Atraskime Higgs’o bozoną iš naujo arba paieškos dvylikamatėje erdvėje

Projekto „Mokslo pieva“ ataskaita

Komandos vadovai:

Jonas Matuzas

Aurelijus Rinkevičius

Tyrimą atliko:

Gediminas Drabavičius

Daumantas Stanikūnas

Agnė Smigelskytė

Ąžuolas Krušna

Vilnius, Kaunas, 2014

Įvadas

Higgs'o bozonas yra fundamentali elementarioji dalelė, standartiniame modelyje, kuri, plačiai tariant, suteikia visoms kitoms dalelėms masę ir dėl sąveikos tarp dalelių, kurią įgalina Higgs'o bozonas, susikūrė tokia visata, kokią matome dabar.

Originaliai, „Į Higgs'o bozoną panašios dalelės" atradimas buvo paskelbtas 2012m. liepos 4d. CERN mokslininkų, kur vienas iš mūsų vadovų – Aurelijus Rinkevičius – yra doktorantas, tad turėjome išskirtinę prieigą prie eksperimentinių duomenų. Dviratį išradinėti nutarėme todėl, kad yra žinomas labai galingas įrankis tokiai analizei atlikti, kuriuo nebuvo naudojamasi darant originalų atradimą. Tas įrankis yra Bajesiniai neuroniai tinklai (BNN).

BNN yra labai plačiai naudojami biologijoje, bioinformatikoje, duomenų klasifikavime, paveiksliukų apdorojime, inžinerijoje ir netgi teisėje. Kadangi CERN turi paketą paruoštą tokioms analizėms atlikti (jis turėjo daugiau pasisekimo ieškant kitų dalelių, nei Higgs'o bozono, kuris eksperimentų metu yra sugeneruojamas ypač retai) nutarėme jį pritaikyti būtent šiai užduočiai.

Projekto metu turėjome išmokti naudotis Linux komandine eilute, rašyti BASH shell kodą, o taip pat nutarėme plačiau susipažinti su mašininiu mokymusi ir neuroniniais tinklais. Visi komandos nariai, projekto metu klausė Andrew Ng, pasaulinio mašininio mokymosi eksperto, paskaitų Coursera tinklapyje. Bei bandėme išmokti kurti savus neuroninius tinklus Octave, Matlab ir R terpėse. Iš šių pasižaidimų išplaukė kompiuteriniai algoritmai išmokstantys traukti kvadratinę šaknį, atpažinti ranka rašytus skaitmenis ir atpažinti žmogaus veidą (ne atpažinti žmogų, bet tai, kad paveikslėlyje ar filmuke yra žmogaus veidas) ir jį nurodyti.

O pagrindinės užduoties sprendimas kulminavo tuo, kad BNN buvo apmokytas atpažinti Higgs'o bozono signalą ir sugebėjo jį išskirti iš triukšmo, kurio apstu tikruose eksperimentuose. Deja, CERN duomenys pasirodė ne tokie laisvai prieinami, kaip tikėjomės, tad teko visus darbus atlikti su sintetiniais duomenimis – tokiais, kurie buvo sugeneruoti tikrų signalų pagrindu, tačiau nebuvo paimti tiesiogiai iš eksperimento. Taigi galime teigti, kad turime algoritmą, pakartotinai atpažinti Higgs'o bozono signalui iš triukšmingų duomenų. Žinoma, tai reikėtų patikrinti su tikrais eksperimentiniais duomenimis.

Teorija ir metodai

Komandos nariai, mokymosi metu susipažino su Linux OS ir jos Terminal valdymu, Bash Shell Scripting, Python programavimo kalba bei R, Matlab ir Octave programavimo aplinkomis. Pagrindinis darbas atliktas naudojant Radford M. Neal sukurtą BNN paketą.

Pirmojoje darbo stadijoje, buvo susipažinta su minėtu paketu, paleidžiant jį pagal kūrėjo instrukcijas ir naudojant pavyzdinius duomenis. Šis pavyzdys buvo supaprastinta mūsų galutinio darbo versija. Duomenys jame buvo 7 dimensijose, o duomenų masyvas buvo sudarytas iš 5000-10000 tūkstančių taškų. Tad įsisavinus pavyzdžio veikimą, mums šį paketą reikėjo pritaikyti darbui 16 dimensijų (kiekvienas taškas turi po 16 parametrų) ir daug didesniems duomenų kiekiams.

