Gradientové učenie hlbokých neurónových sietí

Truc Lam, Bui
Školiteľ: RNDr. Kristína Malinovská, PhD.

27. júna 2018

Ciele práce

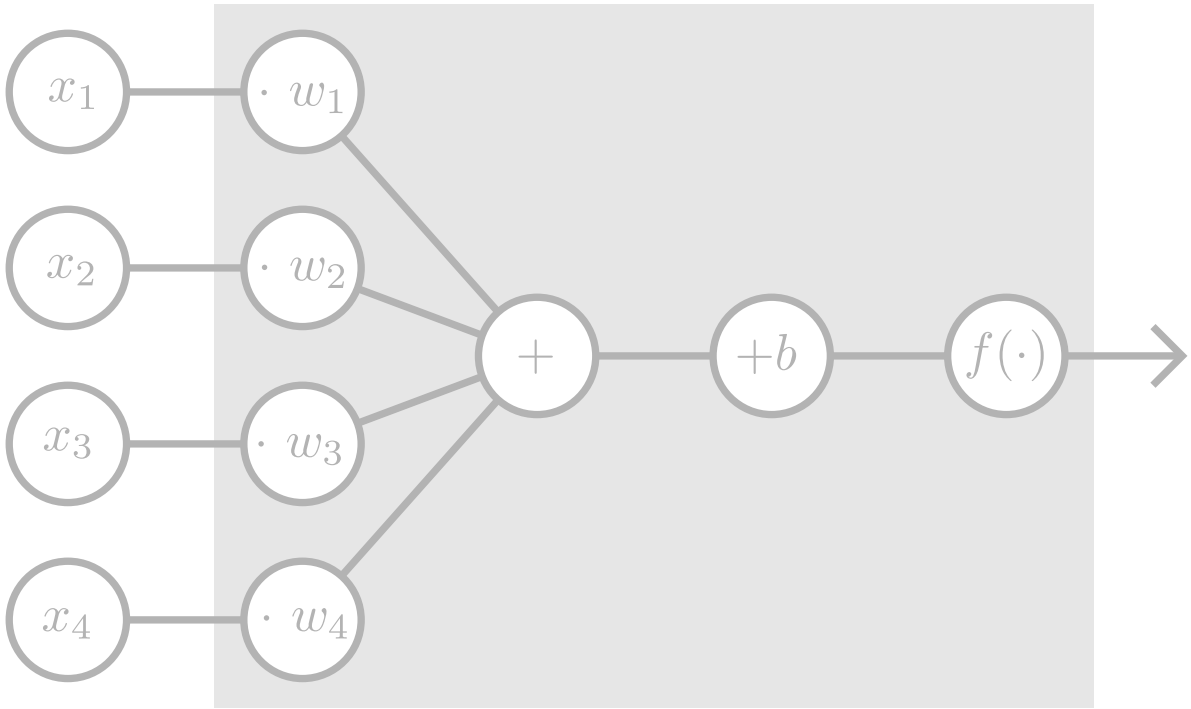
1. Preskúmať rôzne nové spôsoby, akými sa dá hlboké učenie urýchliť.
2. Experimentálne ich vyhodnotiť.

Stručne o neurónových sieťach

Čo? Načo?

