Katastrofalno pozabljanje pri inkrementalnem učenju konvolucijske nevronske mreže

Jakob Božič, Danijel Skočaj

Fakulteta za računalništvo in informatiko, Univerza v Ljubljani
E-pošta: jakob.bozic@gmail.com, danijel.skocaj@fri.uni-lj.si

Catastrophic forgetting during incremental learning of convolutional neural network

Catastrophic forgetting is a well-documented phenomenon which occurs during incremental learning of artificial neural networks. When trained on a new task, the network very rapidly and almost completely forgets how to perform previously learned tasks. We investigate the main causes of catastrophic forgetting in a deep convolutional neural network for image classification, how fast it occurs and how intensive it is. Different approaches to updating network parameters, aimed at preventing or at least alleviating the catastrophic forgetting are proposed and evaluated.

1 Uvod

V zadnjih letih so glavni akter na področju računalniškega vida postale (globoke) umetne nevronske mreže, ki na določenih problemih že dosegajo ali celo presegajo človeške zmožnosti. Obsežne podatkovne zbirke so delno rešile problem potrebe po velikih učnih množicah, razne regularizacijske tehnike dobro preprečujejo preveliko prilagajanje učnim podatkom, katastrofalno pozabljanje oz. inkrementalno učenje pa ostaja eden izmed odprtih problemov.

Do katastrofalnega pozabljanja pride, ko želimo obstoječo nevronsko mrežo, ki rešuje določen problem, naučiti reševanja novega problema. Mreža ob učenju novega problema zelo hitro in skoraj popolnoma pozabi, kako se rešuje prejšnji problem. Na primer, če lahko z obstoječo mrežo prepoznavamo števila, želeli bi pa jo uborabiti tudi za prepoznavanje črk, bi mreža ob učenju prepoznave črk pozabila, kako se prepozna števila.

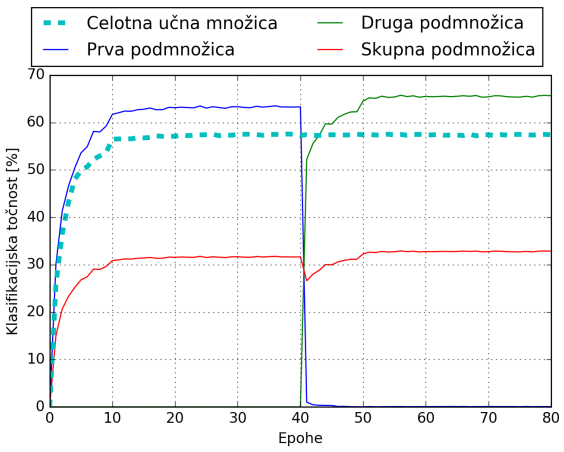
V tem prispevku se bomo posvetili proučevanju tega problema. Podrobno analizo katastrofalnega pozabljanja v globokih konvolucijskih nevronskih mrežah bomo opravili skozi različne eksperimente, ugotovitve pa bomo uporabili za razvoj različnih pristopov k osveževanju parametrov mreže, s katerimi želimo katastrofalno pozabljanje preprečiti ali vsaj omiliti.

Katastrofalno pozabljanje je bilo opisano že v 80. letih prejšnjega stoletja [9, 10], sicer na zelo plitvih polno povezanih nevronskih mrežah in preprostih problemih. Prvi pristopi k zmanjševanju so se pojavili relativno kmalu. Tako v [3] kot v [4] problem rešujejo z ortogonalizacijo vhodnih podatkov.

Nedavno so se pojavile nove metode, ki bolj ali manj uspešno naslavljajo katastrofalno pozabljanje v globokih konvolucijskih nevronskih mrežah. Nekatere se zanašajo na spreminjanje kriterijske funkcije, da kaznuje spreminjanje parametrov, ki so ocenjeni kot pomembni za prej naučene naloge, npr. [5, 2, 1] . V [8] avtorji za vsak problem naučijo binarno masko vseh parametrov, ki določa, ali se parameter pri uporabi upošteva ali ne. Naši pristopi temeljijo na manipuliranju gradienta med vzvratnim razširjanjem.

