 FPGA-val (Speed Grade 1 és 2), Intel Programmable Acceleration kártya Intel Arria 10 GX FPGA-val
VPU plugin (libmyriadPlugin.so)	NCS és NCS2, Intel Vision Accelerator eszközök Myraid processzorral
GNA plugin (libGNAPLugin.so)	Intel Speech Enabling Developer Kit, Amazon Alexa Premium Far-Field Developer Kit, Intel Pentium Silver J5005, Intel Celeron J4005, Intel Core i3-8121U
Multi-Device plugin (libMultiDevicePlugin.so)	egy modell több eszközön történő parallel inferenciájára
Heterogenous plugin (libHeteroPlugin.so)	modell automatikus szétosztása több eszközre (akkor szükséges, ha néhány hálózati réteg típust nem minden eszköz támogat)

3.2. táblázat. Plugin típusok és támogatott hardverek

A fejlesztés során is használt python API-n keresztül mutatom be az Inference Engine-t. Az interfészt a következő osztályok szolgáltatják:

`ExecutableNetwork` osztály reprezentál egy modell példányt, amit betöltöttük a pluginba és készen áll az inferenciára;

`IECore` egy Inference Engine objektum. Segítségével egy egységes interfészen keresztül kezelhetjük a pluginokat;

⁹Windows operációs rendszeren természetesen .dll kiterjesztésű és a fájlnev nem tartalmazza a "lib" karakter-sort és ez igaz a többi programkönyvtár elnevezésére is

INetLayer tartalmazza a főbb információkat a modell rétegeiről, továbbá eszköz a rétegek néhány paraméterének módosításához;

INetwork minden információt tartalmaz az IR fájlokból beolvasott modellről, és annak paramétereinek módosítását teszi lehetővé;

IEPlugin a plugin-ok fő interfésze. Célja, hogy inicializálja és bekonfigurálja a plugin-t;

InferRequest egy interfészt kínál a következtetés kérésére az **ExecutableNetwork** osztályon keresztül, továbbá kezeli a predikció végrehajtását és a kapott adatokat;

InputInfo a modell bemeneti rétegének tulajdonságait tartalmazza: **layout** karaktersorozat, a (bemeneti) tenzor elrendezése¹⁰, **shape** szám n-es a tenzor alakja, **precision** az adatok számbábrázolás szerinti típusa;

LayersStatsMap a Python **dict** osztályának leszármaztatása, mely újraimplementálja a **update()** metódust, hogy módosítható legyen a rétegek finomhangolásának statisztikái;

LayerStats a rétegek finomhangolási statisztikáinak konténere;

OutputInfo a modell bemeneti rétegének tulajdonságait tartalmazza hasonlóan az **InputInfo** osztályhoz;

Egy elemi alkalmazás implementációjához, mely inferenciát végez az NCS2 eszközön szükség van az **INetwork** és **IECore** vagy **IEPlugin** osztályokra továbbá létrejön egy **ExecutableNetwork** egyed, amivel az hálózatot futtatjuk. A modell feltöltése az eszközre, annak alkalmazásban történő futtatását az eszközön a 3.4.1 kód szemlélteti.

3.1. kód. Inference Engine használata PYthon-ból

```

1 from openvino.inference_engine import INetwork, IEPlugin
2 net = INetwork(model= '<dir>/model.xml', weights= '<dir>/model.bin')
3 plugin = IEPlugin(device= 'MYRIAD')
4 exec_net = plugin.load(network= net)
5 i_layer_key = list(net.inputs.keys())[0]
6
7 if net.inputs[i_layer_key].shape == list(samples.shape) :
8     res = exec_net.infer(inputs={i_layer_key : samples})

```

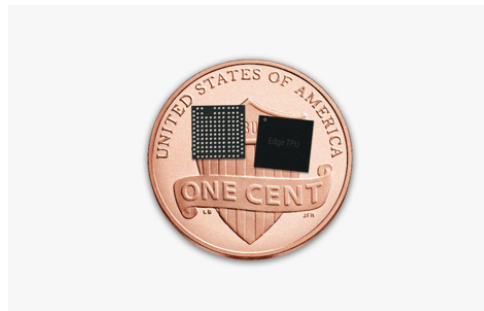
Az **IEPlugin** rendelkezik egy **load()** metódussal, mely egy **INetwork** objektumtól megkapja a hálózat adatait és példányosít egy **ExecutableNetwork** egyedet. Ezen objektum **infer()** metódusával végeztethetjük az inferenciát szinkron módon. Lehetőség van aszinkron üzemmódra az **async_infer()** metódussal. Mindkét metódus használata esetén át kell adni egy táblázat típusú paramétert (Python **dict** osztály) melynek kulcsértékei a bemeneti rétegek nevei (ez egyezik az **INetwork** **inputs** attribútumának kulcsával), a kulcsokhoz rendelt értékek pedig a minták megfelelő alakú, **numpy.ndarray** típusú tenzora. A **res** változó szintén egy táblázat