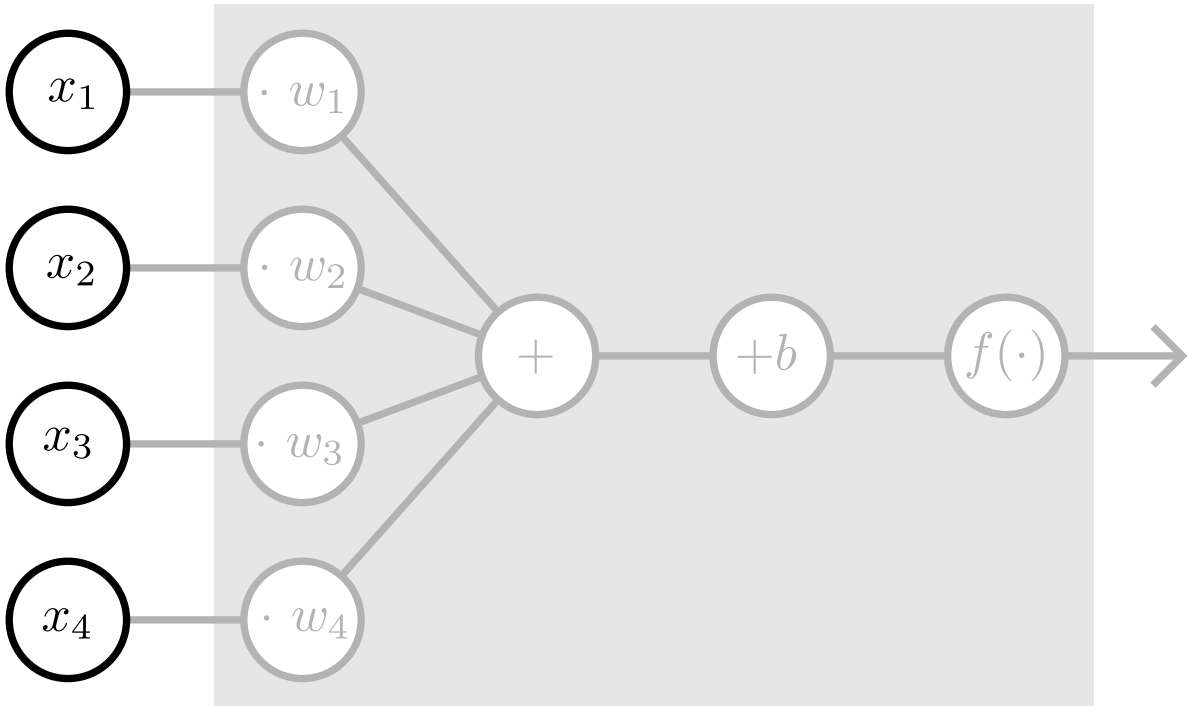
Umelý neurón

- parametrizovaná transformácia vstupov



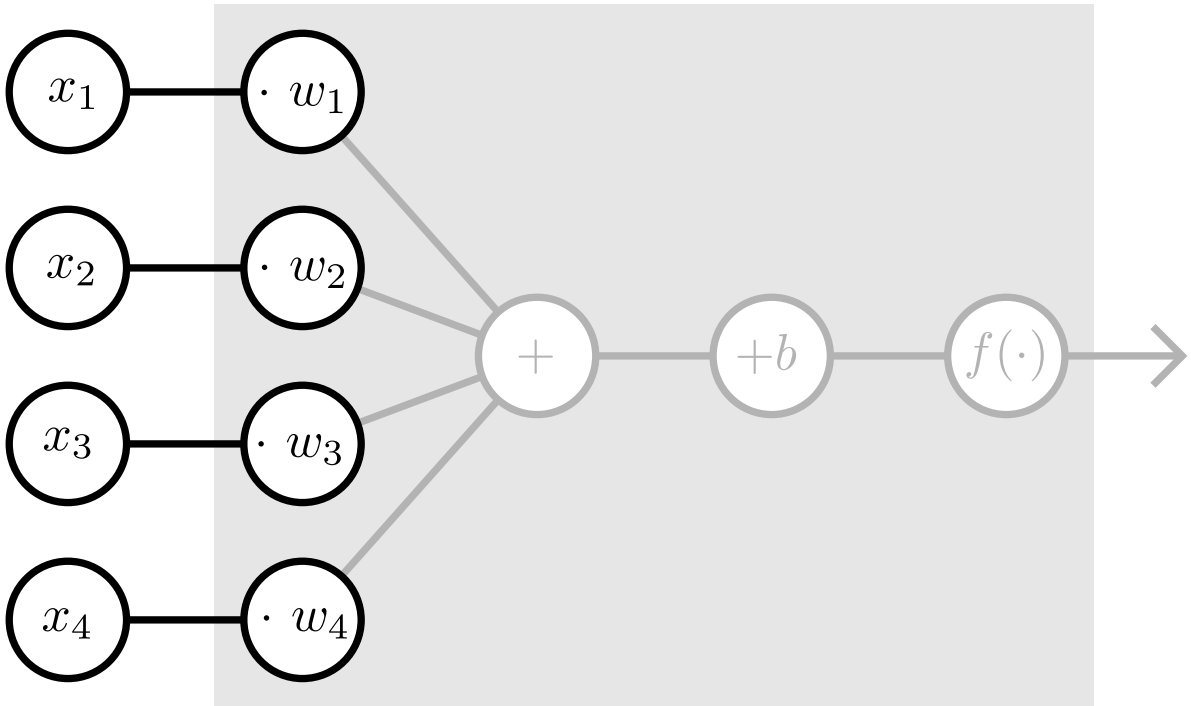
Umelý neurón

- parametrizovaná transformácia vstupov



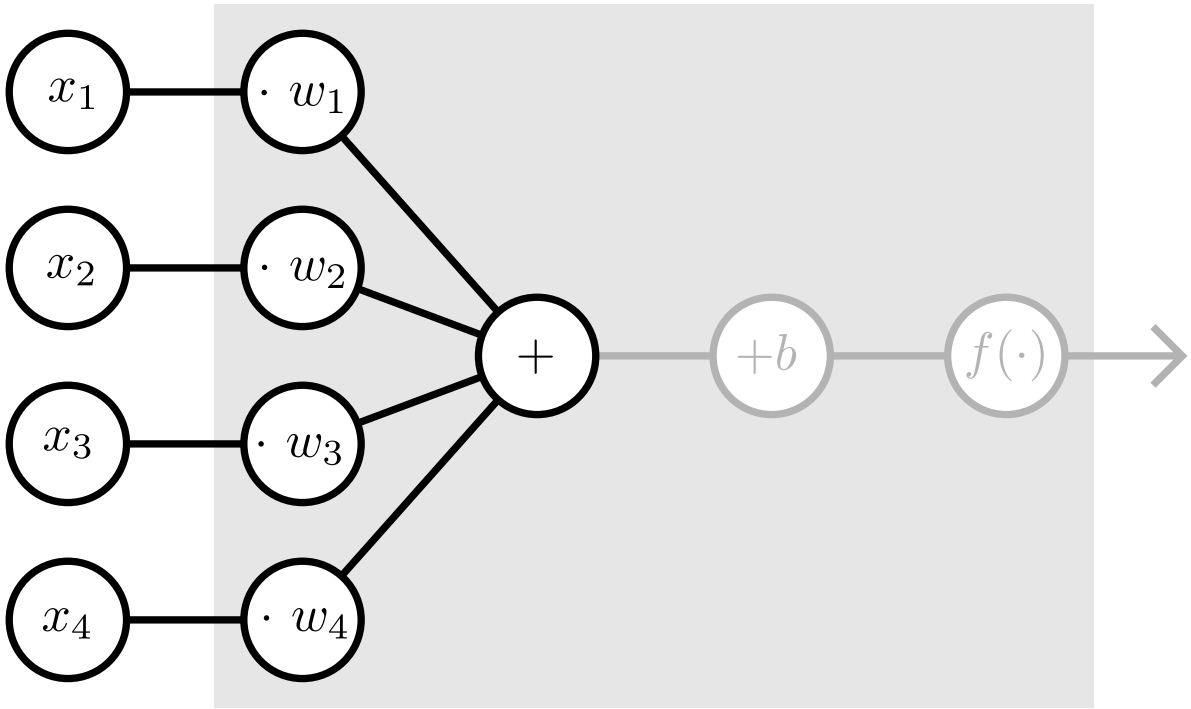
Umelý neurón

- parametrizovaná transformácia vstupov



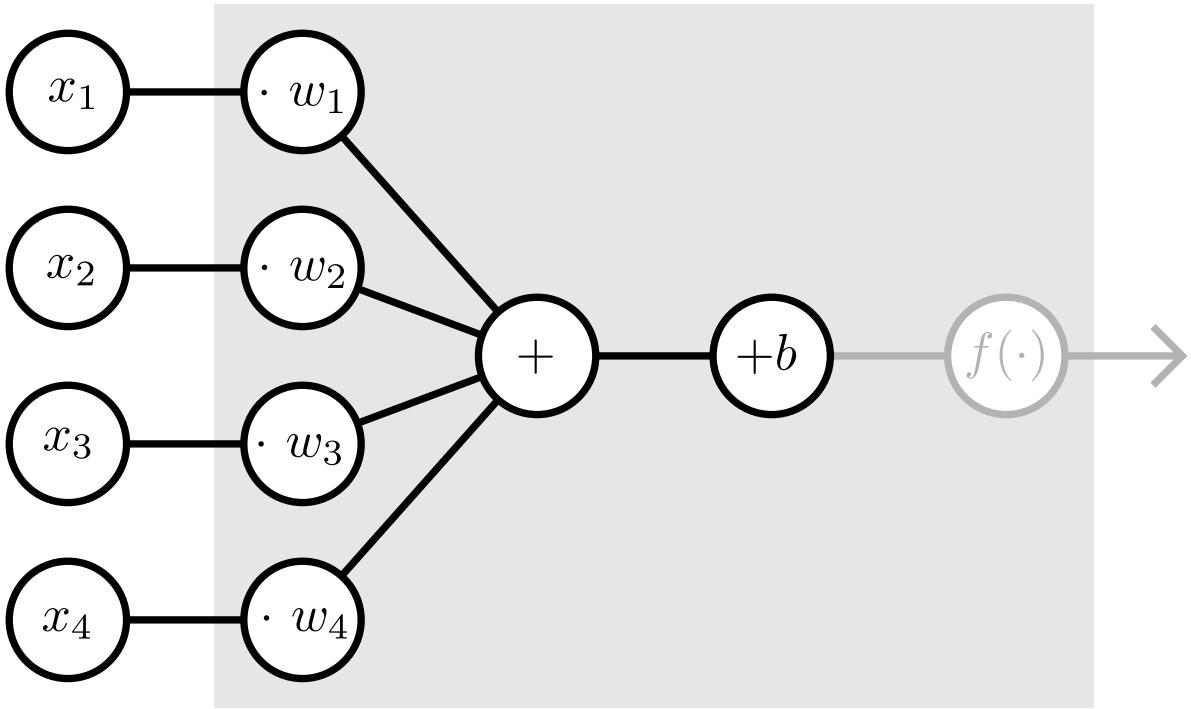
Umelý neurón

- parametrizovaná transformácia vstupov



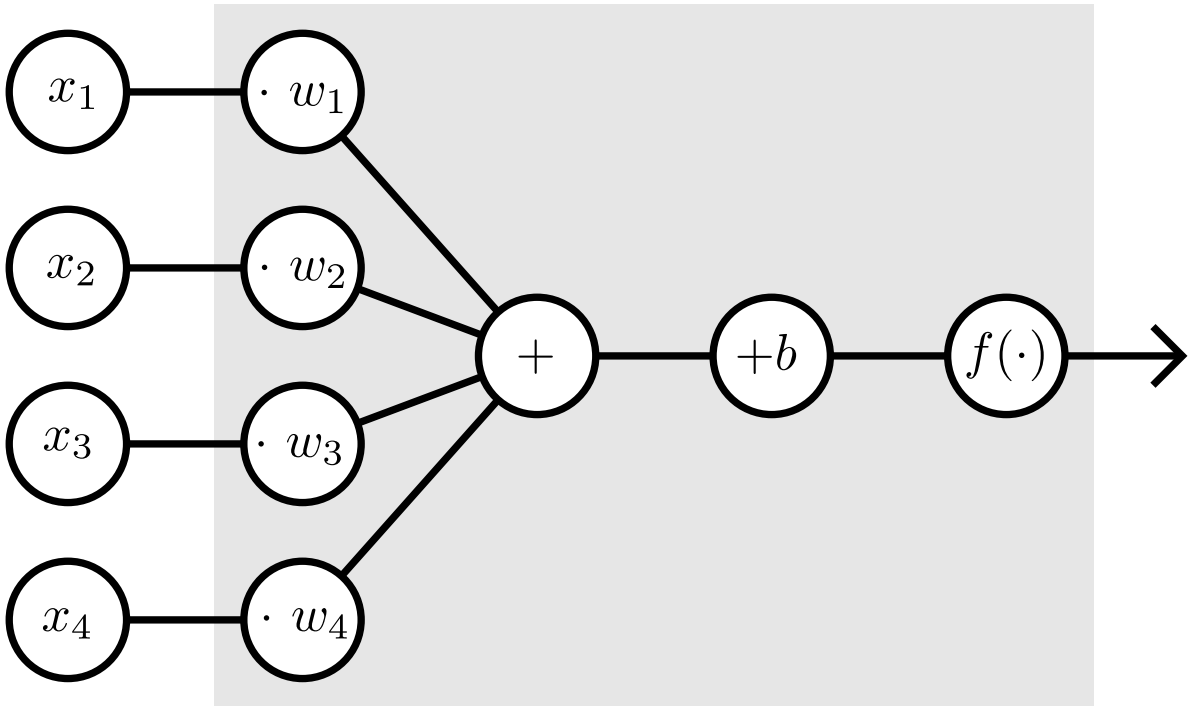
Umelý neurón

- parametrizovaná transformácia vstupov



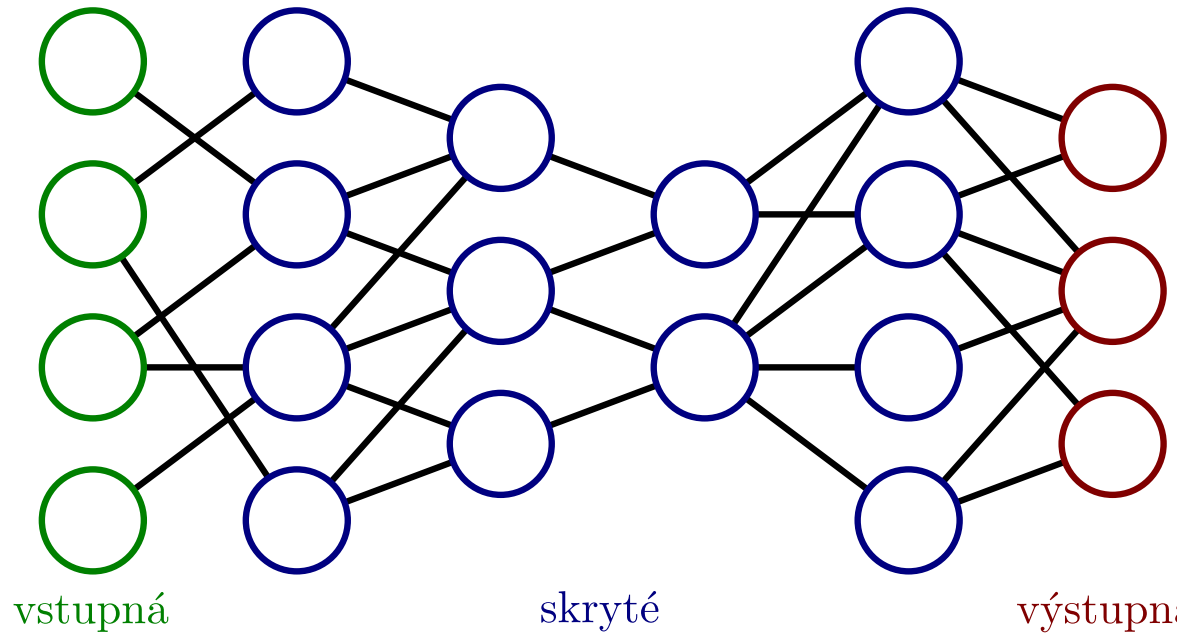
Umelý neurón

- parametrizovaná transformácia vstupov



Umelé neurónové siete

- skladanie sietí z umelých neurónov
- dopredné neurónové siete: neuróny organizované do vrstiev



Výpočtový graf

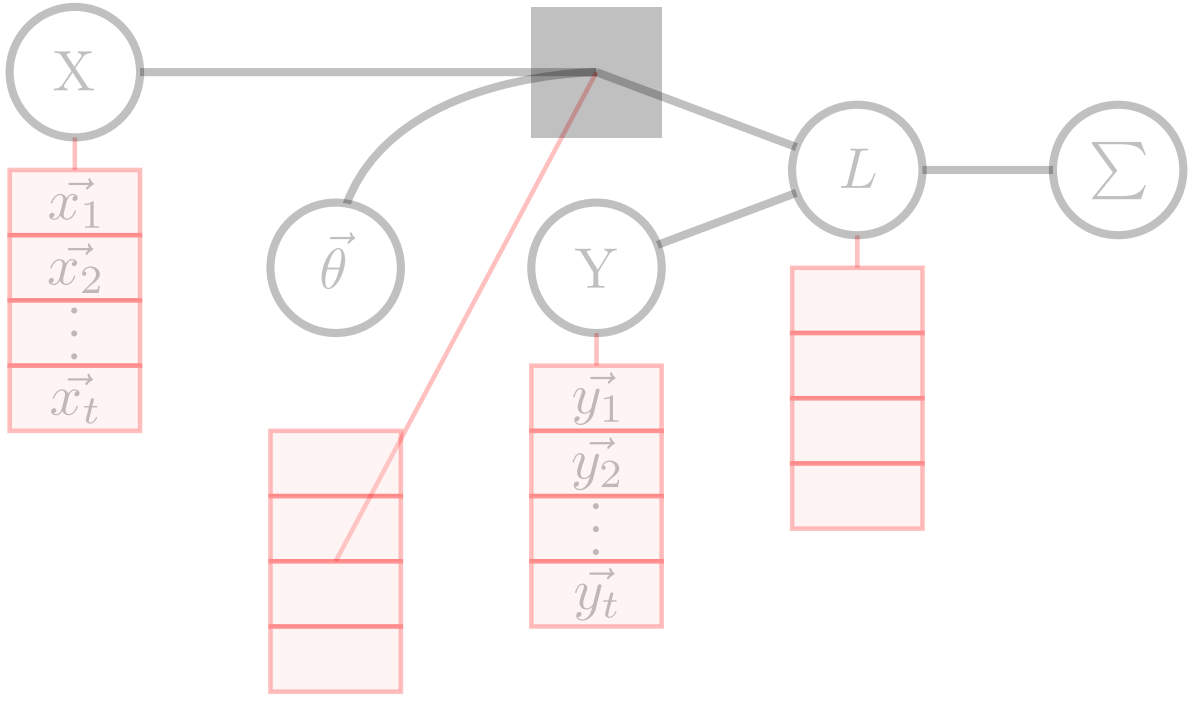
- orientovaný graf
- vrcholy reprezentujú aplikácie funkcii alebo hodnoty
- hrany reprezentujú vstupy do vrcholov
- operácie aj s vektormi a tenzormi