2 Katastrofalno pozabljanje

Inkrementalno učenje bi nam omogočalo, da bi lahko obstoječo nevronsko mrežo uporabili tudi za nove naloge, ne da bi morali ob tem mrežo ponovno naučiti tudi že do tedaj osvojenih nalog. Ponovno učenje je lahko zelo dolgotrajno, v kolikor pa iz kateregakoli razloga starih učnih podatkov nimamo več na voljo, sploh ni mogoče. Želeli bi, da bi ob inkrementalnem učenju na dveh ločenih podmnožicah dosegli enako ali vsaj približno tako dobre rezultate, kot če bi imeli že na začetku na voljo vse podatke.



Slika 1: Klasifikacijske točnosti ob učenju na celotni zbirki CIFAR-100 ter ob učenju na dveh podmnožicah.

Na sliki 1 so predstavljene klasifikacijske točnosti, ki jih dobimo ob učenju na celotni podatkovni zbirki CIFAR-100 (svetlo modra prekinjena) ter ob učenju na dveh podmnožicah te zbirke (temno modra in zelena). Podmnožici

dobimo tako, da zbirko razbijemo na dva dela, vsak del vsebuje polovico razredov in vse njim pripadajoče primere. Rdeča črta predstavlja povprečje klasifikacijskih točnosti obeh podmnožic. V idealnem scenariju bi se po 40. epohi, ko se začne učenje na drugi podmnožici, rdeča črta začela približevati svetlo modri, vendar zaradi katastrofalnega pozabljanja klasifikacijska točnost na prvi učni podmnožici strmoglavi in se posledično to niti približno ne zgodi.

3 Zasnova eksperimentov

Za analizo katastrofalnega pozabljanja smo zasnovali globoko konvolucijsko nevronsko mrežo. Osnovni gradnik mreže je sestavljen iz konvolucije + ELU + paketne normalizacije (angl. Batch Normalization) + konvolucije + ELU + paketne normalizacije + združevanja z maksimizacijo (angl. Max Pooling) + izpadne plasti (angl. Dropout). Osnovni gradnik se ponovi trikrat, na koncu je dodana še polno povezana plast. Mrežo tako skupno sestavlja 25 plasti.

Za evalvacijo pristopov smo uporabili podatkovno zbirko CIFAR-100 [6], zanjo smo se odločili, ker ima relativno veliko primerov (60.000), dimenzije (32×32) pa niso prevelike in smo zato lahko izvedli veliko eksperimentov. Ta mreža sicer ne dosega tako dobrih rezultatov kot trenutno najboljše arhitekture, kar pa za to raziskavo ne predstavlja problema. Želimo namreč spoznati razloge za katastrofalno pozabljanje, zaradi splošnosti arhitekture pa lahko domnevamo, da se naše ugotovitve prenesejo tudi na preostale nevronske mreže.

V vseh eksperimentih učenje poteka v dveh fazah, z dvema učnima podmnožicama, ki nimata nobenih skupnih primerov. Mrežo najprej 40 epoh učimo na prvi podmnožici, nato pa začnemo z drugo fazo učenja, v kateri učimo na drugi podmnožici. V obeh fazah uporabimo optimizacijsko metodo Adam, začetna stopnja učenja znaša 0,001 in se zmanjša za faktor 10 vsakih 10 epoh. Za kriterijsko funkcijo uporabljamo križno entropijo (angl. Cross Entropy).

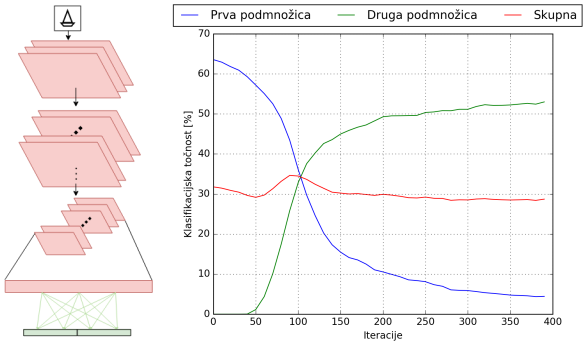
4 Eksperimenti

Zgornjo mejo za klasifikacijsko točnost, ki jo lahko dosežemo, predstavlja klasifikacijska točnost, ki jo dobimo, če mrežo učimo na celotni učni množici (svetlo modra prekinjena črta na sliki 1), znaša pa 56.7%. Tej vrednosti bi se želeli čim bolj približati.

