Učenje dubokih reprezentacija ocenom i maksimizacijom uzajamne informacije

Mladen Nikolić

Grupa za mašinsko učenje i primene
Matematički fakultet
Univerzitet u Beogradu

Pregled

Neki pojmovi teorije informacija

Ocena uzajamne informacije neuronskom mrežom

Učenje dubokih reprezentacija - DeepINFOMAX

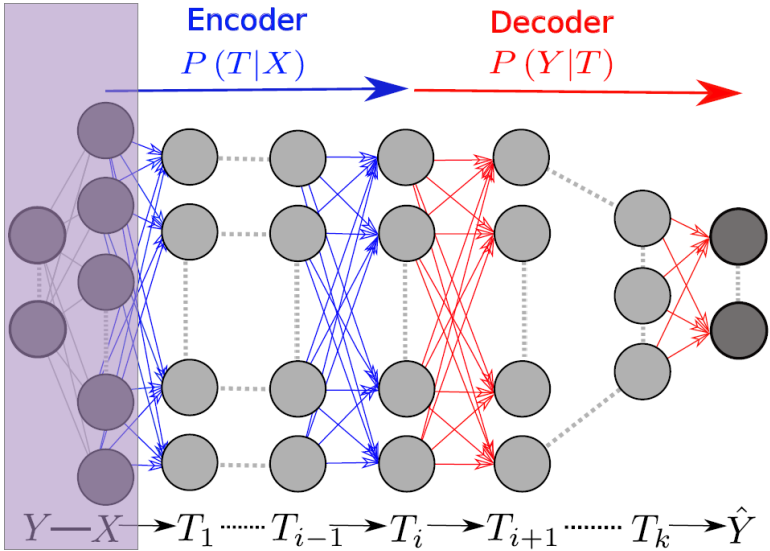
Motivacija

- Kada razmišljamo i donosimo odluke, retko se izražavamo na senzornom nivou (npr. pikseli), već na nivou apstraktnih reprezentacija
- Većina bitova informacije na senzornom nivou može biti beznačajna na semantičkom nivou (npr. prepoznavanje govornika)
- Ipak, funkcije cilja algoritama nenadgledanog učenja se tipično izražavaju u terminima ulaza
- Primarni cilj je definisati metod nenadgledanog učenja reprezentacije u terminima naučene reprezentacije

Ideje

- Ključna ideja je učiti reprezentaciju (tj. enkoder) maksimizujući uzajamnu informaciju između ulaza i reprezentacije
- Dodatno, bitna je mogućnost nametanja neke strukture reprezentacije pomoću apriornih raspodela
- Ukoliko se ne cilja na rekonstrukciju celog ulaza, možda je bolje maksimizovati uzajamnu informaciju reprezentacije sa delovima ulaza

Enkoder



Pregled

Neki pojmovi teorije informacija

Ocena uzajamne informacije neuronskom mrežom

Učenje dubokih reprezentacija - DeepINFOMAX

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, pravljenje 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?

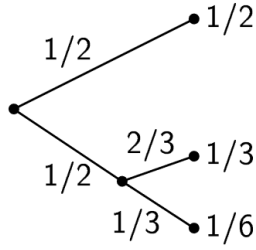
Kvantifikovanje količine informacije

- 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?

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

Uzajamna informacija

- Kako meriti medjusobnu zavisnost dve promenljive X i Y?

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

- Zašto ne Pirsonov koeficijent korelacije?

Jensen-Šenonovo odstupanje

$$D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m)$$

$$m = \frac{p+q}{2}$$

- Simetrično je
- Ograničeno odozgo sa $\ln 2$
- Neka je Z fer novčić, a X promenljiva koja na osnovu vrednosti Z daje vrednost prema raspodeli p, odnosno q. Tada važi:

$$I(X, Z) = D_{JS}(p||q)$$

Teorija informacija i neuronske mreže

- Informaciono usko grlo je pristup obučavanju neuronskih mreža pomoću uzajamne informacije
- Neki vidovi regularizacije neuronskih mreža počivaju na uzajamnoj informaciji
- Ključni problem - neefikasno izračunavanje nad neprekidnim visokodimenzionalnim promenljivim