Učenie s učiteľom, všeobecne

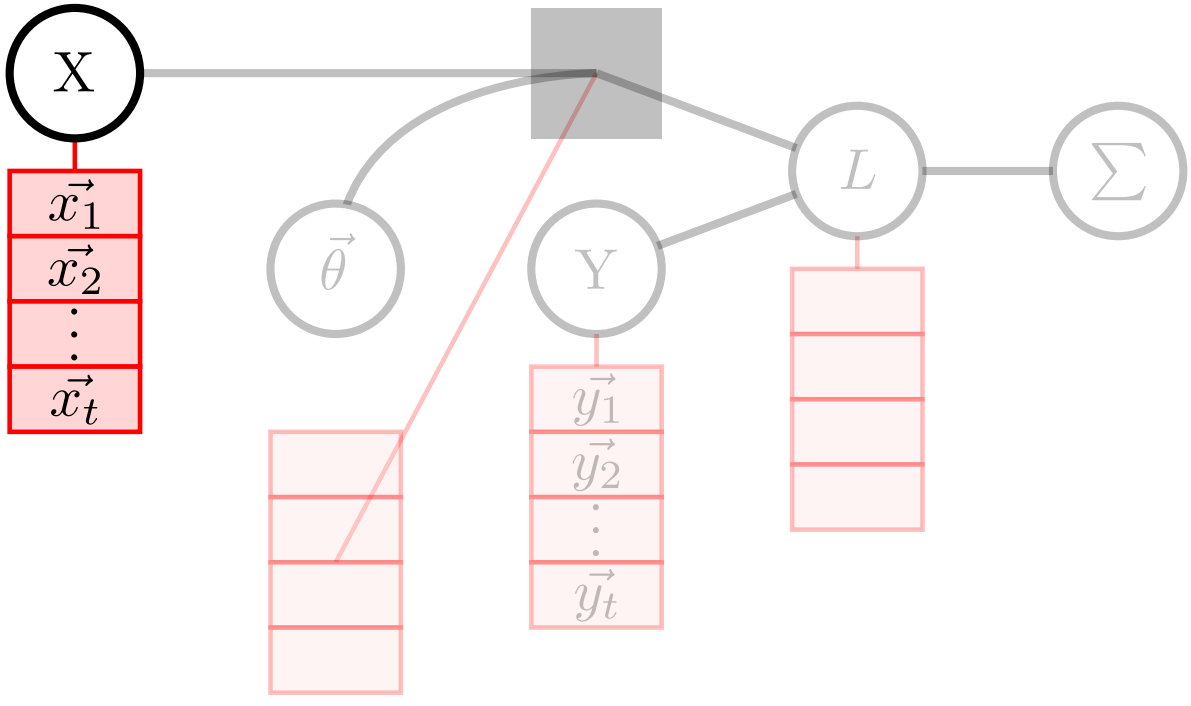
- trénovacie dáta: dvojice [vstup, správny výstup]
- minimalizujeme celkovú "chybu"

$$\sum_{i=1}^{t} L(\text{správny výstup}, \text{náš výstup})$$

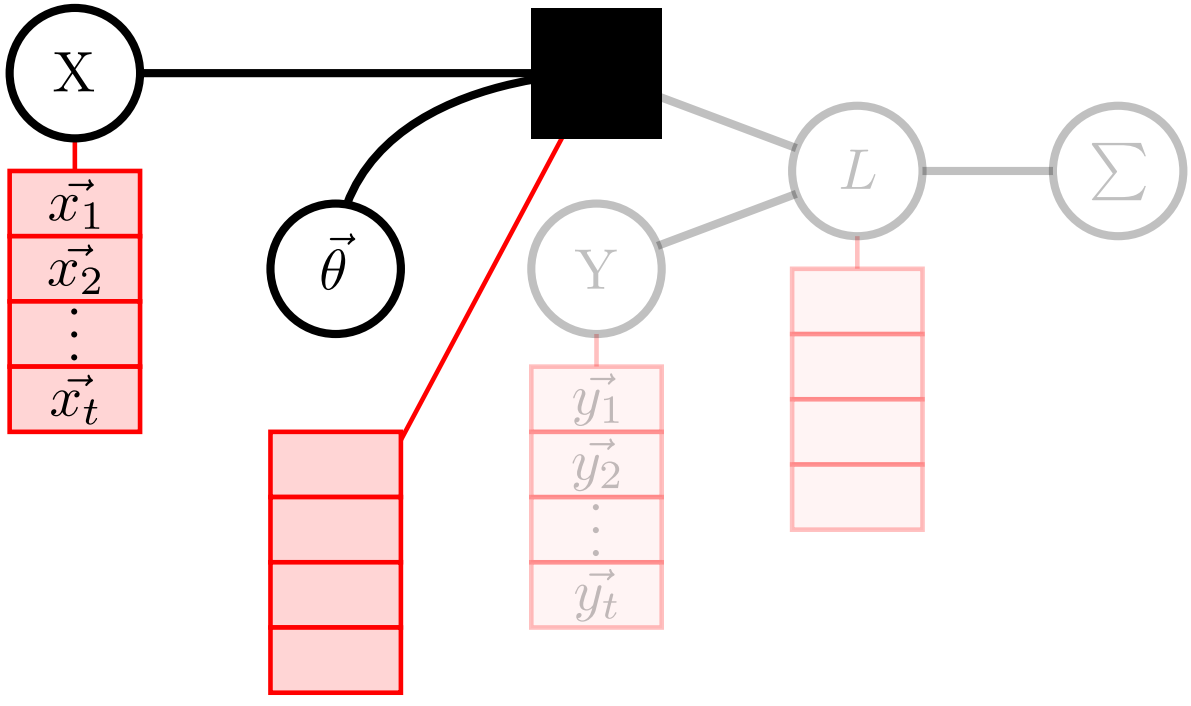
Učenie s učiteľom, výpočtový graf



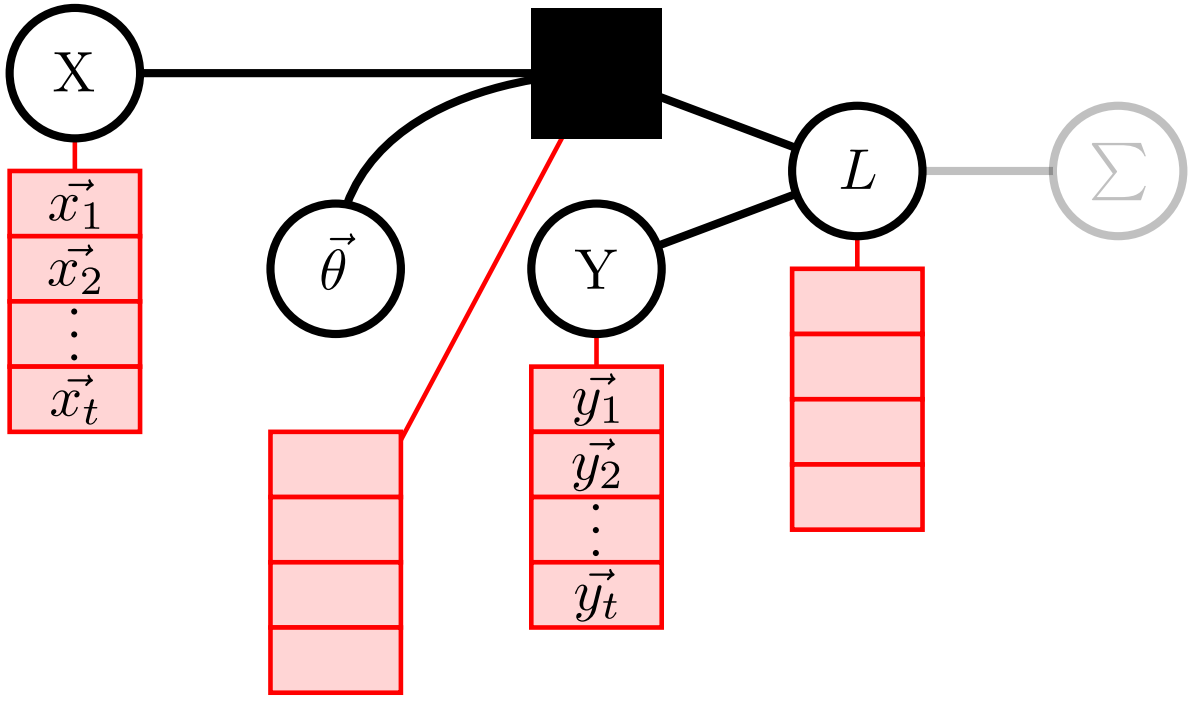
Učenie s učiteľom, výpočtový graf



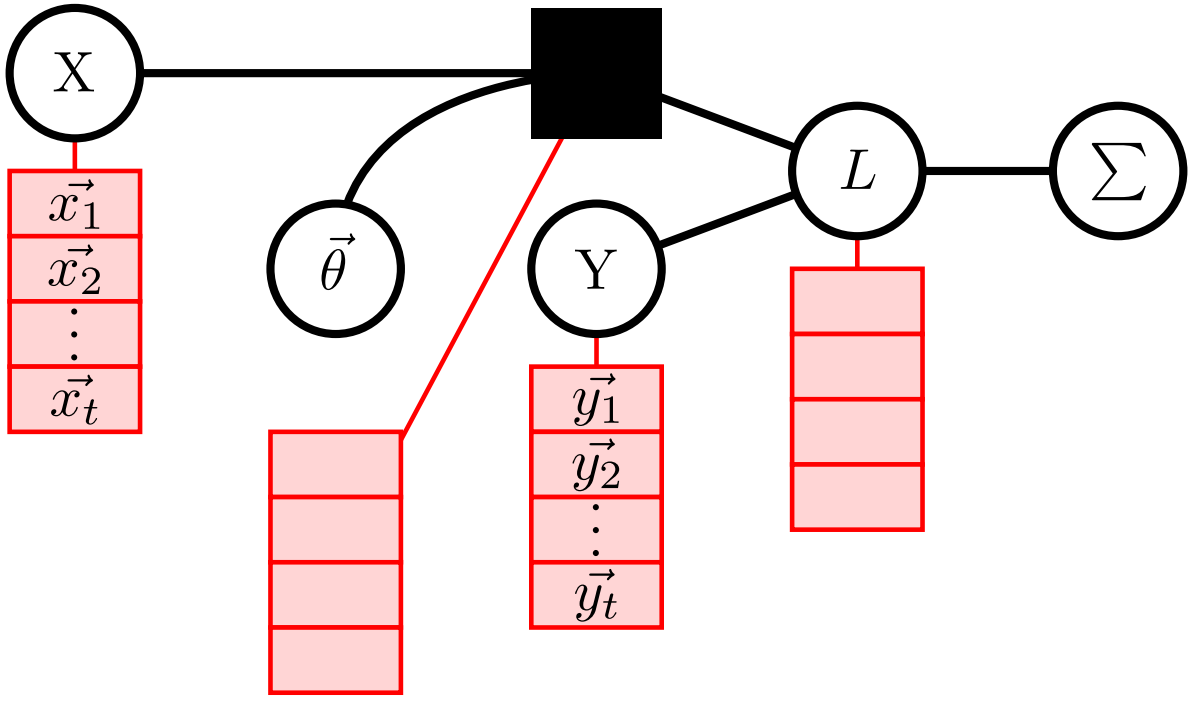
Učenie s učiteľom, výpočtový graf



Učenie s učiteľom, výpočtový graf



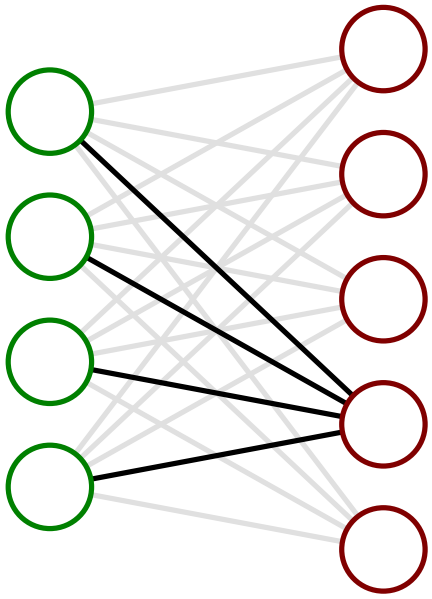
Učenie s učiteľom, výpočtový graf



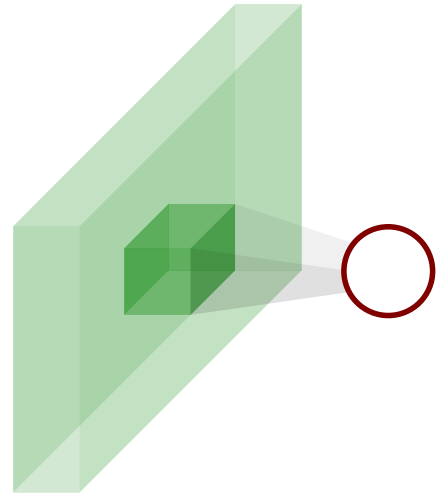
Učenie s učiteľom, výpočtový graf

- upravíme parametre siete podľa gradientu
- algoritmus spätnej propagácie chýb (angl. backpropagation)
- využitie tejto informácie—rôzne optimalizačné algoritmy