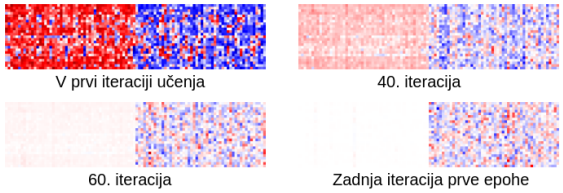
4.1 Osnovni pristop

Na sliki 1 vidimo, kako močno je katastrofalno pozabljanje, če v drugi fazi učenja ne spreminjamo stanja mreže. Zato smo najprej preverili, ali lahko katastrofalno pozabljanje zmanjšamo z zamrznitvijo vseh razen zadnje plasti v drugi fazi učenja. Ugotovili smo, da ima to na katastrofalno pozabljanje zanemarljiv vpliv, saj mreža že znotraj ene epohe druge faze učenja pozabi praktično vse prej naučeno.

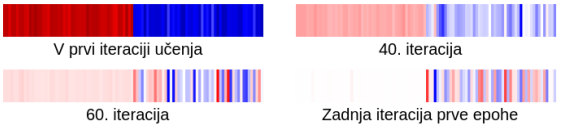
Shema levo na sliki 2 prikazuje, kateri parametri so zamrznjeni (rdeča) in kateri ne (zelena) v drugi fazi učenja. Slika 2 prikazuje, kaj se dogaja s klasifikacijskimi



Slika 2: Shema in klasifikacijske točnosti znotraj prve epohe druge faze.



(a) Uteži



(b) Odmiki

Slika 3: Spreminjanje parametrov v zadnji plasti mreže znotraj prve epohe druge faze učenja.

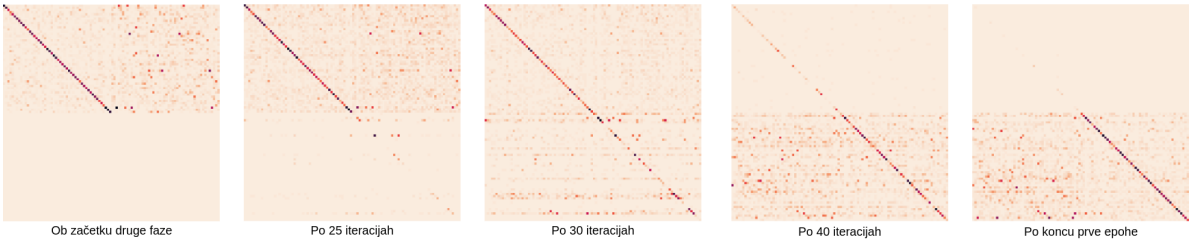
točnostmi v prvi epohi druge faze učenja. Skupna klasifikacijska točnost ob koncu prve epohe znaša 29,3% in se nato le še zmanjšuje, zgornji meji, ki znaša 56,7% in je predstavljena tudi v tabeli 1, se niti najmanj ne približa.

Matrike zamenjav na sliki 4 prikazujejo, kako mreža uvršča primere v razrede znotraj prve epohe druge faze učenja. Vidimo, da mreža začne zelo hitro prepoznavati vse primere, kot da pripadajo razredom iz druge učne podmnožice, kar je glavni razlog za padec klasifikacijske točnosti na prvi podmnožici.

Da bi bolje razumeli, zakaj pride do tega, smo preverili, kako se spreminjajo parametri (uteži in odmiki) v zadnji plasti nevronske mreže.

