

Półnadzorowane syjamskie sieci neuronowe

Adrian Musiał
we współpracy z dr. Markiem Śmieją

13 stycznia 2021

Wstęp

Celem projektu było zreprodukowanie wyników [1] oraz zbadanie metod uczenia półnadzorowanego zaproponowanych w [2] i [3] w kontekście syjamskich sieci neuronowych. Rozważaliśmy sytuację, w której liczba danych etykietowanych jest mała, mamy jednak dostęp do znacznie większego zbioru danych nieetykietowanych.

Krótki opis rozważanych modeli

Autorzy [1] proponują model półnadzorowany SEVEN – jest to syjamska sieć neuronowa o strukturze autoenkodera. Funkcja kosztu składa się z kary związanej z przyporządkowaniem pary do niewłaściwej klasy (obliczanej dla danych etykietowanych) oraz z rekonstrukcją obrazka na podstawie jego nisko wymiarowej reprezentacji (obliczanej zarówno dla danych etykietowanych, jak i bez etykiet).

Strategia opisana w [2] opiera się na nadawaniu w trakcie treningu pseudoetykiet (pseudo-labels) danym nieetykietowanym i stopniowym zwiększaniu wagi tychże danych w funkcji kosztu. Model jest wstępnie trenowany tylko na danych etykietowanych.

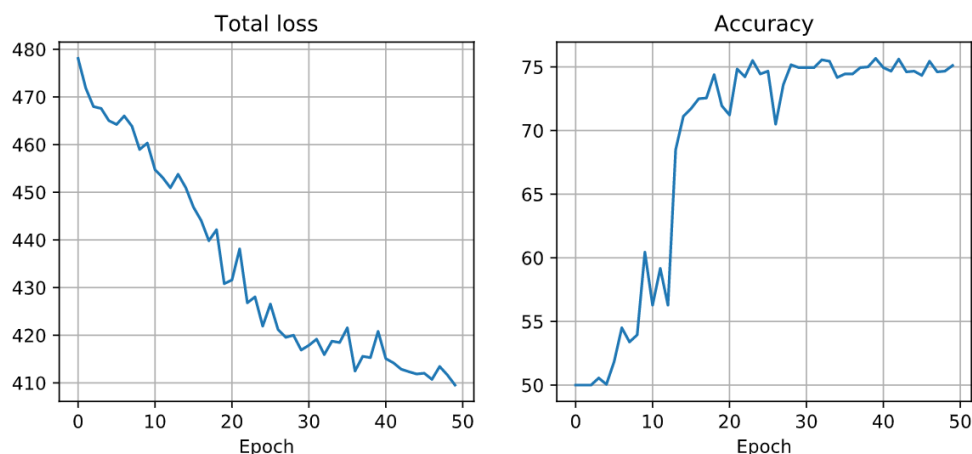
Podobne podejście proponują autorzy [3]. W etykietowaniu planowanym (curriculum labeling) pseudoetykiety nadawane są najpierw dla 20% najpewniejszych predykcji na zbiorze danych nieetykietowanych. Wartość ta wzrasta o 20% w każdej kolejnej iteracji do momentu nadania pseudoetykiet wszystkim parom.

Wyniki

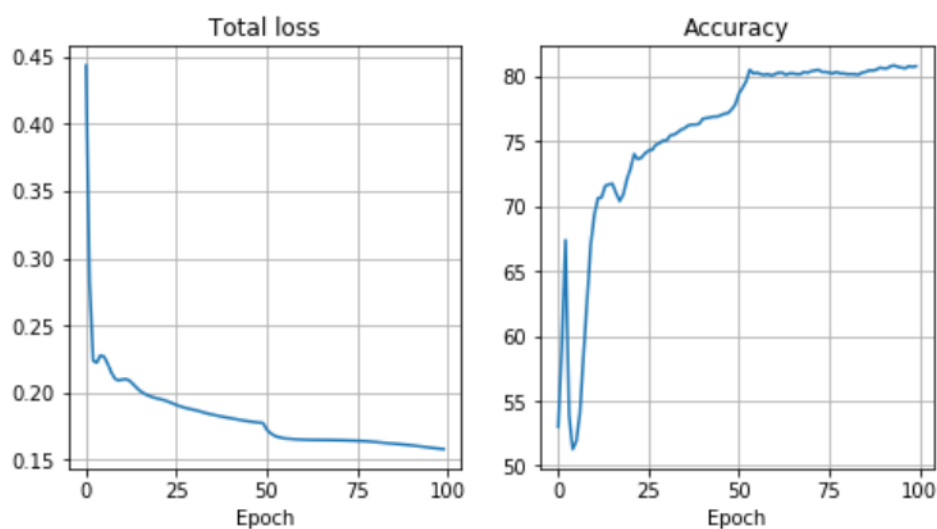
Do testów użyliśmy zbioru danych MNIST. Zbiór treningowy składał się z 180 par obrazków: 90 par pozytywnych („na obu obrazkach jest ta sama cyfra”) oraz 90 negatywnych. Dodatkowo losowaliśmy istotnie większy zbiór par nieetykietowanych. Zbiór testowy zawierał 900 par pozytywnych i 900 negatywnych, a jako metrykę przyjęliśmy dokładność.

