Pogled na obučavanje dubokih neuronskih mreža iz perspektive teorije informacija

Mladen Nikolić

Grupa za mašinsko učenje i primene
Matematički fakultet
Univerzitet u Beogradu

Pregled

Neobično ponašanje neuronskih mreža

Neki pojmovi teorije informacija

Informaciono usko grlo (IB)

IB i mašinsko učenje

IB i duboke neuronske mreže

Pregled

Neočekivano dobri praktični rezultati

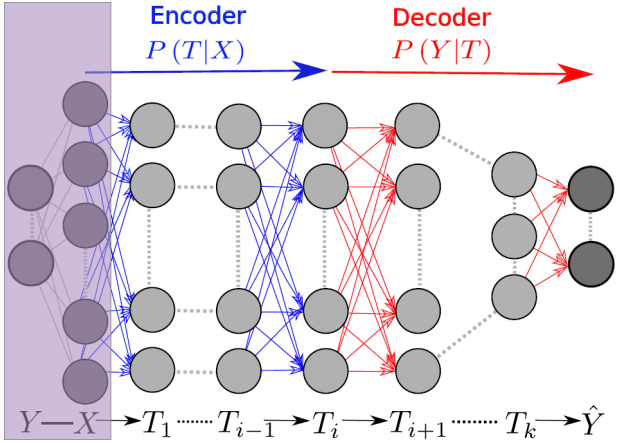
- Opšte je zapažanje da mreže daju bolje rezultate u primenama nego što bi se očekivalo na osnovu njihove nekonveksnosti i visoke fleksibilnosti
- Ovo zapažanje je pokrenulo istraživanja u različitim pravcima
- Jedan pravac je ispitivanje svojstava lokalnih minimuma greške neuronske mreže
- Drugi je ispitivanje njihovog kapaciteta za preprilagođavanje

Neobična otpornost na preprilagođavanje

- Eksperimentalno je pokazano da se velike neuronske mreže mogu prilagoditi trening skupu sa greškom 0, a ipak imati značajnu moć generalizacije
- Ovo je neintuitivno i traži objašnjenje
- Model se dobija kroz optimizacioni proces. Šta se u njemu dešava?
- Neka od skorašnjih zapažanja dolaze iz perspektive teorije informacija, ali postoje i druga

Skriveni slojevi kao nove promenljive

- Duboke neuronske mreže se sastoje od slojeva neurona koji konstruišu nove reprezentacije ulaza, a na izlazu predviđaju ciljnu promenljivu



Pregled

Informacija

- Kada neko saznanje smatramo informativnim?

Informacija

- Kada neko saznanje smatramo informativnim?
- Kada postoje alternative tom saznanju, odnosno neizvesnost koja biva umanjena tim saznanjem
- Možemo razmatrati izvođenje nekog eksperimenta, donošenje izbora od strane druge osobe, itd.
- U svakom slučaju informacija je tesno povezana sa neizvesnošću ishoda

Kvantifikovanje količine informacije

- Može li se količina informacije/neizvesnost kvantifikovati?
- Eksperiment ima n ishoda sa verovatnoćama $p_1, \ldots, p_n$
- Količina informacije koja se saopštava otkrivanjem ishoda koji se desio zavisi od datih verovatnoća
- Što je neizvesnost veća, veća je količina informacije saopštena otkrivanjem ishoda

Kvantifikovanje količine informacije

- Verovatnoće ishoda u slučaju fer novčića su 0.5 i 0.5
- Koliko bitova informacije je potrebno da bi se saopštio ishod?

Kvantifikovanje količine informacije

- Verovatnoće ishoda u slučaju fer novčića su 0.5 i 0.5
- Koliko bitova informacije je potrebno da bi se saopštio ishod?
- A ako su verovatnoće ishoda 0 i 1?

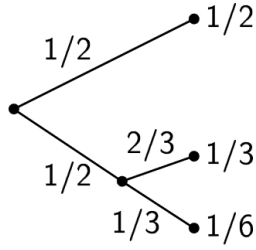
Kvantifikovanje količine informacije

- Verovatnoće ishoda u slučaju fer novčića su 0.5 i 0.5
- Koliko bitova informacije je potrebno da bi se saopštio ishod?
- A ako su verovatnoće ishoda 0 i 1?
- A 0.25 i 0.75?

