Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра математики

ФУНКЦІЯ ВТРАТ У ЗАДАЧАХ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ З МАЛОЮ НАВЧАЛЬНОЮ ВИБІРКОЮ

Текстова частина до курсової роботи
за спеціальністю „Системний аналіз”

Керівник курсової роботи
к.т.н Швай Надія Олександрівна

_____
(підпис)
“____” _____ 2021 р.

Виконала студентка
Судорженко Анна

“____” _____ 2021 р.

Київ 2021

ЗМІСТ

ВСТУП 3

РОЗДІЛ 1: АНАЛІЗ ЗАДАЧІ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ З МАЛОЮ НАВЧАЛЬНОЮ ВИБІРКОЮ 4

1.1 Задача класифікації зображень 4

1.2 Визначення функції втрат. 5

1.2.1 Функції втрат $L1$ та $L2$ 5

1.2.2 Функція втрат hinge 6

1.2.3 Функція втрат cross entropy (log loss) 7

1.2.4 Функція втрат softmax з маржею 8

1.2.5 Функція втрат cosine 10

РОЗДІЛ 2. РОЗРОБКА МОДЕЛІ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ НА МАЛІЙ НАВЧАЛЬНІЙ ВИБІРЦІ ДАНИХ 12

2.1 Постановка задачі класифікації зображень, опис використаної моделі та опис навчальних вибірок 12

2.2 Результати навчання моделі, аналіз отриманих даних 13

ВИСНОВОК 17

ЛИТЕРАТУРА 18

ВСТУП

У задачі класифікації зображень однією з проблем є необхідність великого набору даних для побудови алгоритму з високою оцінкою його якості. На практиці, процес збору великих навчальних вибірок є досить ресурсозатратним процесом. Один з методів вирішення цієї проблеми є вибір правильної функції витрат. На сьогоднішній час найчастіше у задачах класифікації обираються одна функція втрат через недооцінювання її впливу на точність моделі. Незважаючи на велику кількість наукових робіт, в яких описують нові, більш потужні loss функції, їх використання є досить рідкісним явищем.

За мету даної роботи було поставлено дослідження існуючих функцій втрат та аналіз їх впливу на точність моделі у задачі класифікації зображень з малою навчальною вибіркою.

Використання аналізу впливу функції втрат на точність моделі дозволяє отримати додатковий інструмент для покращення якості розпізнавання об’єктів.

Робота складається з двох розділів.

Перший розділ присвячується аналізу задачі класифікації зображень та дослідженню існуючих функцій втрат для бінарних та категоріальних задач класифікації.

У другому розділі наведено результати аналізу впливу обраних функцій витрат на точність задачі класифікації, що навчена на трьох різних малих навчальних вибірках.

РОЗДІЛ 1: АНАЛІЗ ЗАДАЧІ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ З МАЛОЮ НАВЧАЛЬНОЮ ВИБІРКОЮ

1.1 Задача класифікації зображень

Розглянемо математичне формулювання задачі класифікації:

Нехай X – множина описів об'єктів, Y – множина назв класів. Існує невідома цільова залежність – відображення $y^*: X \to Y$, значення якої відомі лише на елементах скінченної навчальної вибірки $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$. Потрібно побудувати алгоритм $a: X \to Y$, здатний класифікувати довільний об'єкт $x \in X$

З визначення стає зрозуміло, що точність алгоритму напряму залежить від кількості елементів в навчальній вибірці. Визнається, що потребу у великих наборах даних є однією з проблем методів глибинного навчання. Перед початком побудови моделі класифікації, необхідно витратити багато ресурсів на процес збору даних та їх розмітку. Одним з підходів зменшення розмірів навчальної вибірки став набір даних ImageNet – широкомасштабна онтологія образів. Вона зіграла неймовірно важливу роль у розвитку комп'ютерного зору та досліджень в області глибинного навчання. ImagetNet побудована як ієрархічна структура, в якій 15 мільйонів розмічений даних розділено на 22 000 категорії. Тобто цей підхід дозволяє використання вагів натренованої моделі на набору даних Imaget-Net та корегування їх під власну задачу шляхом дотренування на своєму наборі даних.

У цій роботі я хочу розглянути метод покращення якості моделі класифікації зображення, що навчена на малій навчальній вибірці шляхом підбору функції втрат.

1.2 Визначення функції втрат.

Функція втрат – це функція, яка характеризує втрати при неправильному прийнятті рішень на основі спостережених даних. Тобто це метод оцінки того, наскільки добре алгоритм моделює вказаний набір даних, наскільки гарно алгоритм працює з заданим набором. Мета функція втрат в нейронній мережі є оцінка та оновлення ваг нейронів з метою поліпшення оцінки на наступному кроці. З цього стає зрозуміло, що правильно обрана функція втрат впливає на точність результуючої моделі після процесу навчання. Значення функції втрат далі будемо називати втратами.