W pierwszej części projektu badaliśmy wpływ rozmiaru zbioru treningowego (zarówno części etykietowanej, jak i tej bez etykiet) na SEVEN. Nie byliśmy w stanie odtworzyć wyników z [1], postanowiliśmy więc rozważyć modyfikacje modelu. Zmieniliśmy pierwszą funkcję kosztu z entropii krzyżowej na stratę kontrastową (contrastive loss [4]), nie spowodowało to jednak znaczącej poprawy.

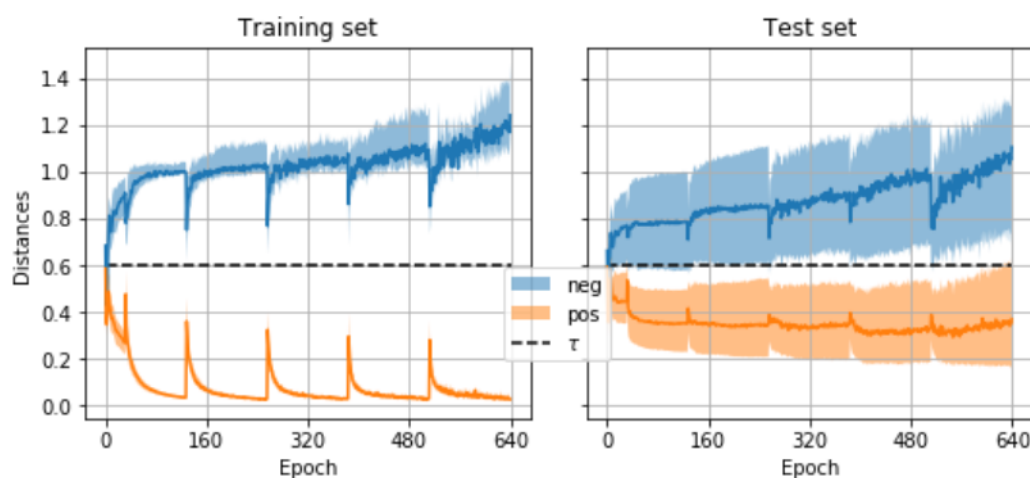
W drugiej części projektu staraliśmy się zaadaptować strategię pseudo-label i curriculum labeling do sieci syjamskich. Zastosowaliśmy tę samą architekturę sieci neuronowej (inną niż w SEVEN) oraz kontrastową funkcję kosztu. Dane etykietowane oraz nieetykietowane wykorzystaliśmy w stosunku 1/20. W obu przypadkach odnotowaliśmy poprawę wyników modelu po uwzględnieniu danych nieetykietowanych w treningu.



Rysunek 1. Przebieg treningu modelu SEVEN ze stratą kontrastową (zbiór testowy). Stosunek danych etykietowanych do nieetykietowanych wynosi 1/10.



Rysunek 2. Przebieg treningu modelu wykorzystującego strategię pseudo-label (zbiór testowy). Dane nieetykietowane używane są od epoki 50.



Rysunek 3. Rozdzielenie klas w trakcie treningu modelu wykorzystującego strategię curriculum labeling. Ciągłe linie to wartości mediany, a zacieniowane obszary wokół nich to IQR.

Źródła

- [1] <https://www.ijcai.org/Proceedings/2017/0358.pdf>
- [2] <https://arxiv.org/pdf/1908.02983.pdf>
- [3] <https://arxiv.org/pdf/2001.06001.pdf>
- [4] <http://yann.lecun.com/exdb/publis/pdf/hadsell-chopra-lecun-06.pdf>
- [5] Kod i wyniki: <https://github.com/adr-mus/ss-snn>