Štandardné stavebné prvky



Obr.: plne prepojená vrstva



Obr.: konvolučná vrstva

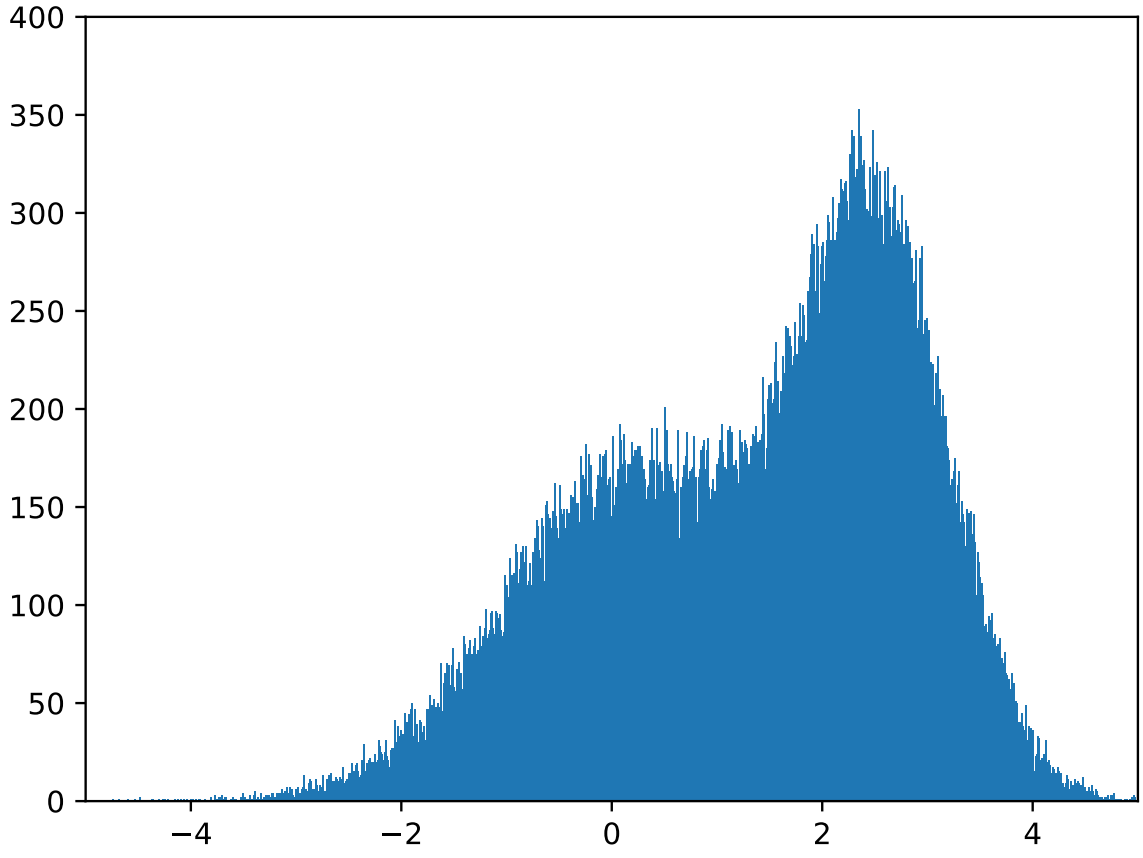
+ aktivačná vrstva

Ďalšie stavebné prvky

- dropout
- dávková normalizácia

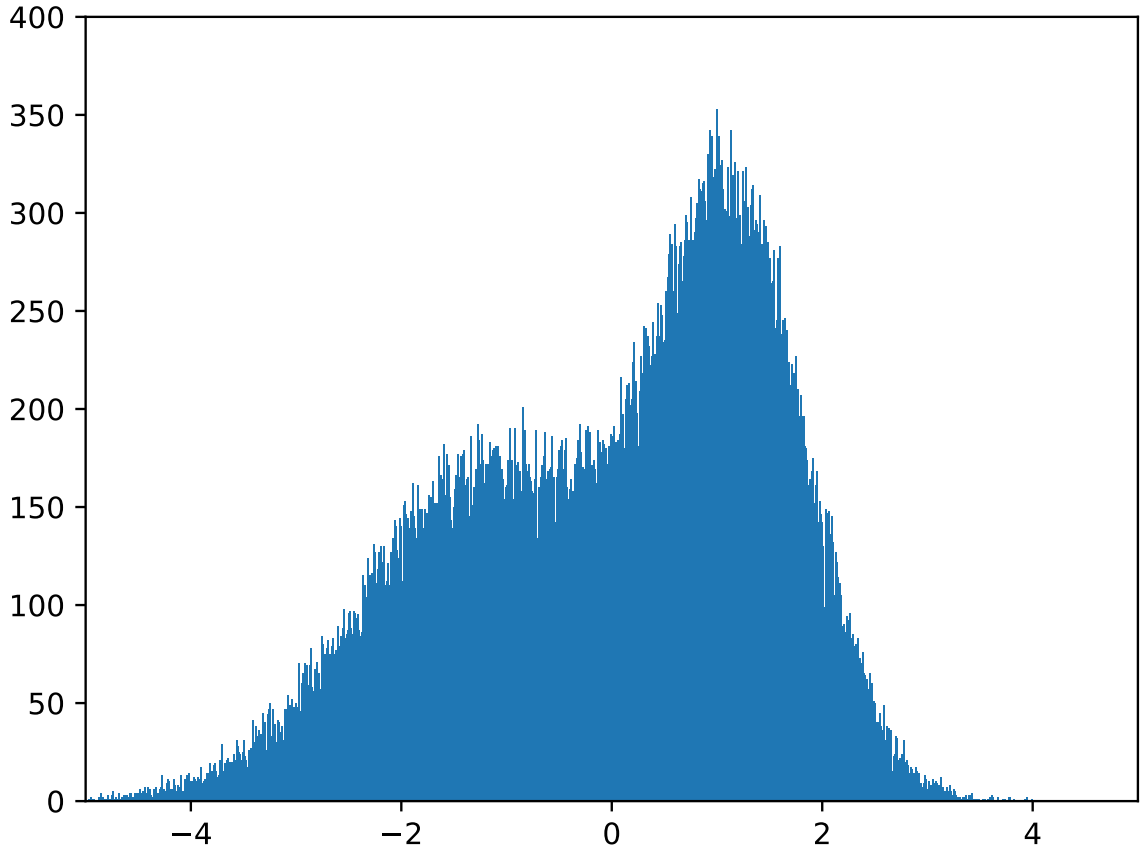
Dávková normalizácia

- $\mathrm{Mean}(\vec{x}) = 0$
- $\mathrm{Var}(\vec{x}) = 1$



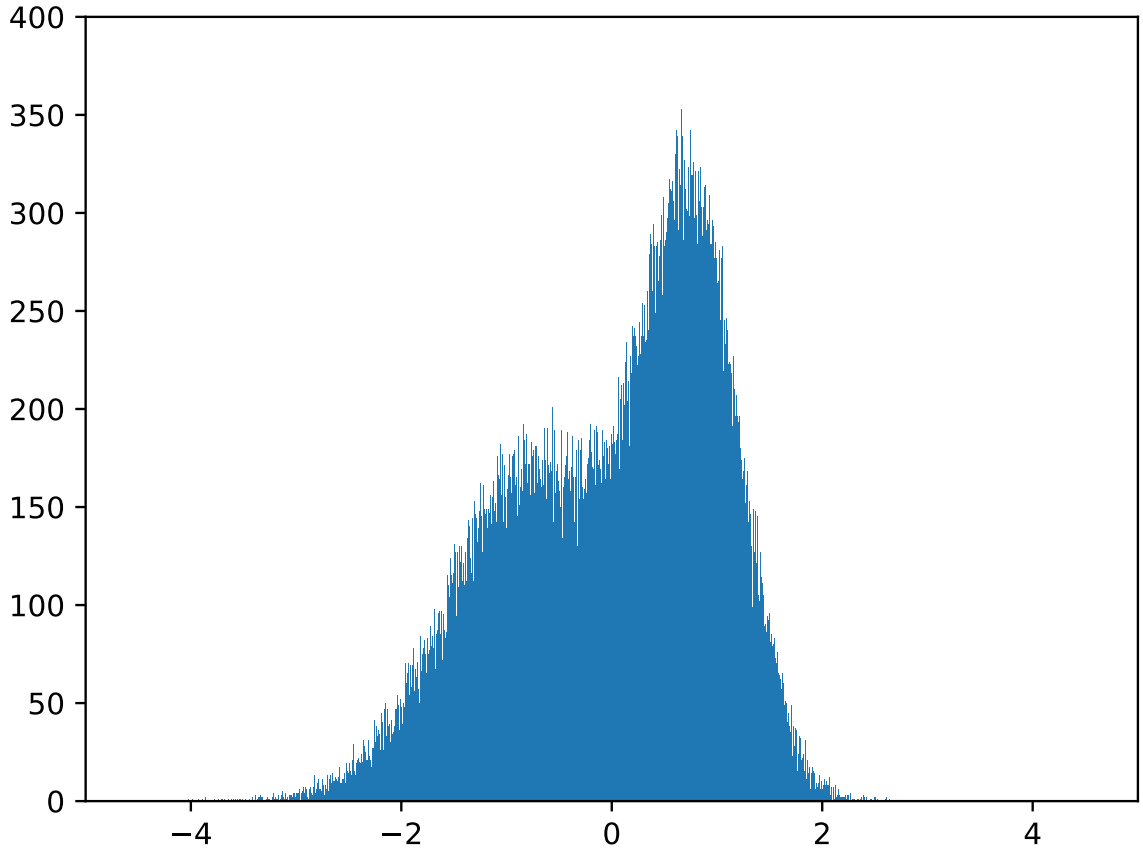
Dávková normalizácia

- $\mathrm{Mean}(\vec{x}) = 0$
- $\mathrm{Var}(\vec{x}) = 1$



Dávková normalizácia

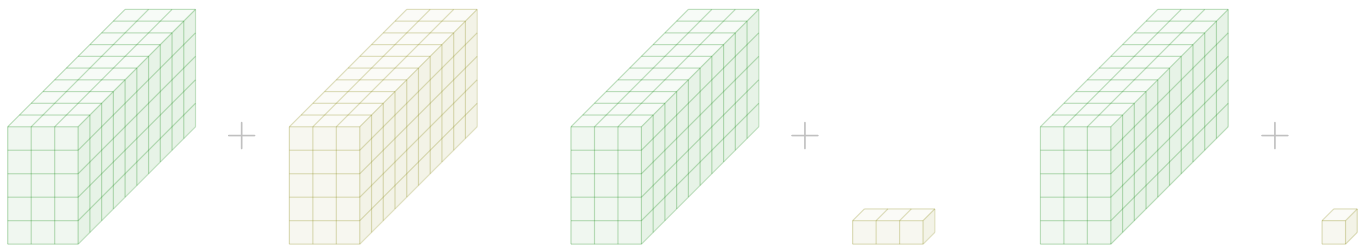
- $\mathrm{Mean}(\vec{x}) = 0$
- $\mathrm{Var}(\vec{x}) = 1$