Higgs'o bozono signalo paieška:

BNN pakete yra naudojamas „Supervised learning" (Prižiūrimo mokymosi) modelis, kuriam yra reikalinga mokymosi duomenų aibėje, kurioje kiekvienas taškas yra įvardintas kaip signalas (1) arba triukšmas (0). Sintetinius duomenis, naudodami kaip mokymosi aibę, galime priversti kompiuterinį BNN algoritmą išmokti atpažinti signalą.

Duomenų paruošimas

Neuroninio tinklo apmokymui skirti duomenys turėjo būti atitinkamai paruošti. Proceso pradžioje, visi sintetiniai duomenys yra sumaišomi, kad nebūtų jokio eiliškumo tarp duomenų identifikuojamų kaip signalas ar triukšmas, tačiau duomenų faile yra pridedamas papildomas stulpelis identifikuojantis priklausomybę vienai arba kitai duomenų grupei. Visi duomenys, tuomet buvo sunormuoti, kad jų reikšmės turėtų vienodą svarumą neuroniniam tinklui mokantis. Taip išvengiame didelių klaidų atsirandančių dėl labai nehomogeniško duomenų pasiskirstymo ir sumažiname skaičiavimui naudojamų kompiuterinių resursų reikalavimus, bei, apskritai, pagreitiname mokymąsi – sumažiname šio proceso trukmę. Tai yra itin svarbu, kadangi net naudojant superkompiuterį, mokymasis truko apie 8 valandas.

Neuronio tinklo sandara

Buvo minėta, kad naudoti sintetiniai duomenys buvo pateikti 16 dimensijų. Tai reiškė, kad tinklo apmokymui buvo reikalingi 16 įvesties mazgų arba „neuronų". Paslėptųjų mazgų skaičius yra pasirenkamas laisvai. Didesnis paslėptų mazgų leidžia neuroniniam tinklui susikurti sudėtingesnes funkcijas arba gluodinti sudėtingesnes duomenų aibes. Kadangi mūsų problema buvo duomenų klasifikavimo – tai yra, išmokyti neuroninį tinklą priskirti duomenį prie triukšmo arba prie signalo – buvo pasirinktas gana nedidelis mazgų skaičius – 20. Kadangi, apmokytas klasifikatorius turėjo tik dvi kategorijas – 1 arba 0, užteko vieno išvesties mazgo.

Rezultatų gavimas

Apmokinę neuronį tinklą, pasitelkėme pagrindinį paketo algoritmą, mes sukūrėme failą, kurį modifikavę pritaikėme jį mūsų problemai spresti. Kaip buvo minėta, reikėjo pakeisti dimensijų skaičių ir įvesties duomenų skaičių. Tai padarę, ir davę algoritmui sintetinius eksperimento duomenis gavome išvesties failą, kuris buvo vienmačių vektorių formato, kuriame kiekviena reikšmė yra tarp 0 ir 1, kurios artumas 1 ar 0 reiškia neuroninio tinklo užtikrintumą, kad taškas yra signalas arba triukšmas.

Atvaizdavimas

Gautas duomenų failas buvo atvaizduotas imant 30 triukšmo arba signalo išvesties reikšmių ir jas sudauginant bei atidedant ant ašies tarp 0 ir 1 histogramų pavidalu. Šios sandaugos daromos, su visais išvesties duomenimis. Skaičius 30 buvo pasirinktas arbitraliai, kadangi tikruose CERn eksperimentuose Higgs'o bozono signalas buvo aptiktas 30 kartų. Jas dauginame, nes laikome šiuos įvykius nepriklausomai, tad vieno iš jų tikimybė yra pavienių įvykių tikimybių sandauga. Rezultatai atvaizduoti pasitelkus Matlab arba R programinius paketus.