¹⁰Elrendezés: a tenzorként reprezentált minták tömbjében az adatok milyen elrendezésben szerepelnek. Például egy képfeldolgozáshoz használt hálózat bemeneti rétege esetén az 'NHWC' elrendezés azt jelenti, hogy a beolvasott tömb 1. tengelye mentén a mintákon(sample Number) való bejárást, 2. és 3. tengelye mentén a magasságit(Height) és szélességit(Width), negyedik tengelye mentén a színcsatornák(Channel) szerinti bejárást végezhetünk

lesz, amiben a kimeneti rétegek neveihez, mint kulcsokhoz van rendelve a következtetés eredménye.

3.5. Google Edge TPU

Az Intelhez hasonlóan a Google is tervezett célprocesszort és hozzá fejlesztő panelt, mély tanulást alkalmazó projektekhez. A Google Edge TPU-ról és a Coral USB Accelerator-ról a Heartbeat internetes magazinban megjelent cikkben értesültem.[19] A Google TPU-ról szerzett ismereteim a wikipédia „Tensor Processing unit” cikkéből szereztem.[33]



forrás:cloud.google.com/edge-tpu/

3.3. ábra. Edge TPU



3.4. ábra. Az Edge TPU-t építették az USB interfésszel ellátott Coral USB Acceleratorba.
forrás:[19]

A Google a *Tensor Processing Unit*ot, egy saját fejlesztésű alkalmazásspecifikus integrált áramkört használ a neurális hálózatokkal kapcsolatos számítások optimalizálására. Ezek lényegében mátrix műveletekre számításra fejlesztett processzorok CISC utasításkészlettel. Három generációt ért meg, az első még csak 8-bites egészekben tudott műveleteket végezni, azonban már a második generáció képes volt lebegőpontos számításokra is. A Tensor processing unit-okkal felszerelt gyorsítókártyákat a Google saját adatközpontjaiban használja főként és elérhetővé teszi őket a *Google Cloud Platform* nevű szolgáltatásán keresztül. Nevéből adódóan a kártya architektúrája lehetővé teszi, hogy tenzorműveletek tudjanak hatékonyan végrehajtani, utasításkészletük kifejezetten támogatja a neurális hálózatokat. Ilyen művelet a konvolúció és az 1.1 alfejezetben tárgyalt aktív függvények alkalmazása a mátrix szorzáson túl.

Az Edge TPU a cég szerverei által használt TPU-hoz viszonyítva kisebb méretben és energiafogyasztásban. Ugyan azt az igényt igyekszik kielégíteni, mint a korábban bemutatott Myriad X processzor. Azzal összehasonlítva viszont képes neurális hálózatok tanítására is korlátozott mértékben. Az Edge TPU részét képezi a Google Coral eszközkészletének, mellyel alternatívát nyújtanak a felhőalapú Cloud AI szolgáltatással szemben. A Coral különböző fajtájú hardvert nyújt a felhasználóknak, mindegyik központi eleme az Edge TPU. Az eszköz programozásához használható API-t az Edge TPU Python könyvtár (edgetpu Python modul) valósítja meg, amely képes TensorFlow Lite-ban alkotott betanított neurális hálózatok futtatására. A könyvtár kulcsfontosságú interfészeit a következő osztályok képezik. Képosztályozási feladatokhoz a ClassificationEngine használatos, vizuális objektumazonosításra a DetectionEngine. Az ImprintingEngine és SoftmaxRegression az ún. *transzfer tanuláshoz* alkalmazható, vagyis előre betanított hálózatok a feladatnak megfelelő újratanítását végezhetjük el vele.