Kvantifikovanje neizvesnosti

- Neka je $H(p_1, \ldots, p_n)$ neizvesnost vezana za eksperiment sa datim verovatnoćama ishoda
- Šta očekujemo od funkcije H?
  - H treba da bude neprekidna
  - Ako su sve verovatnoće jednake, H treba monotono da raste sa n
  - Ako se ishodi zamene novim eksperimentima, H treba da bude težinska suma vrednosti H za te eksperimente

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right)$$



Entropija

- Funkcija koja zadovoljava prethodne zahteve je jedinstveno određena do na multiplikativnu konstantu i naziva se *entropija*

$$H(p_1, \ldots, p_n) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

- Nadalje ne pišemo osnovu logaritma
- Ako je $p_i = 1$ za neko i, onda je $H(p_1, \ldots, p_n) = 0$
- $H(p_1, \ldots, p_n)$ je maksimalno kada su sve verovatnoće p_i jednake

Entropija

- Neka je p raspodela. Onda:

$$H(p) = -\sum_{i=1}^{\infty} p(x_i) \log p(x_i)$$

$$H(p) = -\int p(x) \log p(x) dx$$

(u neprekidnom slučaju osobine mogu biti drugačije nego u diskretnom!)

Entropija

- $-\log p(x)$ se može interpretirati kao količina informacije koju nosi ishod x
- Entropija je očekivana količina informacije slučajne promenljive X sa raspodelom p
- Dužina binarnog kodiranja ishoda sa raspodelom p pri optimalnom kodiranju je između $H(p)$ i $H(p) + 1$
- Meri se u bitovima

Entropija - kodiranje engleskog teksta

- U slučaju da se sva slova i razmak smatraju jednako verovatnim, entropija je $-\sum_{i=1}^{27} \frac{1}{27} \log \frac{1}{27} \approx 4.75$
- Tekst dužine n se onda kodira pomoću $5n$ karaktera
- Ipak, stvarne frekvencije karaktera su različite: $f_e = 0.1270$, $f_t = 0.0906$, ..., $f_z = 0.0007$
- Ako se konstruiše kodiranje koje će dodeliti najkraći kod karakteru e, a najduži karakteru z, prilazi se dužini od $4.22n$
- Ako se uzmu u obzir i zavisnosti između karkatera (npr. q se ne pojavljuje posle z), mogu se dobiti i kraća kodiranja
- Procenjuje se da je u engleskom entropija slova između 0.6 i 1.3

Unakrsna/relativna entropija (cross/relative entropy)

- Šta ako ishode sa raspodelom p kodiramo pomoću optimalnog koda za raspodelu q?

$$H(p, q) = -\int p(x) \log q(x) dx$$

Kulbek-Lajblerovo odstupanje (Kullback-Leibler divergence)

- Koliko bitova informacije se nepotrebno troši ukoliko se za kodiranje ishoda slučajne promenljive sa raspodelom p koristi kod optimalan za promenljivu sa raspodelom q?

$$D_{KL}(p||q) = -\int p(x) \log q(x) dx - H(p) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Uzajamna informacija

- Kako meriti medjusobnu zavisnost dve promenljive X i Y?

$$I(X;Y) = D_{KL}(p_{xy}||p_x p_y) = \int\int p_{xy}(x,y) \log \frac{p_{xy}(x,y)}{p_x(x)p_y(y)} dxdy$$

$$= \int\int p_{xy}(x,y) \log \frac{p_{xy}(y|x)}{p_y(y)} dxdy$$

Uzajamna informacija

- Kako meriti medjusobnu zavisnost dve promenljive X i Y?

$$I(X;Y) = D_{KL}(p_{xy}||p_x p_y) = \int\int p_{xy}(x,y) \log \frac{p_{xy}(x,y)}{p_x(x)p_y(y)} dxdy$$

$$= \int\int p_{xy}(x,y) \log \frac{p_{xy}(y|x)}{p_y(y)} dxdy$$

- Zašto ne Pirsonov koeficijent korelacije?