Na sliki 3a je prikazano, kako se spreminjajo vrednosti uteži v zadnji plasti mreže znotraj prve epohe druge faze učenja. Rdeča barva označuje zmanjšanje, modra pa povečanje vrednosti, intenziteta barve pa označuje velikost spremembe. Posamezna vrstica prikazuje uteži od enega nevrona v predzadnji plasti do vseh nevronov v zadnji plasti, posamezen stolpec pa uteži od vseh nevronov v predzadnji plasti do enega v zadnji. Vseh vrstic je dejansko 2048, kolikor je nevronov v predzadnji plasti, vendar je prikazanih le prvih 25, saj za ostale veljajo podobne zakonitosti. Opazimo, da se uteži do nevronov, ki predstavljajo razrede iz prve učne množice zelo izrazito zmanjšujejo, preostale pa zvišujejo.

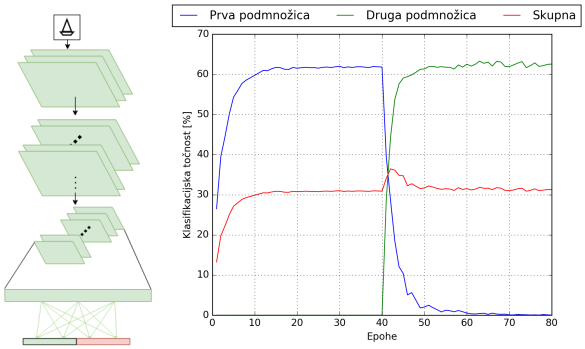
Podobno kot za uteži velja tudi za odmike v zadnji plasti, kar je prikazano na sliki 3b. Vsak stolpec predstavlja odmik enega izmed 100 nevronov v zadnji plasti.



Slika 4: Matrike zamenjav v prvi epohi druge faze učenja. Vrstica predstavlja napovedan razred, stolpec pa dejanski.

4.2 Zamrznitev zadnje plasti

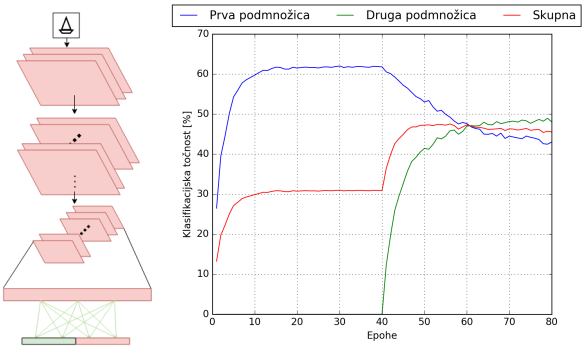
Zaradi izrazitega zmanjševanja vrednosti parametrov nevronov v zadnji plasti, ki predstavljajo razrede iz prve učne podmnožice, smo se odločili, da bomo v drugi fazi učenja zamrznili te parametre. Zamrznitev izvedemo tako, da vse gradiente iz zamrznjenih nevronov nastavimo na 0 med vzvratnim razširjanjem.



Slika 5: Zamrznitev dela zadnje plasti.

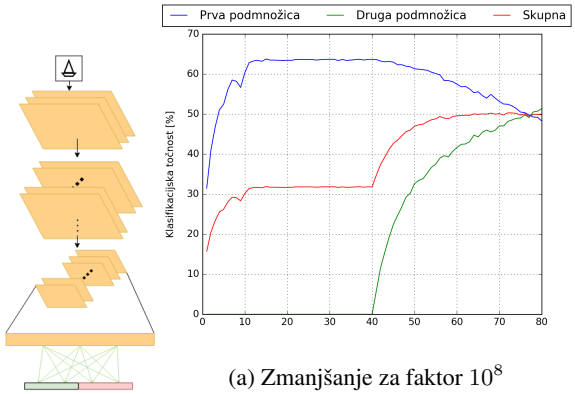
Na sliki 5 vidimo, da zamrznitev dela zadnje plasti katastrofalno pozabljanje sicer rahlo upočasni, vendar je to še vedno močno prisotno. Skupna klasifikacijska točnost doseže 36,4%, kar je še vedno daleč od zgornje meje.

Nadalje smo zamrznili tudi vse ostale plasti, tako da je učenje potekalo le na delu nevronov v zadnji plasti, ki predstavljajo razrede iz druge učne podmnožice.