3.6. Új gyorsítók: Intel Nervana Neural Network Processor

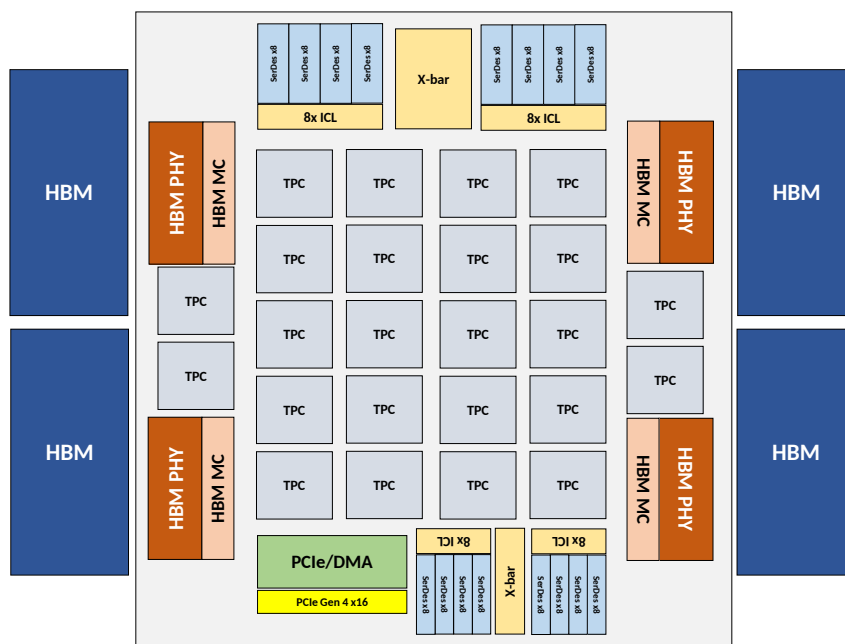
Az Intel idén augusztusban mutatta be a Hot Chips konferencián kifejezetten mély tanulásra Spring Crest és Spring Hill kódnevek alatt fejlesztett rendszerchipeket, a Neural Network Processor for Training-et (NNP-T) és a Neural Network Processor for Inference-et (NNP-I)[20]. Mint nevük is mutatja ezek kifejezetten mély tanulás gyorsításra lettek fejlesztve és úgy tervezték őket, hogy jól skálázható rendszert alkossanak. Ezeket a hardvereket a konferencián elmondottak alapján mutatnám be.[23][24] Ebben az alfejezetben szereplő ábrák a konferencián levetített fóliákból származnak.[21][22]

3.6.1. Neural Network Processor for Training

Az NNP-T egy egylapkás rendszer vagy rendszerchip, melynek fejlesztésénél fontos szempont volt az egyensúly a számítási teljesítmény, a perifériák közötti kommunikáció sebessége és memória használat között. Ezt támogatja a lapka memóriájába már betöltött adatok minél hatékonyabb újrafelhasználásának elősegítése, továbbá a kötegelt adatfeldolgozás optimalizálása. Könnyen skálázható rendszert lehet építeni belőle, ugyanis ezt beépített megoldások segítik. A jövőbeli munkafolyamatokat is támogatására a számítási egységek utasításkészlete programozható.

Hardver tulajdonságai

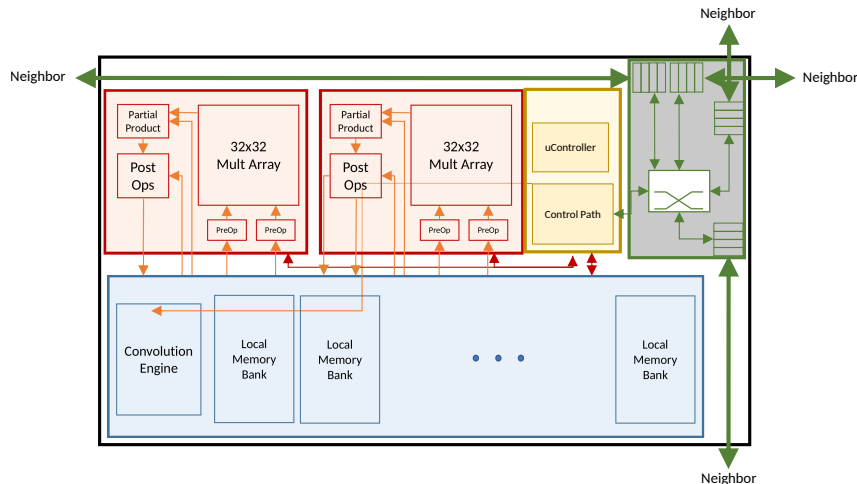
Rendszerchip A chip 4. generációs 16 sávú PCIe végponttal rendelkezik. A nagy memóriakapacitás lehetőségét lapkán 4 darab HBM2¹¹ memória foglalattal oldották meg. A lapkák közötti kommunikációra SerDes interfész 64 sávval is implementálva lett. A számításokat 24 darab tenzor processzormag végzi (TPC), és a vállalat szerint számítási kapacitásuk összesen 119 TOP/s (billió utasítás másodpercenként). A lapkán elhelyeztek 60 MB osztott memóriát is.