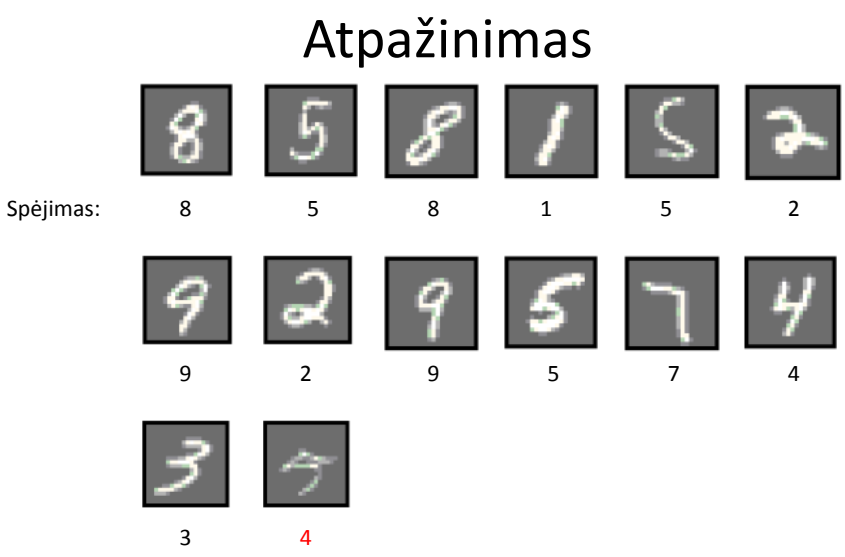
Rezultatai

Projekto metu buvo sukurti keli mašininio mokymosi algoritmai naudojantis R ir Matlab programavimo terpėmis, kurie leido geriau įsisavinti neuroninių tinklų ir, apskritai, mašininio mokymosi idėjas. Keli iš jų pateikiami toliau.

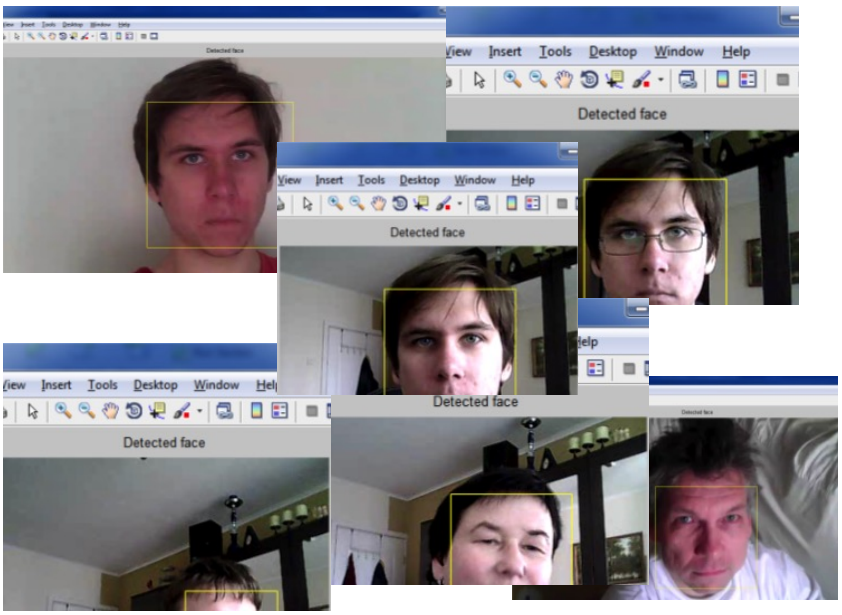
Neuroninio tinklo rezultatai bandant trauki kvadratinę šaknį

Duomenys:	Atsakymai:	Rezultatai:	Paklaidos:
❑ 1	✓ 1	➢ **1.002256499**	*0.002256499*
❑ 4	✓ 2	➢ **1.996889304**	*0.003110696*
❑ 9	✓ 3	➢ **3.002788856**	*0.002788856*
❑ 16	✓ 4	➢ **3.998681298**	*0.001318702*
❑ 25	✓ 5	➢ **4.999641878**	*0.000358122*
❑ 36	✓ 6	➢ **6.001388804**	*0.001388804*
❑ 49	✓ 7	➢ **6.999822078**	*0.000177922*
❑ 64	✓ 8	➢ **7.999051240**	*0.000948760*
❑ 81	✓ 9	➢ **9.001323620**	*0.001323620*
❑ 100	✓ 10	➢ **9.993539852**	*0.006460148*