Markovljev lanac

- Niz promenljivih $X_1, X_2 \ldots$ čini Markovljev lanac ukoliko važi

$$p(X_{n+1}|X_n, \ldots, X_1) = p(X_{n+1}|X_n)$$

- Za Markovljev lanac važi nejednakost obrade informacija

$$I(X_i; X_{i+1}) \geq I(X_i; X_{i+2})$$

Pregled

Relevantna informacija

- Ako je dat vektor promenljivih X, šta čini relevantnu informaciju u njima?
- U odnosu na šta?
- Recimo u odnosu na drugu promenljivu Y
- Jedan odgovor je koncept minimalne dovoljne statistike

Dovoljna statistika

- Ukoliko za statistiku T važi $p(Y|T(x), x) = p(Y|T(x))$, onda je ona dovoljna statistika za Y
- Intuitivno, poznavanje celokupnih podataka x ne daje ništa više informacija o Y nego poznavanje vrednosti dovoljne statistike na tim podacima
- Može se pokazati $I(Y; x) = I(Y; T(x))$
- Za normalnu raspodelu $\mathcal{N}(0, \sigma^2)$ sledeće statistike su dovoljne
  - $T(x_1, \ldots, x_n) = (x_1, \ldots, x_n)$
  - $T(x_1, \ldots, x_n) = (x_1^2, \ldots, x_n^2)$
  - $T(x_1, \ldots, x_n) = \left(\sum_{i=1}^{m} x_i^2, \sum_{i=m+1}^{n} x_i^2\right)$
  - $T(x_1, \ldots, x_n) = \sum_{i=1}^{n} x_i^2$

Minimalna dovoljna statistika

- Dovoljna statistika T je minimalna ukoliko za svaku drugu dovoljnu statistiku T' postoji funkcija f tako da važi $T(x) = f(T'(x))$
- Intuitivno, minimalnom je čini to što se može izračunati iz bilo koje druge dovoljne statistike - druge statistike možda nose više informacije koja nije bitna za Y i koja se može izgubiti u ovom izračunavanju
- Time se maksimalno kompresuje relevantna informacija iz x
- Prema nejednakosti obrade infromacije važi $I(T(x); x) \leq I(T'(x); x)$
- Minimalna dovoljna statistika za Y sadrži jednako informacija o Y koliko i x, ali sadrži najmanje informacije o x!

Minimalna dovoljna statistika

- Pokazano je da minimalna dovoljna statistika u slučaju mnogih raspodela ne postoji
- Da li je moguće definisati približnu minimalnost?

Informaciono usko grlo (IB)

- Potrebno je naći reprezentaciju T za X koja nosi što više informacije o Y, a što manje o X
- Potrebno je naći minimum funkcionala

$$\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)$$

Rešenje IB problema

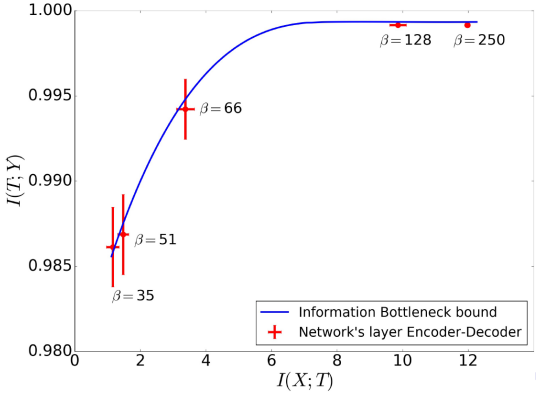
- Pod pretpostavkom Markovljevog lanca $Y \to X \to T$, za dato $p(x, y)$, postoji fromulacija problema u vidu sistema jednačina:

$$\begin{cases} p(t|x) = \frac{p(t)}{Z(x;\beta)} \exp\left(-\beta D_{KL}(p(y|x)||p(y|t))\right) \\ p(t) = \sum_x p(t|x)p(x) \\ p(y|t) = \sum_x p(y|x)p(x|t) \end{cases}$$