3.5. ábra. Az NNP-T egylapkás rendszer (SoC) felépítése

Tenzor processzor (TPC) Egy TPC magon belül két darab mátrixszorzó tömb van, melyek egyenként 32×32-es méretű mátrixok tud szorzást végezni, a konvolúció műveletet pedig külön egység végzi. Minden mag továbbá rendelkezik 2,5 MB Cache memóriával, melyek sávszélessége 1,4 TB/s. A 3.6 ábrán látható zöld színnel jelzett útválasztó a magok közötti kommunikációért felelős.

¹¹High Bandwidth Memory: DRAM modulok térben egymásra pakolását alkalmazó RAM interfész típus. HBM2 esetén a legnagyobb kapacitás 8GB lehet egy tokban. Leginkább videó kártyákban használatos

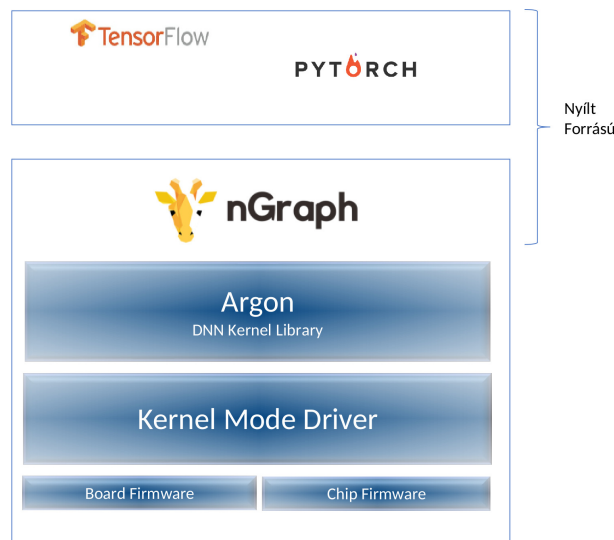


3.6. ábra. Egy TPC felépítése

hardverkártya tulajdonságai Az Intel az NNP-T rendszerchipet egyedi PCIe és OAM alak-tényezőjű kártyán fogja forgalmazni, melyre egy rendszerchipet szerelnek összesen 32 GB HBM2-2400 típusú memóriával – melynek elméleti sávszélessége 1,22 TB/s – BGA tokozás-ban. A rendszerchip órajele az 1,1GHz-et is elérheti. A kártya 4. generációs PCI Express x16, valamint SPI, I2C csatlófelülettel és általános célú I/O kivezetéssel rendelkezik. A központi egység léghűtésű átlagos használat esetén 150 és 250 W közötti fogyasztással. A kártyán kivezetésre került 64 SerDes sáv, melynek aggregált sávszélessége 3,58 Tb/s. Ezen interfész segít megvalósítani több lapka összehangolt működését, mellyel maximum 1024 csomópontot lehet összekötni úgy, hogy azok összes TPC magja egy nagy hálózatként kezelhető, ezzel elérve a masszív párhuzamosítást.

Az NNP-T programozásához használatos szoftver együttes

Az alkalmazásfejlesztés az ilyen rendszerchipekkel szerelt kártyákra egyedi eszközkészlettel lehetséges. A vállalat ajánl egy teljes fejlesztői szoftverkészletet nyílt forrású komponensekkel. A 3.7 ábra az egyes komponenseket hardver és felhasználó közötti viszonya szerint rétegezve mutatja be. Ezek közül az nGraph-ot és a magas szintű keretrendszerek közül a TensorFlow-t, már bemutattam korábban, illetve pár szóban összefoglalom az Argon programkönyvtárat.