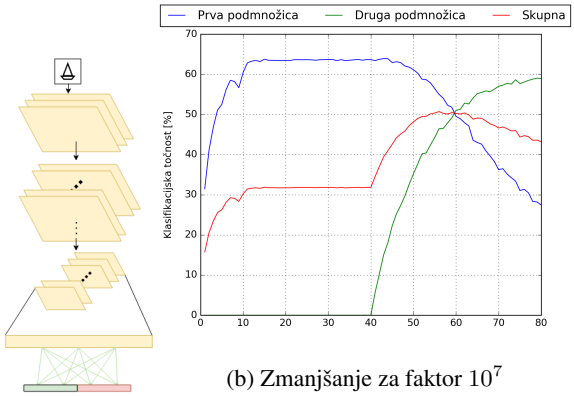


Slika 6: Zamrznitev dela zadnje plasti in vseh ostalih.

Na sliki 6 vidimo, da se ob zamrznitvi tudi vseh preostalih plasti katastrofalno pozabljanje zelo upočasni, skupna klasifikacijska točnost v drugi fazi učenja se občutno poveča in doseže 47,5%, s čimer smo mnogo bližje zgornji meji.



(a) Zmanjšanje za faktor 10^8



(b) Zmanjšanje za faktor 10^7

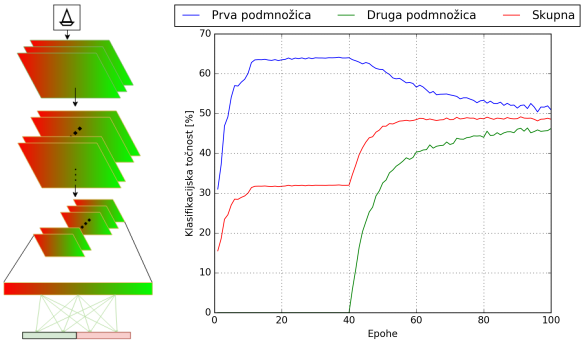
Slika 7: Variabilna stopnja učenja.

4.3 Variabilna stopnja učenja

Kombinacijo obeh verzij zamrznitve zadnje plasti predstavlja uporaba variabilne stopnje učenja. V drugi fazi učenja v zadnji plasti zamrznemo parametre nevronov, ki predstavljajo razrede iz prve učne množice, za vse preostale plasti pa uporabimo zmanjšano stopnjo učenja. Ugotovili smo, da je za izogib katastrofalnemu pozabljanju potrebno ogromno zmanjšanje, vsaj za faktor 10^7. Na slikah 7a in 7b je prikazano spreminjanje klasifikacijskih točnosti ob uporabi variabilne stopnje učenja. Rumena barva na shemah na sliki 7 označuje parametre, za katere velja zmanjšana stopnja učenja v drugi fazi učenja. S faktorjem zmanjšanja 10^8 se katastrofalno pozabljanje močno upočasni in dosežemo klasifikacijsko točnost 50,3%. Najvišjo klasifikacijsko točnost dosežemo s faktorjem zmanjšanja 10^7 in sicer 50,7%, s čimer smo od zgornje meje oddaljeni le 6 odstotnih točk. Previdni moramo sicer biti, da učenje v drugi fazi ustavimo dovolj zgodaj.

4.4 Metoda MAS

Memory Aware Synapses (MAS) [1] je ena izmed obstoječih metod za odpravljanje katastrofalnega pozabljanja. V drugi fazi učenja v kriterijsko funkcijo doda regularizacijski del, s katerim se kaznuje spremembe parametrov, ki so ocenjeni kot bolj pomembni za delovanje na prvi podmnožici. Oceno pomembnosti parametra izračunamo na prvi učni podmnožici. Avtorji predvidijo, da za vsako podmnožico naučimo ločeno zadnjo plast mreže, ki jo moramo ob uporabi ustrezno nastaviti, kar posledično pomeni, da moramo za vsak testni primer vedeti, ali pripada razredu iz prve ali druge podmnožice, kar v praksi zelo omejuje uporabo. Metodo smo priredili tako, da deluje tudi z enotno zadnjo plastjo, s čimer smo odstranili to omejitev. Pri izračunu ocen pomembnosti parametrov tako upoštevamo samo izhode nevronov v zadnji plasti, ki predstavljajo razrede iz trenutne podmnožice, ko mrežo učimo na drugi podmnožici pa zamrznemo parametre nevronov, ki predstavljajo razrede iz prve učne množice.