У наступних підрозділах ми розглянемо деякі з них.

1.2.1 Функції втрат L_1 та L_2

Функція L_1 або середня абсолютна похибка – сума абсолютних різниць між цільовими значеннями та прогнозованими змінними:

$$L_1 = \left\| y^{true} - y^{pred} \right\|_1 = \frac{\sum_{i=1}^{n} \left| y_i^{true} - y_i^{pred} \right|}{n}$$

Функція L_2 або середньоквадратична похибка – сума квадратів відстаней між цільовими значеннями та прогнозованими змінними:

$$L_2 = \left\| y^{true} - y^{pred} \right\|_2^2 = \frac{\sum_{i=1}^{n} \left(y_i^{true} - y_i^{pred} \right)^2}{n}$$

Зазвичай ці функції використовуються в задачах регресії і вважається, що їх не слід використовувати у задачах класифікації. Проте як доведено у роботі [2] оцінка L_1, яку застосовують до оцінок ймовірності призводить до мінімізації очікуваної ймовірності неправильної класифікації, оцінка L_2 також мінімізує очікувану ймовірність неправильної класифікації, але регулюється з половиною очікуваного квадрата L_2 норми оцінок ймовірності прогнозів.

Для задачі класифікації функції L_1 та L_2 будуть мати наступний вигляд:

$$L_1 \circ \sigma = \left\| y^{true} - \sigma(y^{pred}) \right\|_1$$

та

$$L_2 \circ \sigma = \left\| y^{true} - \sigma(y^{pred}) \right\|_2^2$$

відповідно, де σ – оцінка ймовірності.

1.2.2 Функція втрат hinge

Завісні втрати (hinge loss) – функція втрат, яка використовується для максимізації розділової класифікації і має таке представлення:

$$hinge = \sum_{j=1}^{n} \max\left(0, \frac{1}{2} - y_j^{true} y_j^{pred}\right),$$

де y_j^{true} приймає значення 0 або 1.

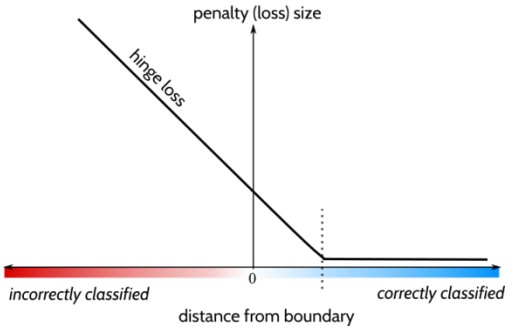

Рис 1. Візуалізація функції витрат hinge loss

На Рисунку 1 ми бачимо візуалізацію роботи функції витрат hinge loss. Вісь X представляє собою відстань від границі будь-якого окремого екземпляра, а вісь Y представляє розмір втрат (штраф). Пунктирна лінія на осі X представляє число 1. Це означає, коли відстань екземпляра від границі більше або дорівнює 1, розмір втрат буде рівним 0. Якщо відстань від границі дорівнює 0 – розмір втрат буде рівним 1. Як можна помітити розмір втрат у правильно класифікованих точках буде невеликий, а неправильно класифіковані точки навпаки будуть мати достатньо великий розмір втрат. Від'ємна відстань від границі призводить до великих втрат на шарнірі, бо це означає, що ми на неправильній стороні границі і екземпляр буде класифіковано неправильно.

1.2.3 Функція втрат cross entropy (log loss)

Функція втрат cross entropy має такий вигляд:

$$H(P, Q) = -\sum_x P(x) \log Q(x),$$

де $P(x)$ – розподіл правильних відповідей, а $Q(x)$ – розподіл ймовірностей прогнозів моделі.

У випадку бінарної класифікації функція втрат cross entropy буде мати наступний вигляд:

$$H_p(Q) = -\frac{1}{N}\sum_{i=1}^{n} y_i \log\left(p(y_i)\right) + (1 - y_i) \log\left(1 - p(y_i)\right))$$

У цьому випадку точки $y = 1$ додаємо $\log\left(p(y)\right)$ (логарифмічну ймовірність того, що $y = 1$) до втрат, а у випадку коли $y = 0$ додаємо $\log\left(1 - p(y)\right)$. Ця функція сприяє наближенню розподілу прогнозування до цільового, штрафуючи не лише за помилкові прогнози, але і за невпевнені. Бінарна cross entropy обчислюється незалежно для кожного компонента вектору передбачених значень і не впливає на значення інших компонент.