3.7. ábra. Fejlesztéshez használható szoftveregyüttes

NNP-T Programozási modell

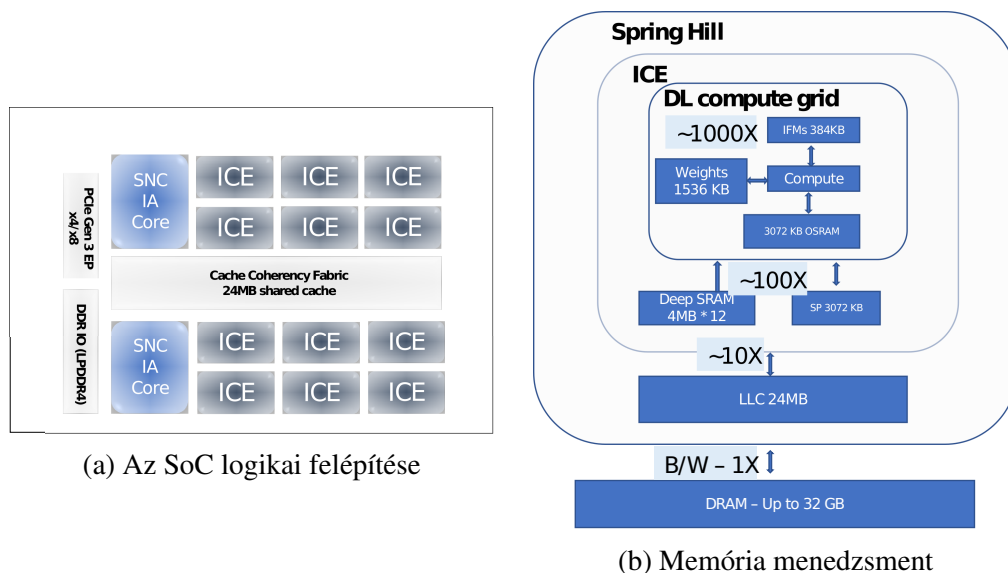
Az Utasításkészlet szintjén is neurális hálózatokra lett optimalizálva a rendszerchip azzal, hogy a hálózat tanításakor alkalmazandó tenzor műveletekhez natív utasítások tartoznak, úgy mint transzponálás, tenzor darabolása, stb. Az utasításkészlet is ezen utasításokra van korlátozva, azonban ez bővíthető egyedi mikrokontroller kutasításokkal. A bővíthetőség azt a célt szolgálja, hogy a jövőbeli mély tanulásra épülő munkafolyamatok támogatottsága is megoldható legyen utasítás szinten. Az Intel hierarchikus elosztott programozási modellt használ, ami jól skálázható, így alkalmazható egy rendszerchip TPC magjaira, vagy akár több lapka együttes programozáskor. Ezt segíti, hogy a magok közti kommunikáció és szinkronizáció elemi utasításai konzisztensek lapkán belül és több lapka között. Azon túl, hogy a hardver az adatokat tenzoroként kezeli, a tenzorokon belüli elemi változókat Bfloat16¹² és 32 bites lebegőpontos típusként tudja értelmezni

3.6.2. Neural Network Processor for Inference

Betanított neurális hálózatok futtatására alkalmazható rendszerchip, melyet az NNP-T-vel együtt kezdett el fejleszteni a vállalat. Ennél a lapkánál a cél a kis fogyasztás volt. A gyártó szerint Wattonkénti 4,8 billió számítást képes végezni másodpercenként. A számítási teljesítmény a fogyasztás hangolásával természetesen változó. (Jelenleg M.2 és PCIe alaktényezőjű kártyákon forgalmazzák. Előbbi 12 wattos teljesítményű, míg az utóbbi teljesítménye 75 W. Számítási kapacitásuk 50 és 170 TOP/s)

¹²Az IEEE 754 félszeres, azaz 16 bites pontosságú lebegőpontos típustól annyiban tér el, hogy az exponens 5 helyett 8 bites. Ez az Intel állítása szerint azon túl, hogy memória hely takarékos, tanításkor jól konvergáló változókat biztosít

Hasonlóan az NNP-T-hez ez a rendszerchip is egyedi processzormagokat tartalmaz, összesen 12-t. A mag megnevezése *Inference Compute Engine*, röviden ICE. Az lapka (3.8a ábra) ki egészül még 24 MB cache memóriával, melyen az összes mag osztozik. Az ICE magok mellett két Intel IA Core processzormag is részét képezi a rendszerchipnek. Ezek párban az ICE-k munkáját hangolják össze a fordítótól érkező utasítások alapján, – ide kapcsolódik, hogy *Ahead Of Time* fordítás végezhető az eszközre – tehát a programozó által közvetlenül ez a két processzormag elérhető. Egy ICE két számítást végző blokkból áll: egy *Deep Learning compute grid*-nek nevezett több számítási egységből álló hálózattól, – melyek a neurális hálózatok modellezése során igényelt műveletek javát végzik – illetve egy programozható vektor processzorból AVX és VNNI utasításkészlettel. A vektor processzor jóvoltából tetszőleges utasításokkal bővíthető a mag.



(a) Az SoC logikai felépítése

(b) Memória menedzsment

3.8. ábra

Egy ICE 4 MB lokális SRAM memóriát kapott. Ezen ICE magok képesek 16 bites lebegőpontos valamint 8, 4, 2 és 1 bites egész változókat kezelni. Az NNP-I lapka 3. generációs PCIe (x4 vagy x8) végponttal kapcsolódik a befogadó kártyához, továbbá egy LP-DDR4 memória ültethető a hordozó lapkára az S . Azért, hogy az adatmozgatást minimalizálják a memória allokáció 4 szinten van rendezve hierarchikusan, ahogy az 3.8b ábrán látható. Ezt a rendszerchipet ugyanazzal a szoftver együttessel tudjuk programozni, mint az NNP-T SoC-t.