1 pav. Neuroninis tinklas išmokęs traukti kvadratinę šaknį

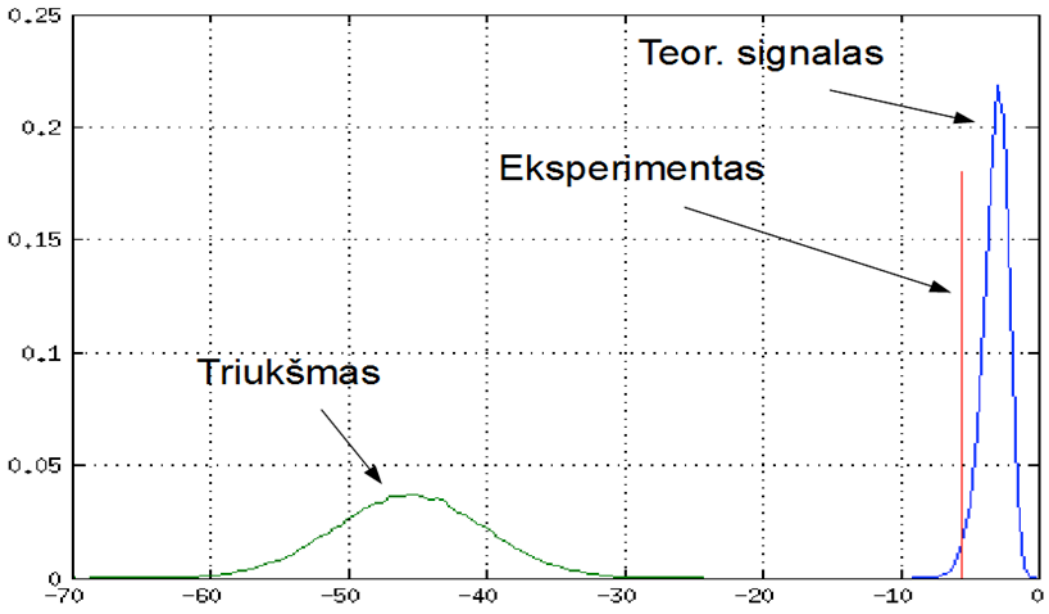


2 pav. Neuroninis tinklas 97,52 proc. tikslumu atpažįstantis ranka rašytus skaitmenis



3 pav. Algoritmas išmokęs atpažinti ir nuotraukoje surasti žmogaus veidą.

Galutinis darbo rezultatas buvo tai, kad Bajesinis neuroninis tinklas buvo apmokytas atpažinti Higgs'o bozono signalą ir išskirti jį iš aplinkos triukšmo. Tai yra, buvo apmokytas klasifikatorius, kuris gavęs 16 dimensijų vektorių gali pasakyti ar labiau tikėtina, kad tai triukšmas ar galimas Higgs'o bozono signalas. Deja, projekto metu, nepavyko gauti tikrų eksperimentinių duomenų, tad tinklas nebuvo patikrintas su tikrais eksperimentiniais duomenimis, tačiau su sintetiniais, veikia patikimai.



4 pav. Higgs'o bozono aptikimo simuliuotas eksperimentas. Atidėta logaritminėse koordinatės

Išvados

Užsibrėžtas darbo tikslas buvo pasiektas, tačiau būtų įdomu patikrinti gautą BNN su tikrais duomenimis, kurs signalas yra itin retas. Darbo metu, taip pat išmokome daugiau nei buvome užsibrėžę ir įgavome naudingų praktinių įgūdžių. Nors kol kas, mūsų rašomi mašininio mokymosi algoritmai yra primityvūs ir gali turėti tiktai labai ribotą praktinį pritaikymą, galima tikėtis, kad netolimoje ateityje, jie galėtų būti naudojami spręsti plačiam problemų spektrui.

Tolimesnės darbo kryptys galėtų būti išsiplėtimas į kitas mokslų sritis, kadangi tokio pobūdžio neuroniniai tinklai, potencialiai, gali būti taikomi bet kokio pobūdžio skaitiniams duomenims.

***

Projektą „Mokslo pieva“ organizuoja mokslininkų ir dėstytojų komanda iš Baltijos pažangių technologijų instituto, Kauno technologijos universiteto, Socialinių inovacijų instituto, Vilniaus universiteto ir Vytauto Didžiojo universiteto.

Projekto metu atliekami įvairūs tyrimai iš fizikos, IT, socialinių mokslų bei kitų disciplinų.

2014 m. projekto „Mokslo pieva“ idėją rėmė UAB „Philip Morris Baltic“. Tyrimų temos, išvados bei rekomendacijos išreiškia autorių asmeninę nuomonę.

Daugiau informacijos: www.mokslopieva.lt