У випадку категоріальної класифікації функція втрат cross entropy буде мати наступний вигляд:

$$CE = -\sum_i y_i{}^{true} \log\left(p\left(y_i^{pred}\right)\right)$$

де p – оцінка ймовірності.

Categorical cross entropy визначає міру того, наскільки два дискретних розподіла ймовірностей відрізняються один від одного. Знак мінуса гарантує зменшення втрат, коли розподіли стають ближчими один до одного.

Проте найчастіше функція cross entropy найчастіше зустрічається у інтерпретації softmax.

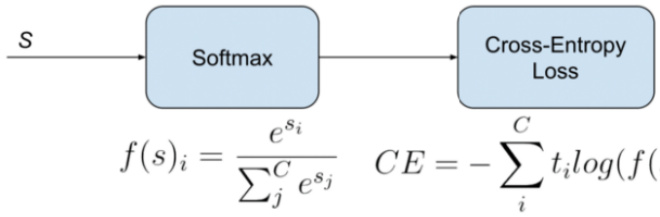

Рис 2. Візуалізація функції втрат softmax

На малюнку 2 ви можете побачити , що функція втрат softmax – це сума значення функції активації softmax та значення функції втрат cross entropy.

1.2.4 Функція втрат softmax з маржею

Перша функція втрат, яка вводіть маржу до функції втрат softmax є L-soft-max. Поле тут має наступну концепцію: воно збільшує віддаленість між класами і, в свою чергу, мінімізує відстань між одними і тим самими класами. Ця внутрішньокласова компактність та міжкласова відокремленість значно підвищує ефективність різноманітних завдань візуальної класифікації та перевірки. Важливо, що у цьому підході також включаємо класифікатор або повністю зв’язаний шар, коли використовуємо термін функції втрат L-Softmax, а не просто функцію активації Softmax та функцію втрат Cross Entopy.

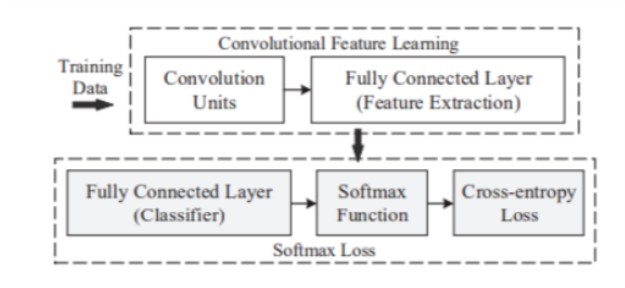

Рис 3. Візуалізація роботи функції втрат L-Softmax

Тоді функція L-softmax буде мати наступний вигляд:

$$L = \frac{1}{N}\sum_i -\log\left(\frac{e^{f_{y_i}}}{\sum_j e^{f_{y_j}}}\right)$$

Вихідне значення повністю зв’язного шару - це лише множення ваг та вихідне значення попереднього шару плюс зміщення. Тут $f_{y_i} = W_{y_i}^T x_i$, де W_{y_i} –

параметри останнього повністю зв'язного шару. Оскільки f – внутрішній добуток між W і x, його також можна сформулювати як $f_j = \|W_j\|\|x_j\| \cos(\theta_j)$, де θ – кут між векторами W і x.

Зважаючи на попередні визначення формула функції softmax приймає такий вигляд:

$$L_i = -\log\left(\frac{e^{\|W_j\|\|x_j\| \cos(\theta_j)}}{\sum_j e^{\|W_j\|\|x_j\| \cos(\theta_j)}}\right)$$

Розглянемо приклад бінарної класифікації. Припустимо ми маємо об'єкт x, який відноситься до класу 1. Функція Softmax буде вимагати $W_1 x > W_2 x$ з метою вірно класифікувати об'єкт до класу 1. Для забезпечення суворішої класифікації та розширення поля прийняття рішень замість $\|W_1\|\|x\| \cos(\theta_1) > \|W_2\|\|x\| \cos(\theta_2)$ необхідно перейти до наступного вигляду $\|W_1\|\|x\| \cos(m\theta_1) > \|W_2\|\|x\| \cos(\theta_2)$, $0 \le \theta_1 \le \frac{\pi}{m}$. Такий підхід дає нам поле для прийняття рішень, що зображено на малюнку 4.