Összefoglalás

A gépi tanulás olyan területeken jelent meg melyek új kihívásokat szültek. A hardvergyártók kifejezetten deep learningre optimalizált megoldásokat kínálnak és a jövőbeli fejlesztések is ebbe az irányba mutatnak. Ezen dolgozatomban bemutatott, mély tanulásban használatos technológiákon látható, hogy egyre változatosabb deep learning rendszereket alapját tudják nyújtani. Olyan keretrendszerek születtek független fejlesztőcsapatok és a hardvergyártók jóvoltából, melyekkel gyorsan és hatékonyan végezhető mély tanuló alkalmazások fejlesztése. Ennek jóvoltából a mindennapi életben is megjelenik ez a technológia, és azon is túl a kutatásokban megjelenő nehezen megoldható számítási problémák váltak kivitelezhetővé. Ezen technológiákat felölelő szakterületek a munkaerőpiacon is igen keresetté válhatnak.

A hardveripar most is sok erőforrást fektet a gyakorlatban használt mesterséges neurális hálózatok hatalmas számítási igényének kielégítésére. Újabb és hatékonyabb eszközöket kínálnak erre a feladatra. Némelyikük a korábban bevált GPGPU architektúrák továbbfejlesztése, míg mások speciális, csakis kifejezetten deep learning architektúrák. Ezek az eszközök a több magos masszívan párhuzamos számításokra alapoznak, melyek jól alkalmazhatóak a neurális hálózatokon.

Ezzel párhuzamosan hibrid számítási rendszereket támogató keretrendszerek is készülnek, szintén a párhuzamos számítás modellre, azontúl az egyes architektúrák eltérő működésére alapozva. Ennek köszönhetően a deep learning alkalmazások változatos hardver architektúrákon futtathatóak, melyek a nyers erőn túl a számítások hatékony végrehajtásával érnek el teljesítménynövekedést. Az ilyen technológiák továbbá lehetőséget biztosítanak, hogy a felhasználók meglévő hardvereiket építsék be deep learning rendszerekbe. Ez a fejlesztési költségek szempontjából nagyon előnyös, ezért a jövőben várhatóan egyre több szerveren találkozhatunk tanuló vagy tanítható alkalmazásokkal.

A nagygépes világgal párhuzamosan megjelentek külön mély tanulásra specializált integrált áramkörök a hordozható és személyi eszközökön is. A Deep Learning így egy „kézzel fogható” technológiává válik. Ez az irány elősegíti a valós idejű mély tanuló alkalmazások fejlesztését, mint például az önvezető járművek, intelligens gyártósorok, személyi applikációk vagy egyes orvosi alkalmazások. Úgy látszik a jövőben erre a szolgáltatók is számítanak: ezen speciális hardverek némelyike építőkövei lehetnek az ún. *Edge Computing* modellnek: míg kezdetben a

felhőszolgáltatásokat nyújtó vállalatok neurális hálózatai teljes egészében szervereiken futottak, mára kezd kibontakozni az az elképzelés, hogy ezeket a számításokat a felhasználóhoz egészen közel lehet vinni, így közel valós idejűek lehetnek a felhő alapú deep learning szolgáltatások is a jövőben.

A szakdolgozatom írása során rengeteg új ismeretre tettem szert. Igyekeztem a legmodernebb hardvereket és legújabb szoftverfejlesztéseket felkutatni. Így szembesültem azzal is, hogy ennek a szakterületnek csak felszínét vizsgáltam még. Ahhoz, hogy tényleg jó rálátásom legyen a jövőbeli fejlesztésekre, a gépi tanulással kapcsolatos ismereteimet mesterképzésben szeretném elmélyíteni.

Az nGraph lefordításának sikertelensége témavezetőm hibrid rendszerére nem tántorított el a korábban megfogalmazott terv kivitelezésétől. A keretrendszer élénk fejlesztését látva van esély rá, hogy a későbbiekben újból megkíséreljük lefordítani a HuSSarra, hogy összevethessük az nGraph teljesítményét más hibrid rendszereken működő fordítókéval.