Pregled

Neki pojmovi teorije informacija

Ocena uzajamne informacije neuronskom mrežom

Učenje dubokih reprezentacija - DeepINFOMAX

Donsker-Varadanova reprezentacija KL odstupanja

- Važi:

$$D_{KL}(p||q) = \sup_{T:\Omega\to\mathbb{R}} \mathbb{E}_p[T] - \log(\mathbb{E}_q[e^T])$$

  gde je supremum po svim funkcijama T takvim da su oba očekivanja konačna
- Ako je $\mathcal{F}$ bilo koji skup funkcija koji zadovoljava te uslove integrabilnosti važi

$$D_{KL}(p||q) \geq \sup_{T\in\mathcal{F}} \mathbb{E}_p[T] - \log(\mathbb{E}_q[e^T])$$

Ocena neuronskom mrežom

- Neka je skup $\mathcal{F}$ skup svih funkcija koje definiše neuronska mreža $T_\theta : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$ sa parametrima $\theta \in \Theta$
- Neka je

$$I_\Theta(X, Z) = \sup_{\theta \in \Theta} \mathbb{E}_{p_{xz}}[T_\theta] - \log(\mathbb{E}_{p_x p_z}[e^{T_\theta}])$$

  pri čemu se u praksi koristi ocena $\hat{I}_\Theta(X, Z)$ u kojoj se očekivanja ocenjuju empirijski
- Pronalaženje supremuma se realizuje gradijentnim usponom

Algoritam MINE

Algorithm 1 MINE

$\theta \leftarrow$ initialize network parameters

repeat

Draw b minibatch samples from the joint distribution:
$(\boldsymbol{x}^{(1)}, \boldsymbol{z}^{(1)}), \ldots, (\boldsymbol{x}^{(b)}, \boldsymbol{z}^{(b)}) \sim \mathbb{P}_{XZ}$

Draw n samples from the Z marginal distribution:
$\bar{\boldsymbol{z}}^{(1)}, \ldots, \bar{\boldsymbol{z}}^{(b)} \sim \mathbb{P}_Z$

Evaluate the lower-bound:
$\mathcal{V}(\theta) \leftarrow \frac{1}{b}\sum_{i=1}^{b} T_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{z}^{(i)}) - \log(\frac{1}{b}\sum_{i=1}^{b} e^{T_\theta(\boldsymbol{x}^{(i)}, \bar{\boldsymbol{z}}^{(i)})})$

Evaluate bias corrected gradients (e.g., moving average):
$\widehat{G}(\theta) \leftarrow \widetilde{\nabla}_\theta \mathcal{V}(\theta)$

Update the statistics network parameters:
$\theta \leftarrow \theta + \widehat{G}(\theta)$

until convergence

Svojstva MINE ocene

- MINE pruža jako konzistentnu ocenu uzajamne informacije
- Pod određenim pretpostavkama, za aproksimaciju u d dimenzionalnom prostoru, sa greškom ε sa verovatnoćom $1-\delta$, potrebno je $O\left(\frac{d\log d+\log(2/\delta)}{\varepsilon^2}\right)$ podataka

Primene

- Izbegavanje gubljenja modova kod generativnih suparničkih modela
- Unapređivanje dvosmernih suparničkih modela
- Implementacija informacionog uskog grla

Pregled

Neki pojmovi teorije informacija

Ocena uzajamne informacije neuronskom mrežom

Učenje dubokih reprezentacija - DeepINFOMAX

Maksimizacija uzajamne informacije

- Reprezentacija treba da čuva informaciju koja se nalazi u ulazima
- T_w je neuronska mreža koja služi za ocenu uzajamne informacije
- $E_\psi : \mathcal{X} \to \mathcal{Y}$ je parametrizovana familija enkodera
- Uzajamna informacija se istovremeno ocenjuje i maksimizuje:

$$(\hat{w}, \hat{\psi}) = \arg\max_{w,\psi} \hat{I}_w(X, E_\psi(X))$$

- Kako i T_w i E_ψ zahtevaju konvolutivne slojeve, oni su deljeni i važi

$$E_\psi = f_\psi \circ C_\psi \qquad T_{w,\psi} = D_w \circ [C_\psi, E_\psi]$$

Formulacija zasnovana na D_{JS}

- Ocena zasnovana na D_{JS} se pokazala bolje od ocene zasnovane na D_{KL}:

$$\hat{I}_w^{JS}(X, E_\psi(X)) = \mathbb{E}_{x\sim p}[-sp(-T_{w,\psi}(x, E_\psi(x)))] - \mathbb{E}_{x,x'\sim p}[sp(T_{w,\psi}(x', E_\psi(x)))]$$

  gde je $sp(z) = \log(1 + e^z)$
- Pretpostavlja se da su performanse bolje zbog ograničenosti D_{JS}, što verovatno umanjuje problem eksplodirajućih gradijenata
- Rezultati se daju za ovu formulaciju

Maksimizacija loklane informacije (1)

- Umesto da se maksimizuje uzajamna informacija sa celim ulazom, može se maksimizovati prosečna uzajamna informacija reprezentacije sa njegovim delovima
- Na ovaj način se preferira informacija koja je deljena među različitim delovima ulaza
- Ukoliko enkoder propušta informaciju koja je specifična za neki deo ulaza, to neće značajno povećati ovako definisan cilj

Maksimizacija loklane informacije (2)

- Neka je $C_\psi(x) = \{C^i_\psi(x)\}_{i=1}^{M\times M}$ koja zadržava strukturalnu (npr. prostornu) informaciju podataka, a i indeksira po delovima ulaza
- Neka je

$$T^i_{w,\psi}(x, E_\psi(x)) = D_w([C^i_\psi(x), E_\psi(x)])$$

$$T^i_{w,\psi}(x', E_\psi(x)) = D_w([C^i_\psi(x'), E_\psi(x)])$$

- Nova funkcija cilja:

$$\frac{1}{M^2}\sum_{i=1}^{M^2} \mathbb{E}_{p_{xz}}[-sp(-T^i_{w,\psi}(x, E_\psi(x)))] - \mathbb{E}_{p_x p_z}[sp(T^i_{w,\psi}(x', E_\psi(x)))]$$

Nametanje apriorne raspodele reprezentaciji

- Često je poželjno da reprezentacije prate određenu raspodelu, da su koordinate nezavisne i slično
- Pretpostavka je da je data apriorna raspodela q koju $E_\psi(x)$ treba da prati ako je $x \sim p$
- Usvaja se suparnički pristup i obučava se funkcija $D_\phi : \mathcal{Y} \to \mathbb{R}$ koja ocenjuje odstupanje q i raspodele izalaza $E_\psi(x)$

$$(\hat{w}, \hat{\psi}) = \arg\max_{\psi} \arg\min_{\phi} [\mathbb{E}_q[logD_\phi(y)] + \mathbb{E}_p[\log(1 - D_\phi(E_\psi(x)))]]$$

- Ukupni cilj se dobija sabiranjem globalne informacije, lokalne informacije i regularizacije apriornom raspodelom

Podaci za evaluaciju

- CIFAR10 i CIFAR100: mali skupovi od 6000 i 600 slika rezolucije 32×32 iz 10, odnosno 100 klasa
- Tiny ImageNet: 100,000 slika rezolucije 64×64 iz 200 klasa
- STL-10: slike rezolucije 96×96 sa 100,000 neoznačenih podataka i po 500 slika po svakoj od 10 klasa

Korišćena arhitektura

- Za CIFAR skupove mreža je bila slična DCGAN mreži, a za ostale je korišćen AlexNet
- ReLU aktivacije i *batch normalization* na svim skrivenim slojeivma, a sigmoidne aktivacije na izlazu
- 64-dimenzionalan izlaz enkodera
- Za funkcije $T_{w,\psi}$ i D_ϕ korišćena su po dva skrivena sloja i od 200 do 1000 skrivenih jedinica
- DIM(G) je arhitektura sa globalnom ciljnom funkcijom, a DIM(L) sa lokalnom
- Apriorna raspodela uniformna na $[0,1]^{64}$