- Rešenje se može oceniti iterativno, ali nije jedinstveno

Informaciona ravan i IB kriva

- Informaciona ravan je određena koordinatama $(I(T;X), I(T;Y))$
- Za dato $p(x,y)$, svakoj raspodeli $p(t|x)$ odgovara jedna tačka u ravni
- Informaciona ravan je podeljena na dva regiona konkavnom krivom na kojoj leže optimalna rešenja za različite vrednosti β
- Tačke iznad krive nisu dopustive, dok one ispod jesu



Pregled

Veza sa mašinskim učenjem

- Na šta liče delovi funkcionala $\mathcal{L}$ u terminima mašinskog učenja?

$$\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)$$

Veza sa mašinskim učenjem

- Na šta liče delovi funkcionala $\mathcal{L}$ u terminima mašinskog učenja?

$$\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)$$

- Greška modela $(-I(T;Y))$ i regularizacija $(I(T;X))$!
- Šta su ishodi pri variranju β od 0 ka $+\infty$?

Veza sa mašinskim učenjem

- Na šta liče delovi funkcionala $\mathcal{L}$ u terminima mašinskog učenja?

$$\mathcal{L}[p(t|x)] = I(T;X) - \beta I(T;Y)$$

- Greška modela ($-I(T;Y)$) i regularizacija ($I(T;X)$)!
- Šta su ishodi pri variranju β od 0 ka $+\infty$?
- Ako važi $\beta = 0$, onda je gdeška koja se pravi nebitna i rešenje može biti trivijalno
- Kako se β povećava, rešenje se kreće ka $T = X$
- Koliko daleko seže ova analogija sa greškom i regularizacijom?

Rizik i empirijski rizik

- U mašinskom učenju, greška modela se formalizuje rizikom

$$R(f) = \int L(f(x), y)p(x, y)dxdy$$

- Zbog nepoznavanja raspodele $p(x, y)$, on se aproksimira empirijskim rizikom

$$E(f) = \frac{1}{n}\sum_{i=1}^{n} L(f(x_i), y_i)$$

na nekom uzorku $\mathcal{D} = \{(x_i, y_i)|i = 1, \ldots, n\}$

Granica greške

- Može se dokazati da sa verovatnoćom μ važi

$$R(f) \leq E(f) + c(n\downarrow, \mu\uparrow, h\uparrow)$$

gde je h Vapnik-Červonenkisova dimenzija

Empirijska ocena funkcionala $\mathcal{L}$

- Izračunavanje optimalnog rešenja za minimizaciju funkcionala $\mathcal{L}$ zahteva poznavanje raspodele $p(x, y)$
- Ona nije poznata, ali je moguće oceniti je pomoću neke empirijske ocene $\hat{p}(x, y)$, kojoj odgovaraju veličine $\hat{I}(T; X)$ i $\hat{I}(T; Y)$
- Ako su $\mathcal{T}$ i $\mathcal{Y}$ konačni domeni promenljivih T i Y, važi:

$$I(T; Y) \leq \hat{I}(T; Y) + O\left(\frac{|\mathcal{T}||\mathcal{Y}|}{\sqrt{n}}\right)$$

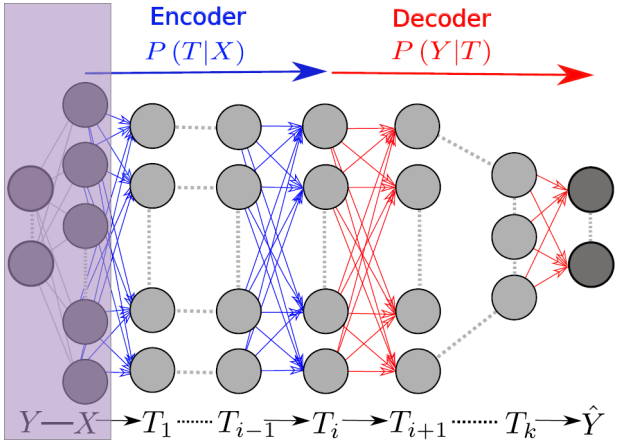
$$I(T; X) \leq \hat{I}(T; X) + O\left(\frac{|\mathcal{T}|}{\sqrt{n}}\right)$$