Navrhované vylepšenia

Posuvná vrstva

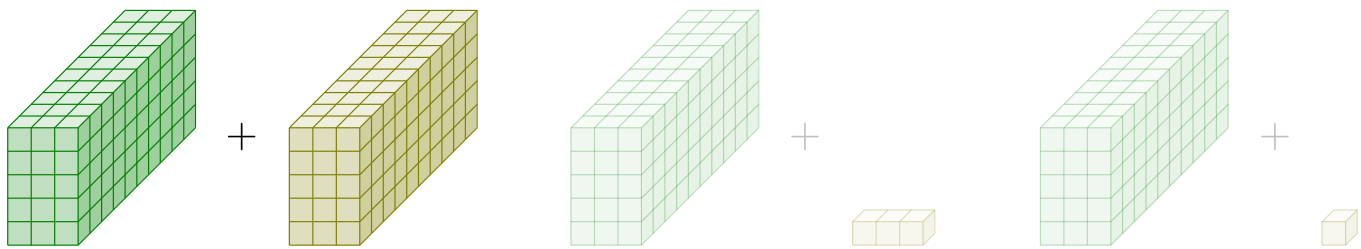
- pripočíta ku vstupom konštanty
- rôznym vstupom môže pripočítať rôzne, ale aj tie isté konštanty
  - posúvanie po prvkoch
  - po kanáloch
  - po vrstvách



individuálne kanálové vrstvové

Posuvná vrstva

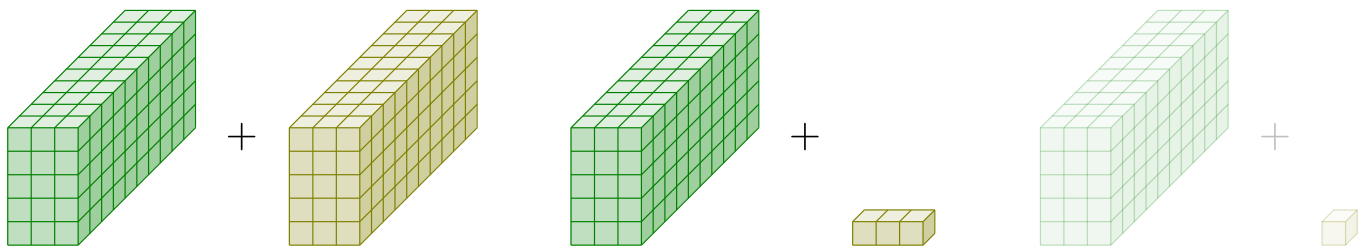
- pripočíta ku vstupom konštanty
- rôznym vstupom môže pripočítať rôzne, ale aj tie isté konštanty
  - posúvanie po prvkoch
  - po kanáloch
  - po vrstvách



individuálne kanálové vrstvové

Posuvná vrstva

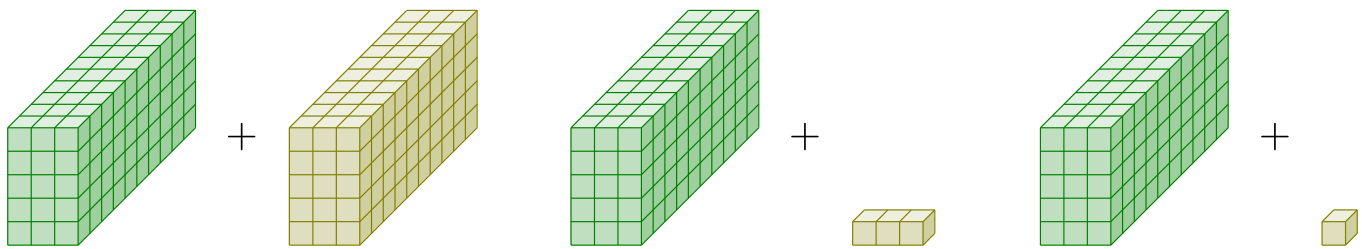
- pripočíta ku vstupom konštanty
- rôznym vstupom môže pripočítať rôzne, ale aj tie isté konštanty
  - posúvanie po prvkoch
  - po kanáloch
  - po vrstvách



individuálne kanálové vrstvové

Posuvná vrstva

- pripočíta ku vstupom konštanty
- rôznym vstupom môže pripočítať rôzne, ale aj tie isté konštanty
  - posúvanie po prvkoch
  - po kanáloch
  - po vrstvách



individuálne kanálové vrstvové

Škálovacia vrstva

- vynásobí vstupy konštantami
- rôzne vstupy môže vynásobiť rôznymi ale aj tými istými konštantami

Pozmenenie architektúry

- Za vstupnou vrstvou a za každou aktivačnou vrstvou, okrem poslednej:
  - (posuvná vrstva)
  - (škálovacia vrstva)

Testované kombinácie

		škálovanie			
		žiadne	individuálne	kanálové	vrstvové
posúvanie	žiadne	✓	✓	✓	✓
	individuálne	✓	✓	✗	✗
	kanálové	✓	✗	✓	✗
	vrstvové	✓	✗	✗	✓

Experimenty

Ako porovnávať?

- trénovacia chyba
- chyba na dátach, ktoré neboli použité pri trénovaní—**validačná chyba**
- (presnosť klasifikácie)

Vylepšenia

- žiadne
- s dávkovou normalizáciou
- naše

Dáta

- CIFAR10, CIFAR100
- 32×32 RGB obrázky, 10 resp. 100 tried
- 50 000 trénovacích, 10 000 testovacích dát
- vstupy normalizované

Dáta

(a) lietadlá

(b) autá

Obr.: Dve triedy z CIFAR10

Architektúry

- úplne konvolučná sieť
- viacvrstvový perceptrón

Dve verzie: s dropoutom/bez dropoutu

Technické detaily 1

- výstupná vrstva má 10 (resp. 100) neurónov
- posledná aktivačná funkcia je softmax, ostatné sú ReLU
- chybová funkcia: cross-entropy
- optimalizačný algoritmus: stochastická metóda najstrmšieho spádu, momentum 0.9

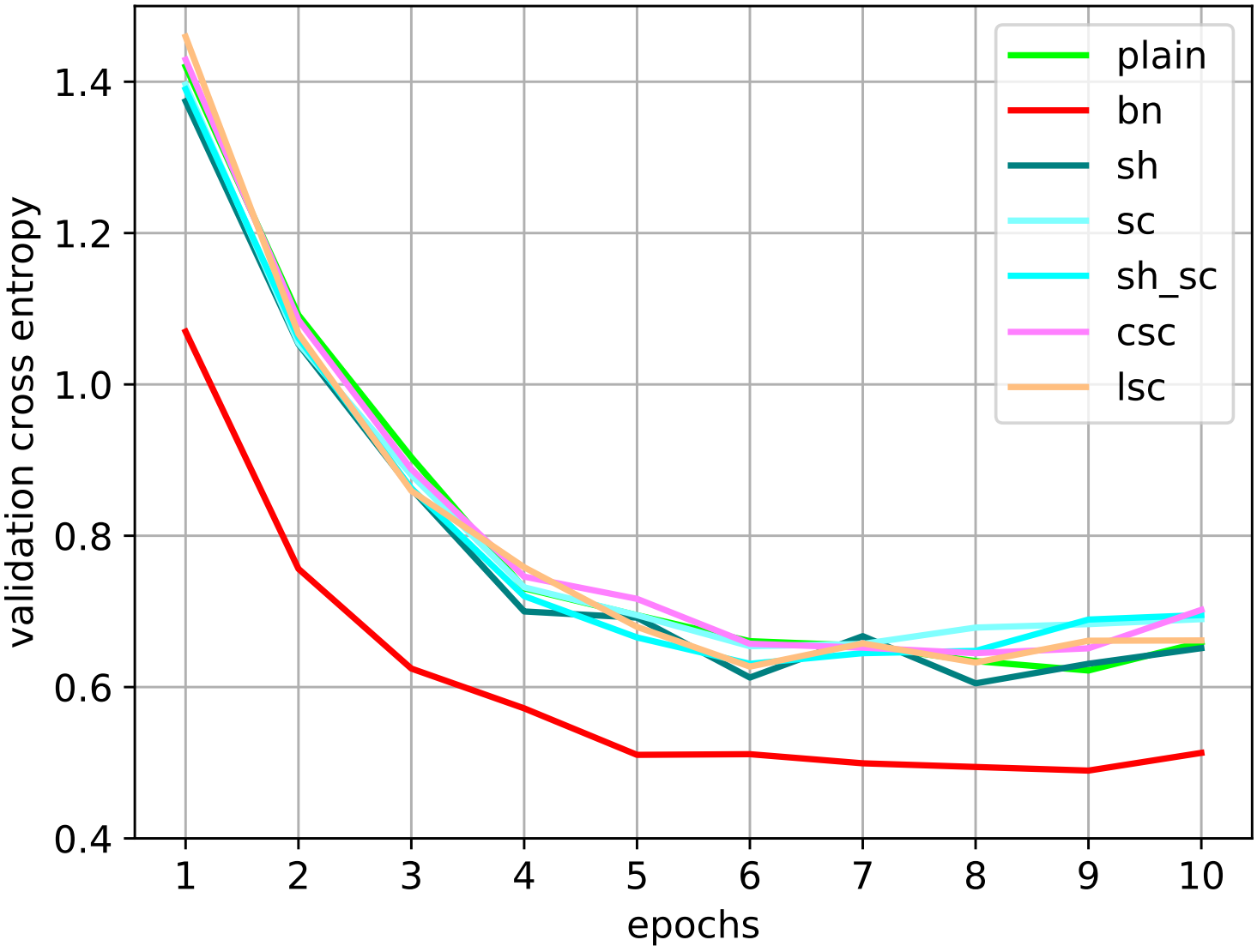
Technické detaily 2

- inicializácia váh: He initializácia, rovnaká inicializácia pre rôzne vylepšenia
- $+-$ manuálne hľadanie rýchlosti učenia (na grid search málo výpočtového času/sily)
- implementované vo frameworku PyTorch
- Microsoft Azure data science, 4 grafické karty