Alternativni modeli

- Varijacioni autoenkoder
- Suparnički autoenkoder
- BiGAN
- Noise as targets

Principi evaluacije

- Korisnost pristupa se meri posredno kroz njihovu uspešnost u sledećim poslovima:
  - Linearnoj klasifikaciji SVM-om (CIFAR10)
  - Nelinearnoj klasifikaciji pomoću mreže sa jednim skrivenim slojem od 200 neurona (svi skupovi)
  - Polunadgledanom učenju obučavanjem jednog sloja nad enkoderom (STL-10)
  - Merenju sličnosti slika (na osnovu postojećeg pristupa) (CIFAR10)
- Neposredno se meri uzajamna informacija MINE-om (CIFAR10) i zavisnost koordinata NDM-om

Rezultati nelinearne klasifikacije (1)

Model	CIFAR10			CIFAR100		
	conv	fc (1024)	$Y(64)$	conv	fc (1024)	$Y(64)$
Fully supervised	75.39			42.27		
VAE	60.71	60.54	54.61	37.21	34.05	24.22
AAE	59.44	57.19	52.81	36.22	33.38	23.25
BiGAN	62.57	62.74	52.54	37.59	33.34	21.49
NAT	56.19	51.29	31.16	29.18	24.57	9.72
DIM(G)	52.2	52.84	43.17	27.68	24.35	19.98
DIM(L)	**70.1**	**70.21**	**63.97**	**48.46**	**46.09**	**36.51**

Rezultati nelinearne klasifikacije (2)

	Tiny ImageNet			STL-10 (random crop pretraining)			
	conv	fc (4096)	$Y(64)$	conv	fc (4096)	$Y(64)$	SS
Fully supervised	36.60			68.7			
VAE	18.63	16.88	11.93	58.27	56.72	46.47	68.65
AAE	18.04	17.27	11.49	59.54	54.47	43.89	64.15
BiGAN	24.38	20.21	13.06	**71.53**	**67.18**	58.48	74.77
NAT	13.70	11.62	1.20	64.32	61.43	48.84	70.75
DIM(G)	11.32	6.34	4.95	42.03	30.82	28.09	51.36
DIM(L)	**33.8**	**34.5**	**30.7**	**71.82**	**67.22**	**61.61**	**75.62**

Rezultati ostalih eksperimenata

Model	Proxies				Neural Estimators	
	SVM (conv)	SVM (fc)	SVM (Y)	MS-SSIM	$\widehat{I}_\rho(X,Y)$	NDM
VAE	53.83 ± 0.62	42.14 ± 3.69	39.59 ± 0.01	0.72	93.02	1.62
AAE	55.22 ± 0.06	43.34 ± 1.10	37.76 ± 0.18	0.67	87.48	0.03
BiGAN	56.40 ± 1.12	38.42 ± 6.86	44.90 ± 0.13	0.46	37.69	24.49
NAT	48.62 ± 0.02	42.63 ± 3.69	39.59 ± 0.01	0.29	6.04	0.02
DIM(G)	46.8 ± 2.29	28.79 ± 7.29	29.08 ± 0.24	0.49	49.63	$0.35(9.96)$
DIM(L+G)	57.55 ± 1.442	45.56 ± 4.18	18.63 ± 4.79	0.53	101.65	$0.5(22.89)$
DIM(L)	63.25 ± 0.86	54.06 ± 3.6	49.62 ± 0.3	0.37	45.09	$0.18(9.18)$

Upiti najbližim susedima

Query DIM(G) DIM(L)

Efekti komponenti ciljne funkcije

- Kvalitet rekonstrukcije visoko zavisi od globalne uzajamna informacije
- Kvalitet klasifikacije visoko zavisi lokalne uzajamne informacije
- Apriorna raspodela vodi manjoj zavisnosti među koordinatama nove reprezentacije u malo popravlja kvalitet klasifikacije

Literatura

- M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, R. D. Hjelm, Mutual Information Neural Estimation
- R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, A. Trischler, Y. Bengio, Learning Deep Rerpresentations by Mutual Information Estimation and Maximization

HVALA NA PAŽNJI!