- Štaviše, preciznost klasifikacije se može odozgo ograničiti veličinom $\exp(-O(I(T; Y)))$

Pregled

Neuronske mreže kao Markovljevi procesi

- Neuronska mreža se može smatrati Markovljevim procesom



Neuronske mreže kao Markovljevi procesi

- Zahvaljujući tome na osnovu nejednakosti obrade podataka važi

$$I(X;Y) \geq I(T_1;Y) \geq \ldots I(T_k;Y) \geq I(\hat{Y};Y)$$

  pri čemu jednakost važi samo ako je svaki sloj dovoljna statistika za Y
- Otud je poželjno za svaki sloj neuronske mreže maksimizovati veličine $I(T_i;Y)$, pritom minimizujući veličine $I(T_i;T_{i-1})$
- Teorijski moguće izvesti algoritmom informacionog uskog grla, što je u praksi previše neefikasno
- Da li je moguće i na neki efikasniji način?

Pitanja vezana za neuronskre mreže

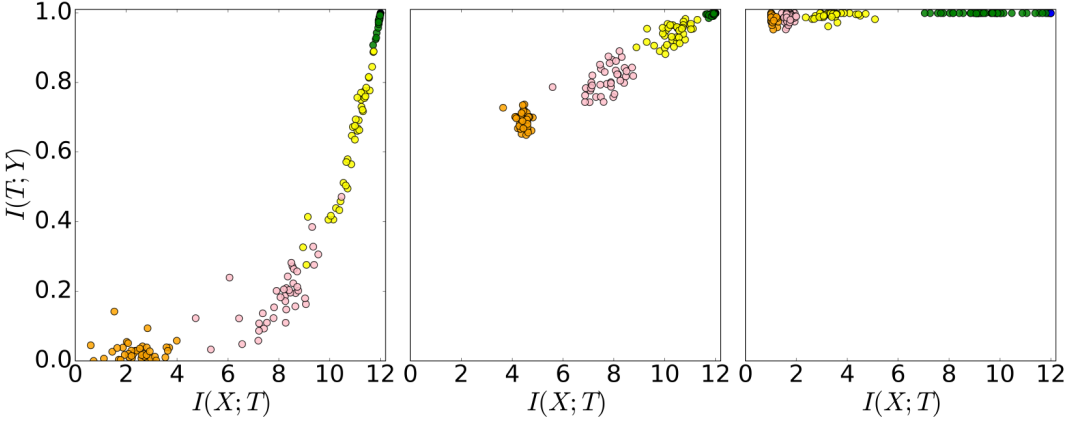
- Kako se u informacionoj ravni ponaša stohastički gradijentni spust?
- Koja je korist od skrivenih slojeva?
- Da li skriveni slojevi čine optimalne reprezentacije u smislu IB?
- Kako se uvidi dobijeni kroz teoriju informacija mogu upotrebiti u praksi?

Eksperimentalna postavka

- Sintetički podaci, kako bi bila poznata raspodela $p(x, y)$ (zajedničku informaciju je moguće i oceniti bez poznavanja $p(x, y)$, ali je to dosta teže)
- Klasifikacioni problem
- Aktivaciona funkcija je tangens hiperbolički, osim poslednjeg sigmoidnog sloja
- Nema regularizacije!
- Koristi se diskretizacija izlaza neurona pri računanju zajedničke informacije
- Eksperimenti su pokretani po 50 puta za različite inicijalizacije parametara mreže i različite trening podatke
- Autori su uvereni da su im zaključci opšti :)