Slika 8: Prilagojena metoda MAS.

Rdeče-zelen gradient na shemi na sliki 8 označuje parametre, za katere velja regularizacija v drugi fazi učenja. Slika 8 prikazuje tudi dobljene klasifikacijske točnosti; z nje je razvidno, da tudi ta metoda močno zmanjša katastrofalno pozabljanje, skupna klasifikacijska točnost doseže 49,2 %. Vrednost regularizacijskega hiperparametra λ znaša 1, kot predlagajo avtorji.

Povzetek vseh dobljenih klasifikacijskih točnosti je predstavljen v tabeli 1. Referenčna skupna klasifikacijska točnost, ki jo dobimo ob učenju na celotni množici znaša 56,7%. V tabeli so tako predstavljena odstopanja od te maksimalne vrednosti.

Tabela 1: Rezultati eksperimentov. Stolpca PM 1 in PM 2 prikazujeta klasifikacijske točnosti na prvi in drugi podmnožici ob najvišji skupni. Odstopanje je izraženo v odstotnih točkah od zgornje meje, ki znaša 56.7%.

Pristop	**PM 1**	**PM 2**	**Skupna**	**Odst.**
Naivni pristop	0,1	65,7	32,9	23,8
Osnovni pristop	2,8	55,7	29.3	27,4
Zamrznitev zadnje plasti	28,6	44,2	36,4	20,3
Zamrznitev vseh plasti	49,2	45,8	47,5	9,2
Variabilna 10^8	52,5	48,2	50.3	6,4
stopnja učenja 10^7	55,0	46,4	50,7	6,0
Prilagojeni MAS	52,0	46,3	49,2	7,5

5 Zaključek

V članku smo predstavili, kateri so glavni vzroki za katastrofalno pozabljanje. Na podlagi teh ugotovitev smo zasnovali različne pristope k osveževanju parametrov nevronske mreže, ki katastrofalno pozabljanje omejujejo. Obstoječo metodo MAS smo prilagodili, da je uporabna tudi v realnem scenariju, kjer ne vemo, kateri podmnožici pripada posamezen primer. Uporaba variabilne stopnje učenja nam omogoča, da ob inkrementalnem učenju dosežemo klasifikacijsko točnost, ki je le za 6 odstotnih točk oz. 10.6% nižja, kot če model učimo na celotni množici (padec z 56,7% na 50.7%).

Avtorji v [11] v podobnem scenariju, ob uporabi zmogljivejše mreže, dosežejo padec z 68,6% na približno 62%, kar predstavlja le rahlo manjše znižanje, kot ga dobimo mi. V istem članku poročajo tudi o rezultatih, ki jih na CIFAR-100 doseže metoda predlagana v [7], padec je tam iz 68,8% na približno 53%, kar je občutno več, kot smo dosegli z evalviranim pristopom. Katastrofalno pozabljanje torej ostaja eden izmed odprtih problemov na področju globokih konvolucijskih nevronskih mrež.

Literatura

[1] Rahaf Aljundi et al. Memory aware synapses: Learning what (not) to forget. *ECCV*, pages 144–161, 2018.

[2] Francisco M. Castro et al. End-to-end incremental learning. *ECCV*, pages 2935–2947, 2018.

[3] Robert French. Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. *Proceedings of the 16th Annual Cognitive Science Society Conference*, pages 335–340, 1994.

[4] Robert M. French. Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. In *Proceedings of the 13th Annual Cognitive Science Society Conference*, pages 173–178, 1991.

[5] James Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526, 2016.

[6] Alex Krizhevsky. Learning multiple layers of features from tiny images. *University of Toronto*, 05 2012.

[7] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40:2935–2947, 2016.

[8] Arun Mallya, Dillon Davis, and Svetlana Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. *ECCV*, pages 2935–2947, 2018.

[9] Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 109 – 165. 1989.

[10] Roger Ratcliff. Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological review*, 97:285–308, 05 1990.

[11] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental Classifier and Representation Learning. *CVPR*, pages 5533–5542, 2017.