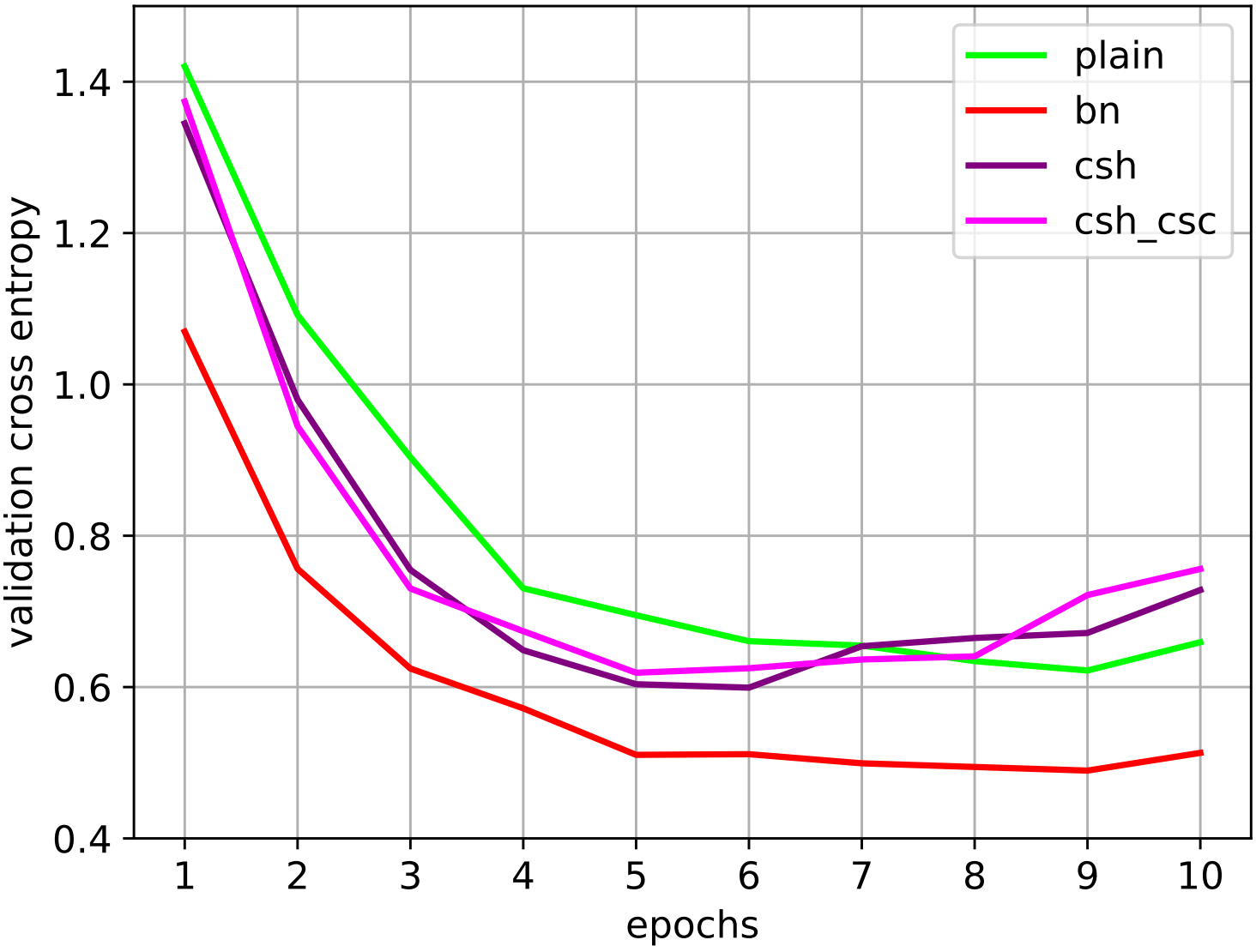
Počiatočný experiment

- CIFAR10
- úplne konvolučná sieť, bez dropoutu
- skúšame všetky vylepšenia

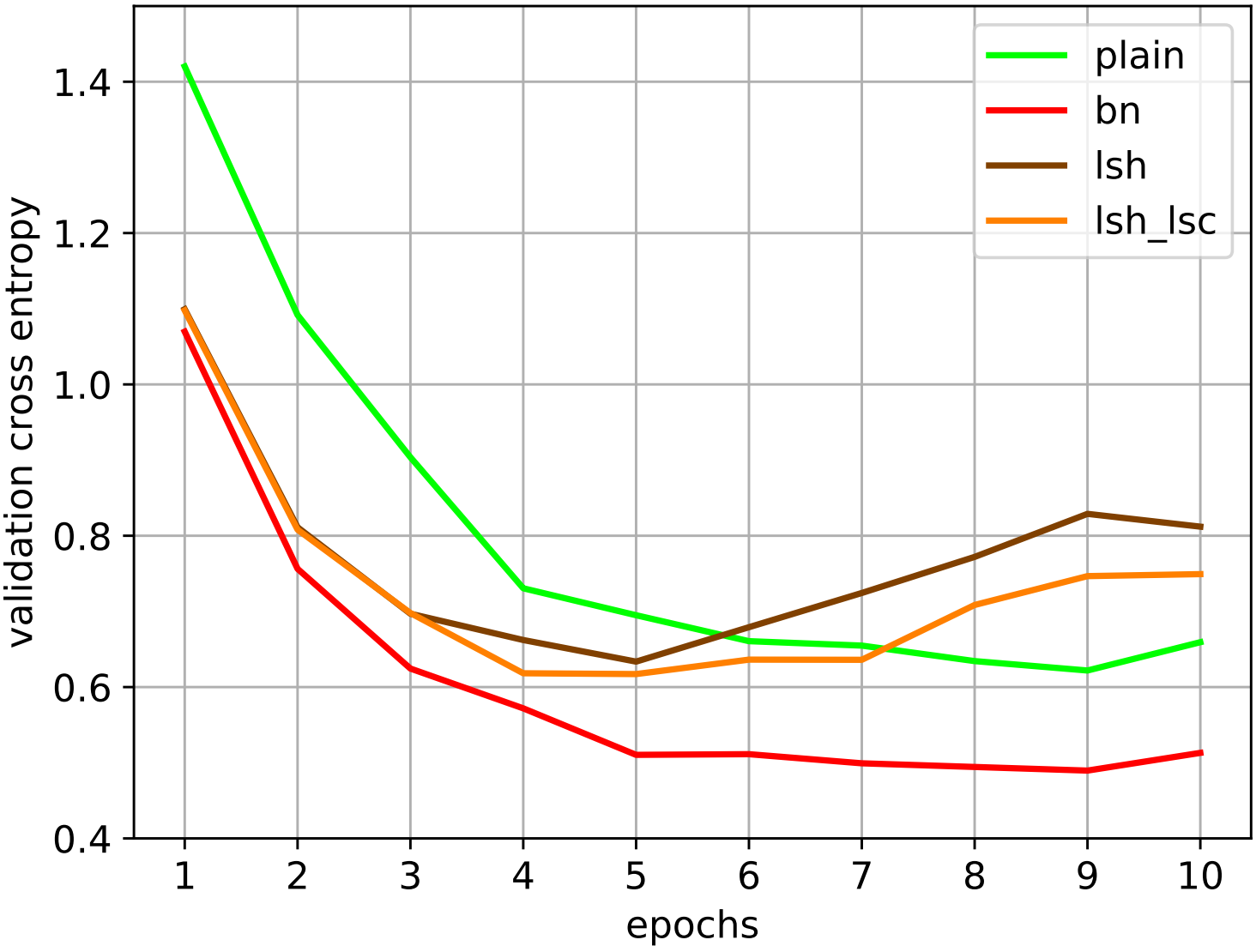
Počiatočný experiment



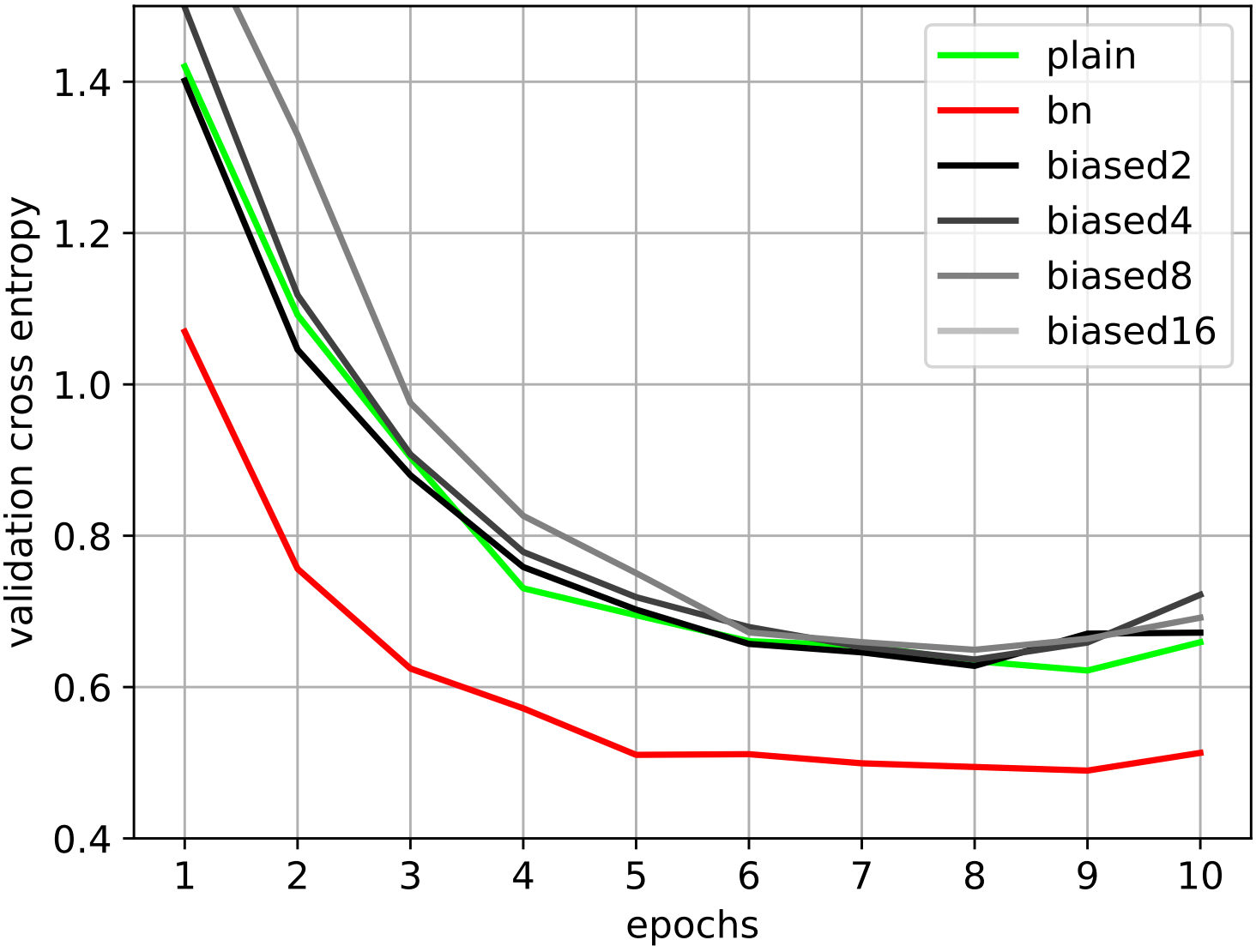
Počiatočný experiment



Počiatočný experiment



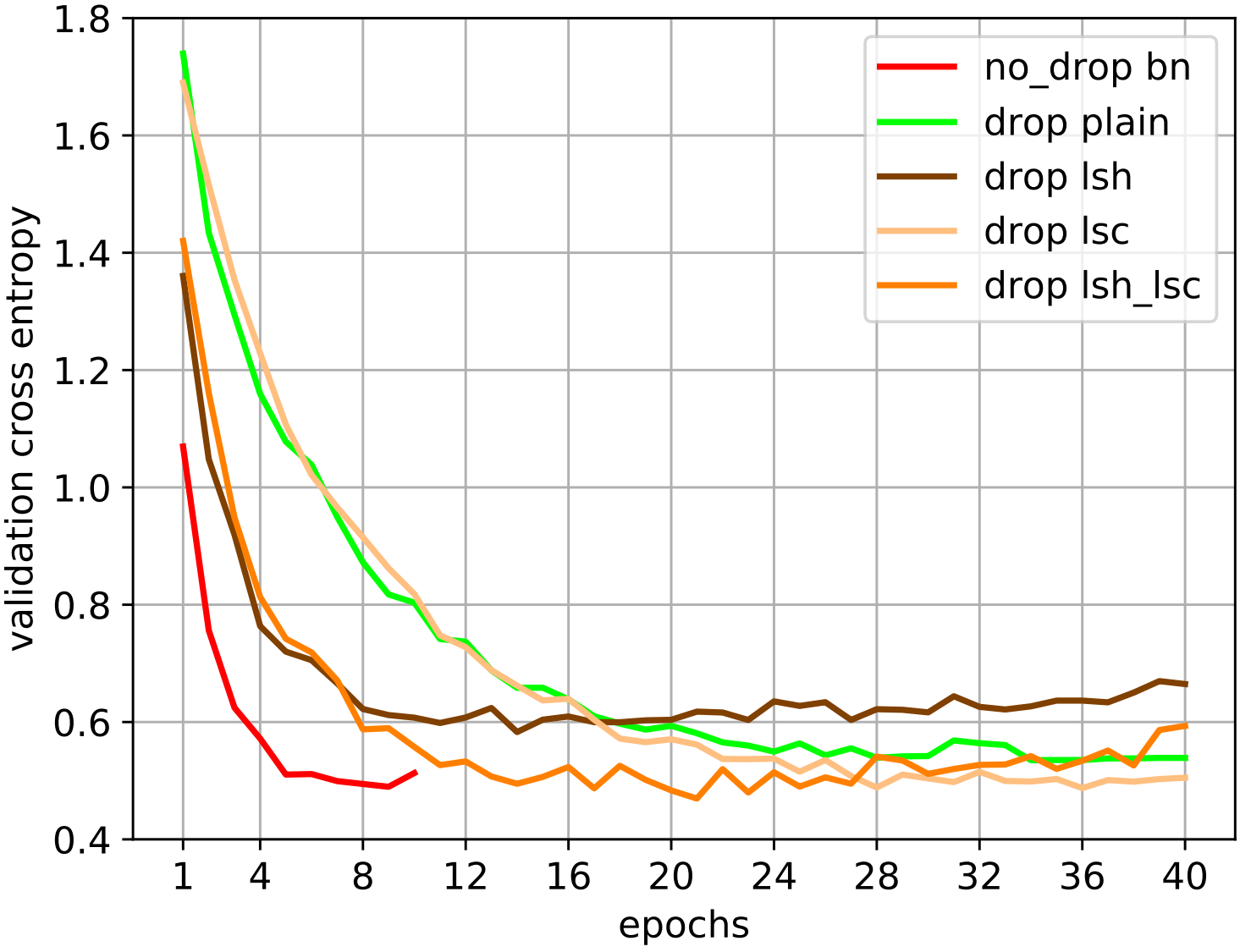
Počiatočný experiment