Irodalomjegyzék

- [1] François Chollet, **Deep Learning with Python**, Manning Publications, 2018.
- [2] Altrichter Márta & Horváth Gábor & Pataki Béla & Strausz György & Takács Gábor & Valyon József, **Neurális hálózatok**, Panem, 2006
- [3] Michael A. Nielsen, **Neural Networks and Deep Learning**, Determination Press, 2015, <http://neuralnetworksanddeeplearning.com/>, [meglátogatva: 2019. október 30.]
- [4] Agulleiro, J. I., Vázquez, F., Garzón, E. M. & Fernández, J. J., **Hybrid computing: CPU+GPU co-processing and its application to tomographic reconstruction**, Ultramicroscopy, 2012, doi:10.1016/j.ultramic.2012.02.003
- [5] Meredith, J., Roth, P., Spafford, K. & Vetter, J., **Performance Implications of Nonuniform Device Topologies in Scalable Heterogeneous Architectures**, IEEE Micro, 2011, doi:10.1109/mm.2011.79
- [6] Li, X., Wan, J., Dai, H.-N., Imran, M., Xia, M. & Celesti, A., **A Hybrid Computing Solution and Resource Scheduling Strategy for Edge Computing in Smart Manufacturing**, IEEE Transactions on Industrial Informatics, 2019, doi:10.1109/tii.2019.2899679
- [7] **NVIDIA TITAN RTX Specs**, Techpowerup, www.techpowerup.com/gpu-specs/titan-rtx.c3311, [meglátogatva: 2019. november 19.]
- [8] **NVIDIA Tesla T4 Specs**, Techpowerup, www.techpowerup.com/gpu-specs/tesla-t4.c3316, [meglátogatva: 2019. november 19.]
- [9] **NVIDIA Tesla V100 SXM2 32 GB Specs** , Techpowerup, www.techpowerup.com/gpu-specs/tesla-v100-sxm2-32-gb.c3185, [meglátogatva: 2019. november 19.]
- [10] Az Nvidia v100 adatlapja, footnotesizeimages.nvidia.com/content/technologies/volta/pdf/tesla-volta-v100-datasheet-letter-fn1-web.pdf, [megtekintve: 2019. november 19.]

- [11] Martín Abadi and Ashish Agarwal and Paul Barham and Eugene Brevdo and Zhifeng Chen and Craig Citro and Greg S. Corrado and Andy Davis and Jeffrey Dean and Matthieu Devin and Sanjay Ghemawat and Ian Goodfellow and Andrew Harp and Geoffrey Irving and Michael Isard and Yangqing Jia and Rafal Jozefowicz and Lukasz Kaiser and Manjunath Kudlur and Josh Levenberg and Dandelion Mané and Rajat Monga and Sherry Moore and Derek Murray and Chris Olah and Mike Schuster and Jonathon Shlens and Benoit Steiner and Ilya Sutskever and Kunal Talwar and Paul Tucker and Vincent Vanhoucke and Vijay Vasudevan and Fernanda Viégas and Oriol Vinyals and Pete Warden and Martin Wattenberg and Martin Wicke and Yuan Yu and Xiaoqiang Zheng, **TensorFlow: Large-Scale Machine Learning on Heterogenous Systems**, 2015, [Elérhető: www.tensorflow.org/about/bib]
- [12] **API Documentation | Tensorflow Core r2.0**, www.tensorflow.org/api_docs, [meglátogatva: 2019. november 21.]
- [13] **TensorFlow Core**, www.tensorflow.org/guide [meglátogatva: 2019. november 21.]
- [14] **Keras Dokumentáció**, keras.io/, [meglátogatva: 2019. október 23.]
- [15] **Introduction — Documentation for the nGraph Library and Compiler stack**, 2019, ngraph.nervanasys.com/docs/latest/introduction.html, [meglátogatva: 2019. október 08.]
- [16] **PlaidML – Home**, 2019, Vertex.AI., vertexai-plaidml.readthedocs-hosted.com/en/stable/, [meglátogatva: 2019. október 20.]
- [17] **cuDNN Developer Guide**, 2019, Nvidia, <https://docs.nvidia.com/deeplearning/sdk/cudnn-developer-guide>, [meglátogatva: 2019. november 19.]
- [18] **OpenVINO™ toolkit Documentation**, 2019, docs.openvinotoolkit.org/2019_R3/index.html, [meglátogatva: 2019. november 4.]
- [19] **Edge TPU: Hands-On with Google’s Coral USB Accelerator**, heartbeat.fritz.ai/edge-tpu-google-coral-usb-accelerator-cf0d79c7ec56, [meglátogatva: 2019. október 23.]
- [20] **At Hot Chips, Intel Pushes ‘AI Everywhere’**, 2019, newsroom.intel.com/news/hot-chips-2019, [meglátogatva: 2019. November 07.]
- [21] Andrew Yang, **Deep Learning Training At Scale Spring Crest Deep Learning Accelerator (Intel® Nervana™ NNP-T)**, 2019, www.slideshare.net/insideHPC/deep-learning-training-at-scale-spring-crest-deep-learning-accelerator/, [letöltve: 2019. november 07.]
- [22] Ofri Wechsler & Michael Behar & Bharat Daga, **Spring Hill (NNP-I 1000) Intel’s Data Center Inference Chip**, 2019, www.slideshare.net/insideHPC/spring-hill-nnpi-1000-intels-data-center-inference-chip/, [letöltve: 2019. november 07.]
- [23] Intel AI, **Introducing Intel Nervana Neural Network Processors for Training**, 2019, www.youtube.com/watch?v=Wj4ifr3DDFg, [megtekintve: 2019. november 10.]
- [24] Intel AI, **Introducing Intel Nervana Neural Network Processors for Inference**, 2019, www.youtube.com/watch?v=fTMIjQePWog, [megtekintve: 2019. november 10.]