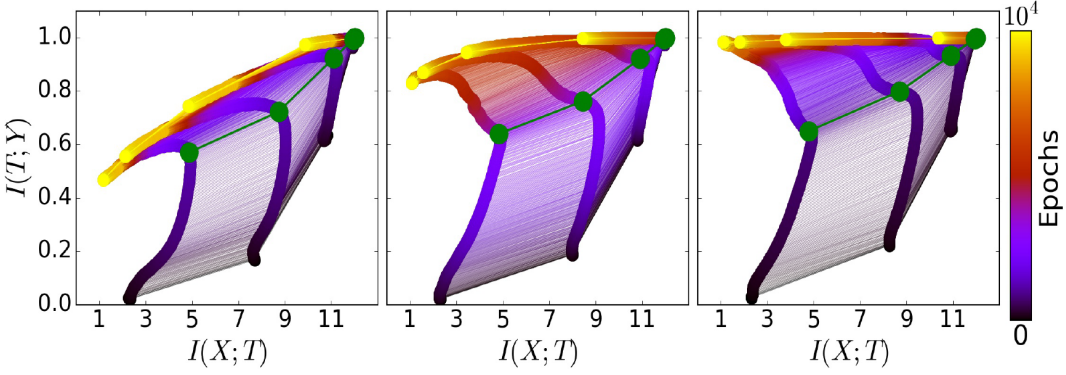
Trening stohastičkim gradijentnim spustom

- Skupovi tačaka iste boje predstavljaju odgovarajuće slojeve u različitim mrežama



Trening stohastičkim gradijentnim spustom

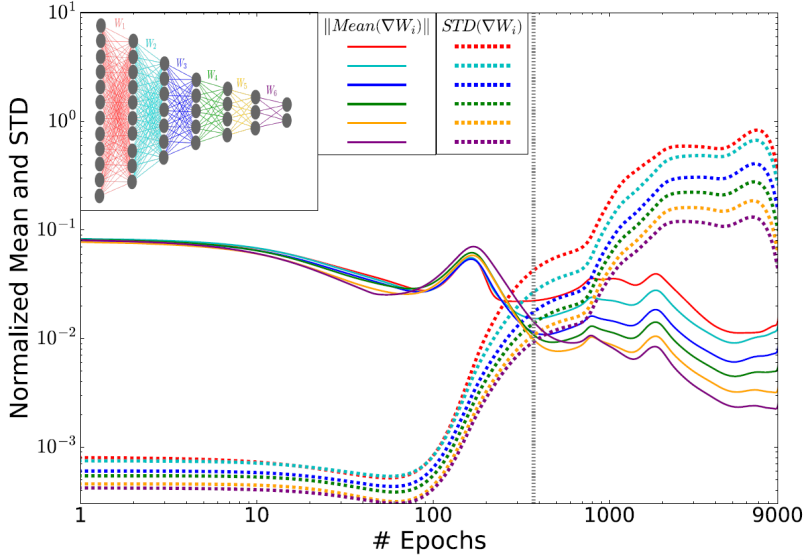
- Putanje uprosečenih pozicija različitih slojeva u toku optimizacije sa 5%, 45% i 85% podataka



Dve faze treninga

- Primećuju se dve faze treninga: kratka faza minimizacije greške i duga faza kompresije
- Faza minimizacije greške je očekivana
- Faza kompresije bez ikakve regularizacije je iznenađujuća i traži objašnjenje