Experiment 2: dropout

- s dropoutom
- z našich skúšame len vrstvové kombinácie
- (CIFAR10, úplne konvolučná sieť)

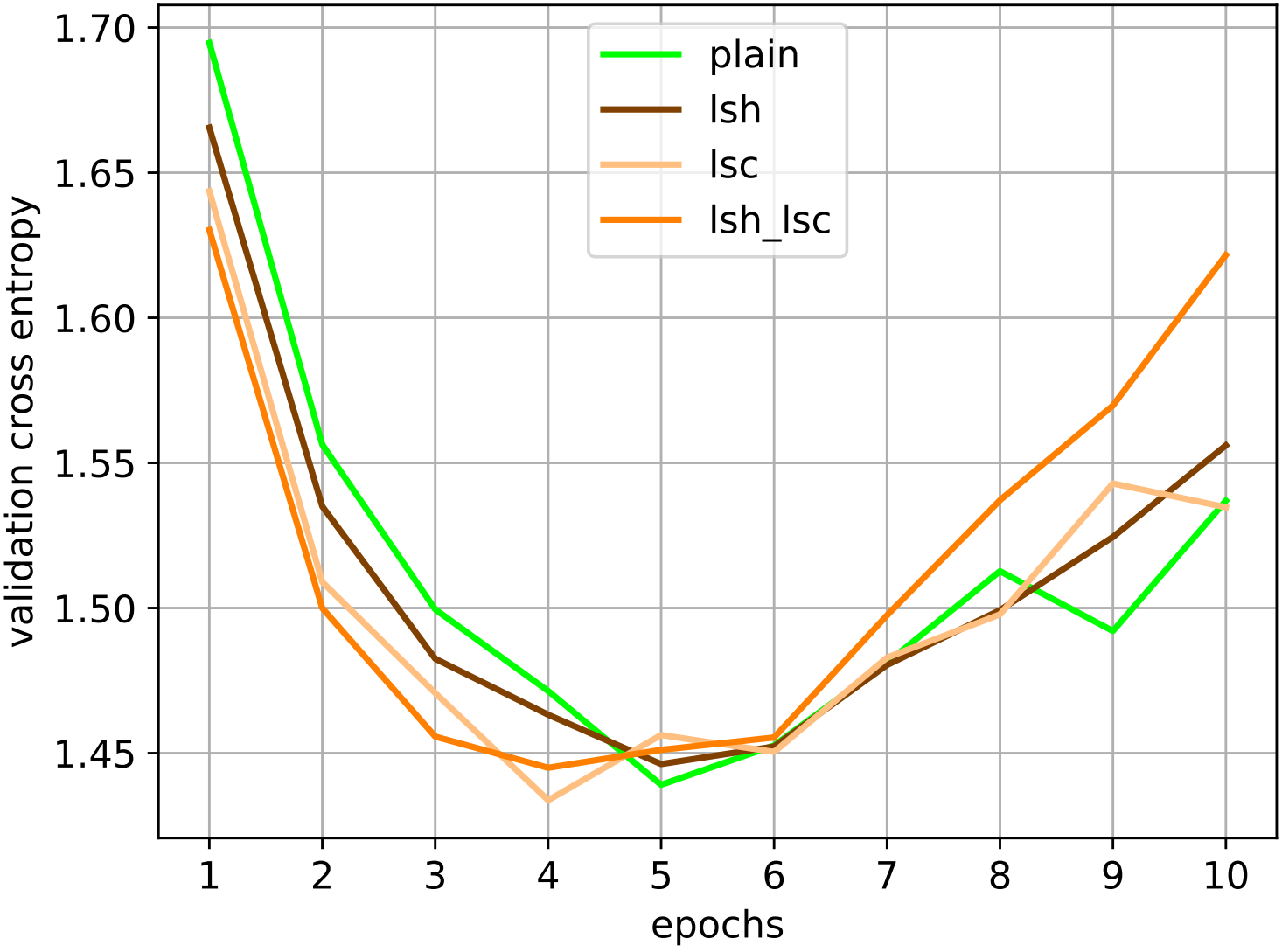
Experiment 2: dropout



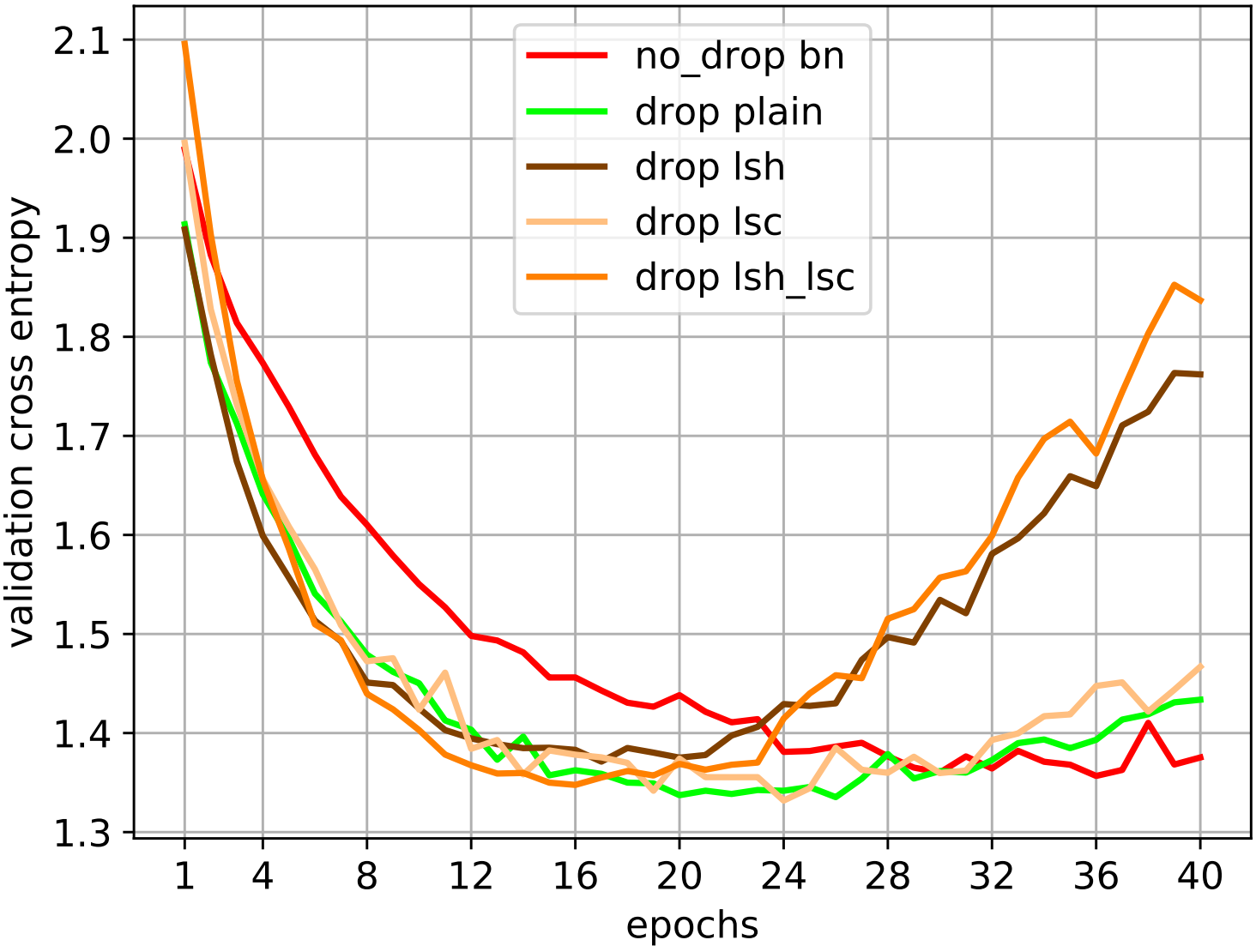
Experimenty 3 a 4: iná architektúra

- viacvrstvový perceptrón, najprv bez dropoutu a potom s dropoutom
- (CIFAR10, skúšame len vrstvové kombinácie)