- [25] **plaidml/plaidml: PlaidML is a framework for making deep learning work everywhere**, GitHub repository, github.com/plaidml/plaidml, [meglátogatva: 2019. október 20.]
- [26] **Graphs and Sessions**, GitHub repository, github.com/tensorflow/docs/blob/master/site/en/r1/guide/graphs.md [meglátogatva: 2019. november 21.]
- [27] **Sigmoid function**, Wikipedia, The Free Encyclopedia, 2019, en.wikipedia.org/w/index.php?title=Sigmoid_function&oldid=910833562, [meglátogatva: 2019. augusztus 21.]
- [28] **Softmax function**, Wikipedia, The Free Encyclopedia, 2019, en.wikipedia.org/w/index.php?title=Softmax_function&oldid=913685380, [meglátogatva: 2019. szeptember 10.]
- [29] **Constant folding**, Wikipedia, The Free Encyclopedia, 2019, en.wikipedia.org/w/index.php?title=Constant_folding&oldid=914455114, [meglátogatva: 2019. október 09.]
- [30] **List of Intel graphics processing units**, Wikipedia, The Free Encyclopedia, 2019, en.wikipedia.org/w/index.php?title=List_of_Intel_graphics_processing_units&oldid=926282215, [meglátogatva: 2019. október 20.]
- [31] **PlaidML**, Wikipedia, The Free Encyclopedia, 2019, en.wikipedia.org/w/index.php?title=PlaidML&oldid=898300482, [meglátogatva: 2019. október 5.]
- [32] **Rectifier (neural networks)**, Wikipedia, The Free Encyclopedia, 2019, [en.wikipedia.org/w/index.php?title=Rectifier_\(neural_networks\)&oldid=923288576](https://en.wikipedia.org/w/index.php?title=Rectifier_(neural_networks)&oldid=923288576), [meglátogatva: 2019. október 29.]
- [33] **Tensor processing unit**, Wikipedia, The Free Encyclopedia, 2019, en.wikipedia.org/w/index.php?title=Tensor_processing_unit&oldid=923241091, [meglátogatva: 2019. november 4.]
- [34] **CUDA**, Wikipedia, The Free Encyclopedia, 2019, <https://en.wikipedia.org/w/index.php?title=CUDA&oldid=924918335>, [meglátogatva: 2019. november 19.]
- [35] Mark Tyson, **Industry sources say discrete Intel Xe cards will arrive mid-2020**, 2019. október 16. hexus.net/tech/news/graphics/135800-industry-sources-say-discrete-intel-xe-cards-will-arrive-mid-2020 [megtekintve: 2019. november 21.]
- [36] Darren Allan, **Intel's first Xe graphics card is officially 'alive' and coming for AMD and Nvidia**, 2019. október 26. www.techradar.com/news/intels-first-xe-graphics-card-is-officially-alive-and-coming-for-amd-and-nvidia [megtekintve: 2019. november 21.]
- [37] Dr. Ian Cutress, **Intel's Xe for HPC: Ponte Vecchio with Chiplets, EMIB, and Foveros on 7nm, Coming 2021**, 2019. november 17. www.anandtech.com/show/15119/intels-xe-for-hpc-ponte-vecchio-with-chiplets-emib-and-foveros-on-7nm-coming-2021 [megtekintve: 2019. november 21.]
- [38] Andy Patrizio, **Intel targets Nvidia (again) with GPU and cross-processor API**, 2019. november 19. www.networkworld.com/article/3454497/intel-targets-nvidia-again-with-gpu-and-cross-processor-api.html [megtekintve: 2019. november 21.]

Köszönetnyilvánítás

Ezúton köszönöm témavezetőmnek biztatását amit a neurális hálózatokkal való ismerkedésem során nyújtott továbbá, hogy mindig számíthattam a segítségére. A mély tanulás iránti lelkesedésével engem is átítatott, mely átsegített a nehézségeken. Nélküle talán sose ismerkedtem volna meg közelebbről ezzel a területtel. Szeretném továbbá megköszönni szüleimnek azt, hogy mellettem álltak, mikor szakterületet váltottam és támogatták egyetemi tanulmányaimat.