Norme i standardne devijacije gradijenata

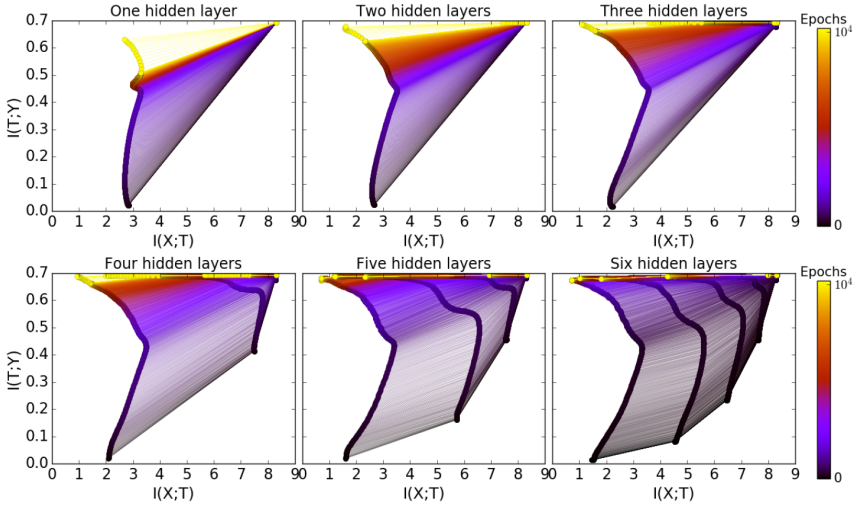


Norme i standardne devijacije gradijenata

- Gradijenti prvo imaju velike norme, a malo osciluju, a potom se norme smanjuju, a više osciluju
- Ove dve faze se podudaraju sa fazama minimizacije greške i kompresije
- Ovakav prelaz je uočljiv i u praktičnim primenama, što sugeriše da i u njima postoje faze minimizacije greške i kompresije

Efekat dodavanja skrivenih slojeva

- Prikaz treninga za mreže sa jednim do šest skrivenih slojeva

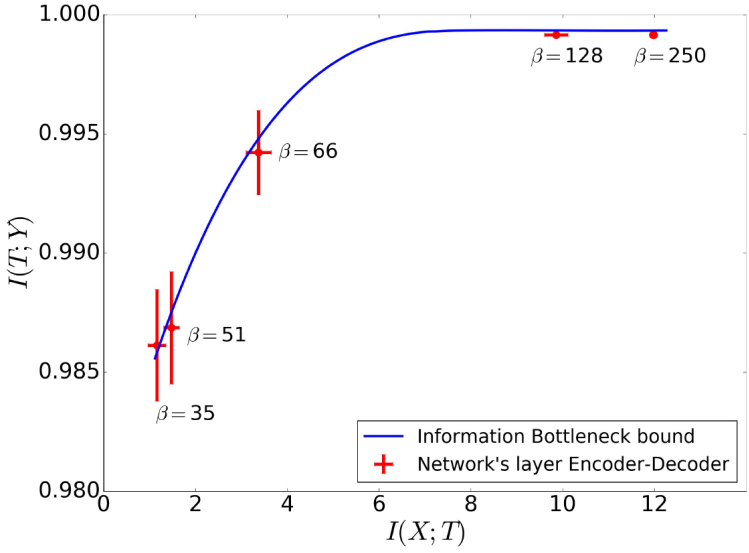


Efekat dodavanja skrivenih slojeva

- Sa svakim skrivenim slojem, konvergencija je sve brža
- Ovo se objašnjava smanjenjem trajanja faze kompresije za svaki sloj, pošto se svaki sloj naslanja na već donekle kompresovani prethodni sloj

Poređenje sa IB krivom

- Skriveni slojevi su praktično optimalni u smislu IB



Praktične posledice

- Stohastički gradijentni spust troši većinu vremena u kompresiji
- Postoji mogućnost da bi drugi algoritmi u tome bili efikasniji (već su poznati kandidati)
- Ukoliko slojevi u opštem slučaju konvergiraju ka IB krivoj, kao što deluje, to znači da među njima važe analitički poznate veze na osnovu IB rešenja koje bi se mogle upotrebiti u treningu
- Na ovome se radi

Zaključci

- Trening neuronskih mreža se sastoji od kratke faze minimizacije greške i duge faze kompresije
- Efikasno se konstruišu reprezentacije koje predstavljaju minimalne statistike u smislu IB
- Visoke performanse su posledica optimalnosti slojeva u smislu IB
- Postoji mogućnost da se ubrza trening zahvaljujući ovim uvidima

Literatura

- N. Tishby, F. Pereira, W. Bialek, The Information Bottleneck Method,
- O. Shamir, S. Sabato, N. Tishby, Learning and Generalization with the Information Bottleneck
- N. Tishby, N. Zaslavsky, Deep Learning and the Information Bottleneck Principle
- R. Schwartz-Ziv, N. Tishby, Opening the Black Box of Deep Neural Networks via Information
- C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding Deep Learning Requires Rethinking Generalization

HVALA NA PAŽNJI!