Experiment 3



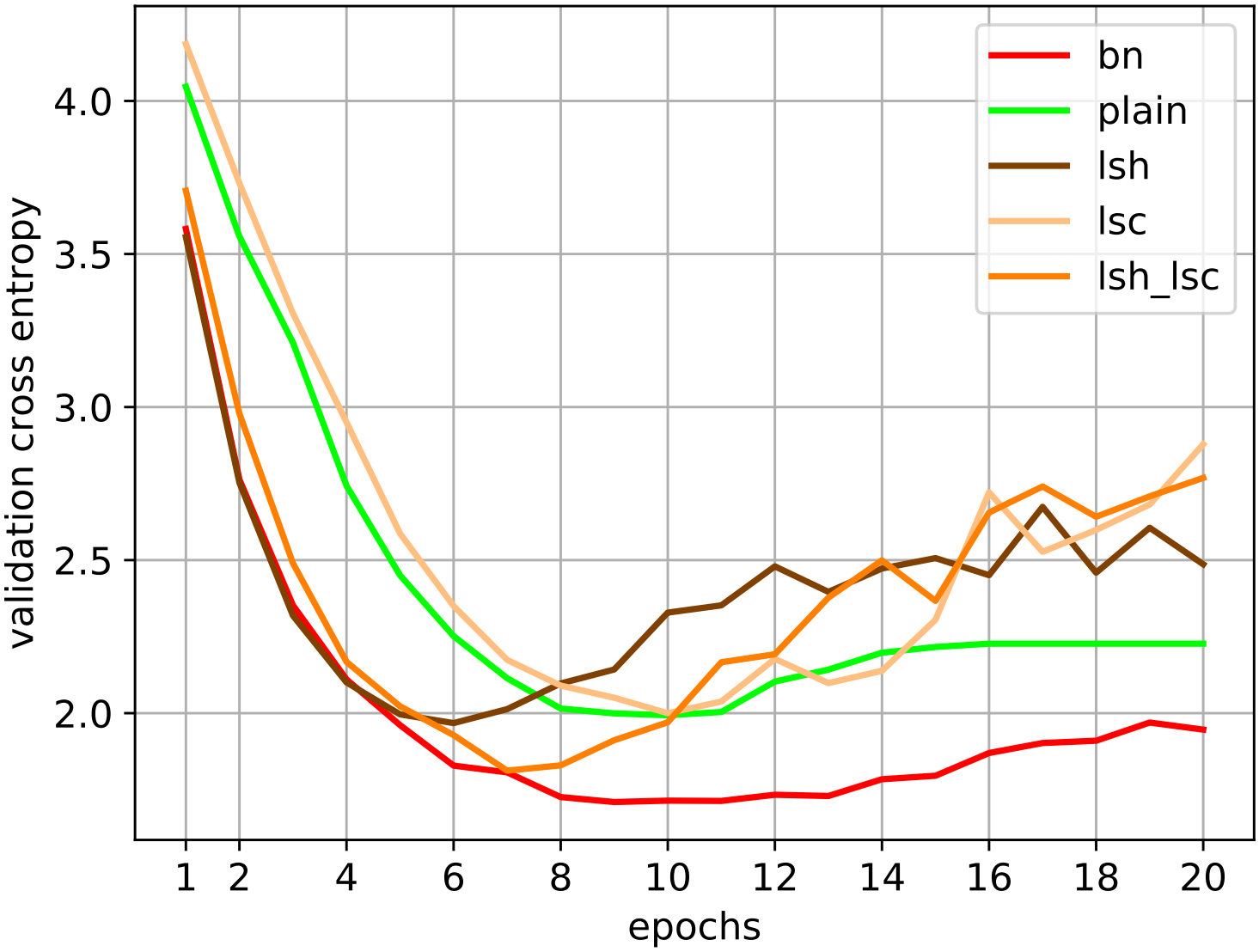
Experiment 4



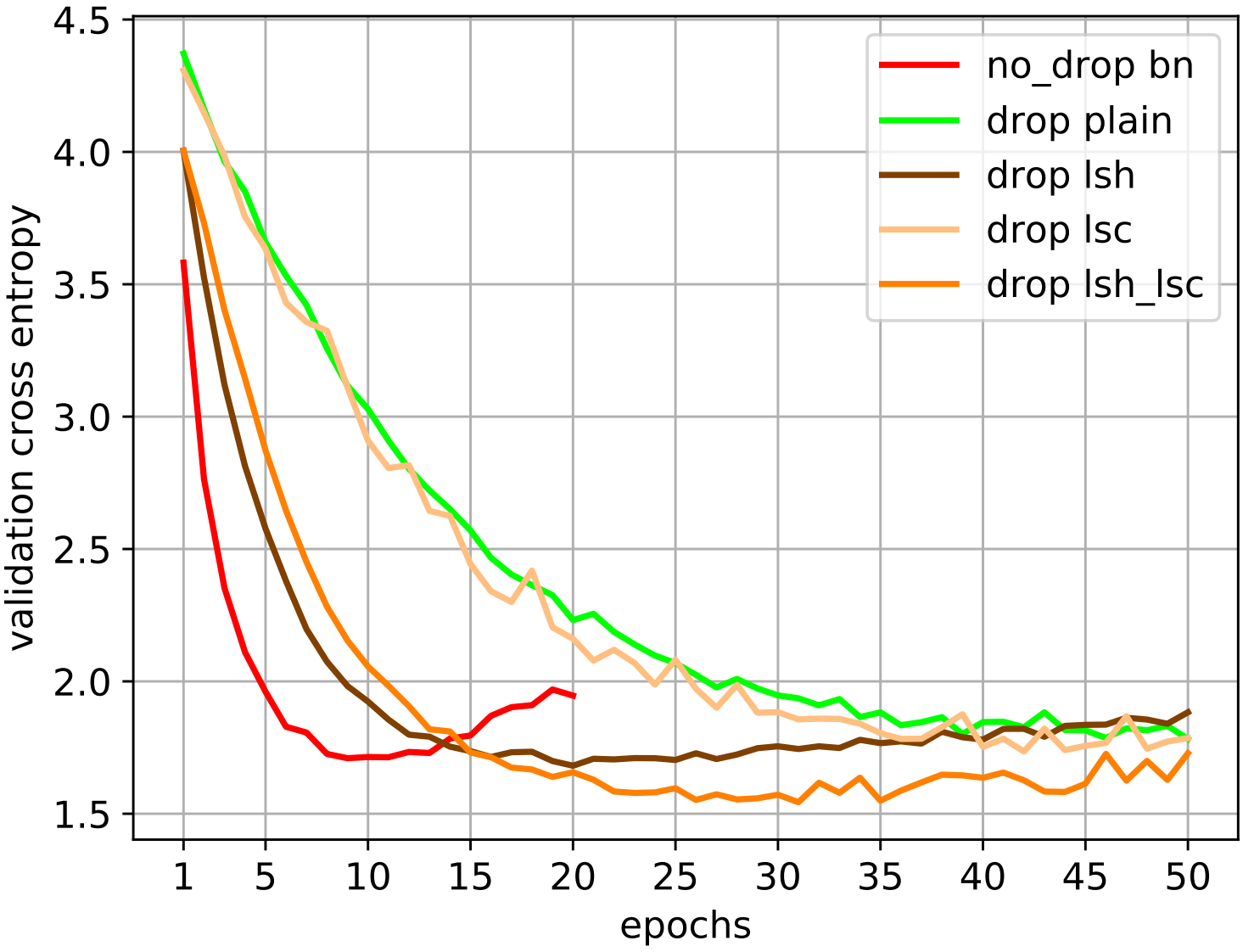
Experimenty 5 a 6: iný dataset

- CIFAR100
- úplne konvolučná sieť, najprv bez dropoutu a potom s dropoutom
- (skúšame len vrstvové kombinácie)

Experiment 5



Experiment 6



Záver

Zhrnutie experimentov

- vrstvové posúvanie-aj-škálovanie je najlepšia kombinácia
- lepšie ako žiadne vylepšenie, ale tendencia sa preučiť
- neporazili sme dávkovú normalizáciu, ale tá sa nie vždy dá použiť (online learning)

Ďalšia práca

- dôvod pozorovaného urýchlenia?
- čo sa naše vrstvy reálne naučia?
- vyhodnotiť na veľkých architektúrach

Otázky školiteľa

Otázka 1

Akým spôsobom ste vyberali experimentálne parametre? Napríklad pre parameter rýchlosť učenia ste testovali iba 3 hodnoty (0.02, 0.04 a 0.08) - prečo?.

Odpoveď 1

- relatívny rozdiel rýchlosti učenia, nie absolútny
- málo otestovaných hodnôt $\leftarrow$ málo výpočtového času
- experimentálne parametre:
  - architektúra
  - (chybová funkcia)
  - regularizácia

Otázka 1

Akým spôsobom ste vyberali experimentálne parametre? Napríklad pre parameter rýchlosť učenia ste testovali iba 3 hodnoty (0.02, 0.04 a 0.08) - prečo?.

Odpoveď 1

- relatívny rozdiel rýchlosti učenia, nie absolútny
- málo otestovaných hodnôt $\leftarrow$ málo výpočtového času
- experimentálne parametre:
  - architektúra
  - (chybová funkcia)
  - regularizácia

Otázka 1

Akým spôsobom ste vyberali experimentálne parametre? Napríklad pre parameter rýchlosť učenia ste testovali iba 3 hodnoty (0.02, 0.04 a 0.08) - prečo?.

Odpoveď 1

- relatívny rozdiel rýchlosti učenia, nie absolútny
- málo otestovaných hodnôt $\leftarrow$ málo výpočtového času
- experimentálne parametre:
  - architektúra
  - (chybová funkcia)
  - regularizácia

Otázka 1

Akým spôsobom ste vyberali experimentálne parametre? Napríklad pre parameter rýchlosť učenia ste testovali iba 3 hodnoty (0.02, 0.04 a 0.08) - prečo?.

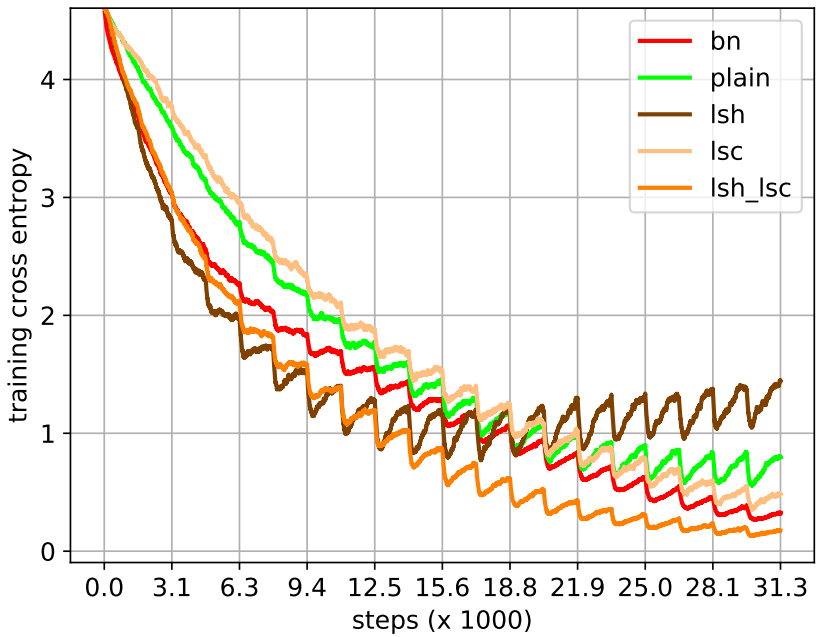
Odpoveď 1

- relatívny rozdiel rýchlosti učenia, nie absolútny
- málo otestovaných hodnôt $\leftarrow$ málo výpočtového času
- experimentálne parametre:
  - architektúra
  - (chybová funkcia)
  - regularizácia

Otázky oponenta

Otázka 1

Vysvetlite dôvod pílovitého priebehu poklesu chyby v grafoch. Koefficient 0.99 nie je vysvetlený.



Obr.: (CIFAR100, úplne konvolučná sieť bez dropoutu.) Exponential moving average of training errors after each step, with coefficient 0.99.

Ďalšie otázky

- Z práce nie je jasné, ako ste hľadali optimálne hodnoty posunu a škálovania (kap. 5) v jednotlivých modeloch. Pomocou gradientu (t.j. analogicky k synaptickým váham)?
- Neskúmali ste skonvergované hodnoty posunu a škálovania aktivít pre jednotlivé behy (s rovnakým seedom)? Boli podobné?
- V závere píšete, že použitie layerwise shift/scale metódy je sľubné, pretože výrazne urýchľuje učenie, no s rizikom preučenia. Ako by ste DNN nanútili silnejšiu regularizáciu?

Ďalšie otázky

- Z práce nie je jasné, ako ste hľadali optimálne hodnoty posunu a škálovania (kap. 5) v jednotlivých modeloch. Pomocou gradientu (t.j. analogicky k synaptickým váham)?
- Neskúmali ste skonvergované hodnoty posunu a škálovania aktivít pre jednotlivé behy (s rovnakým seedom)? Boli podobné?
- V závere píšete, že použitie layerwise shift/scale metódy je sľubné, pretože výrazne urýchľuje učenie, no s rizikom preučenia. Ako by ste DNN nanútili silnejšiu regularizáciu?

Ďalšie otázky

- Z práce nie je jasné, ako ste hľadali optimálne hodnoty posunu a škálovania (kap. 5) v jednotlivých modeloch. Pomocou gradientu (t.j. analogicky k synaptickým váham)?
- Neskúmali ste skonvergované hodnoty posunu a škálovania aktivít pre jednotlivé behy (s rovnakým seedom)? Boli podobné?
- V závere píšete, že použitie layerwise shift/scale metódy je sľubné, pretože výrazne urýchľuje učenie, no s rizikom preučenia. Ako by ste DNN nanútili silnejšiu regularizáciu?