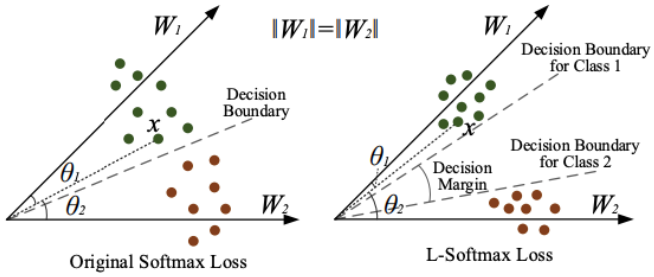

Рис 4. Візуалізація різниці між Softmax та L-Softmax

Тепер функція L-Softmax має вигляд:

$$L_i = -\log\left(\frac{e^{\|W_{y_i}\|\|x_i\| \phi(\theta_{y_i})}}{e^{\|W_{y_i}\|\|x_i\| \phi(\theta_{y_i})} + \sum_{j \ne y_i} e^{\|W_j\|\|x_i\| \cos(\theta_j)}}\right)$$

де $\phi(\theta_{y_i}) = \begin{cases} \cos(m\theta), 0 \le \theta \le \frac{\pi}{m} \\ D(\theta), \frac{\pi}{m} \le \theta \le \pi \end{cases}$

Тут m – ціле додатнє число (чим більше m, тим більше поле класифікації), $D(\theta)$ – монотонно спадаюча функція ($\cos\left(\frac{\pi}{m}\right) = D\left(\frac{\pi}{m}\right)$)

Наступна модифікація функції Softmax є Additive Margin Softmax. Ця модифікація функції пропонує інший підхід до додавання маржі в функцію Softmax. На відміну від множення m до θ, вона вводить поле адитивно шляхом зміни $\phi(\theta) = \cos(\theta) - m$. Additive Margin Softmax нормалізує вагу та зміщення і вводить параметр s, який масштабує значення косинусу. Функція втрати Additive Margin Softmax приймає такий вигляд:

$$L_i = -\frac{1}{N}\sum_{i=1}^{n}\log\left(\frac{e^{s(\cos(\theta_{y_i})-m)}}{e^{s(\cos(\theta_{y_i})-m)} + \sum_{j=1,j\neq y_i}^{c} e^{s\cos(\theta_j)}}\right)$$

Отже межа прийняття рішення буде сформована таким чином:

$$m = \cos(\theta_{W_1,P_1}) - \cos(\theta_{W_1,P_2})$$

де P_1 – ознака класу 1, а P_2 – ознака класу 2 (рис. 5).

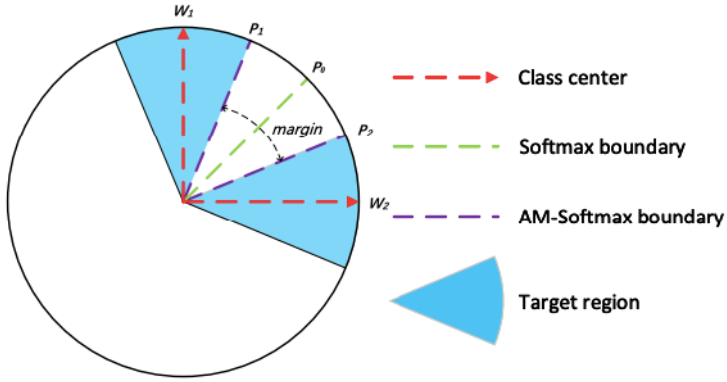

Рис. 5 Звичайна границя прийняття рішень softmax та additive margin softmax

1.2.5 Функція втрат cosine

Косинусова подібність двох векторів a і b розмірності d, $a, b \in \mathbb{R}^d$ базується на куті між цими двома векторами:

$$\sigma_{cos}(a,b) = \cos(a\angle b) = \frac{\langle a,\ b\rangle}{\|a\|_2 * \|b\|_2}$$

де $\langle . \rangle$ – добуток векторів, $\| . \|_p$ - L^p норма.

Функція втрат cosine має наступне представлення:

$$L_{cos}(x,y) = 1 - \sigma_{\cos}(f(\theta), \varphi(y))$$

Вважаємо зміну φ фіксованою та ставимо за мету вивчити параметр θ нейронної мережі $f(\theta)$ шляхом максимізації подібності косинусів між характеристиками зображень та класами.

Чим вище значення косинусної подібності, тим вища значення точності моделі. Цілком протилежний вектор має значення косинусної подібності -1, цілком ортогональний вектор має значення косинусної подібності 0, а повністю ідентичні вектори мають значення 1.

РОЗДІЛ 2. РОЗРОБКА МОДЕЛІ КЛАСИФІКАЦІЇ ЗОБРАЖЕНЬ НА МАЛІЙ НАВЧАЛЬНІЙ ВИБІРЦІ ДАНИХ

2.1 Постановка задачі класифікації зображень, опис використаної моделі та опис навчальних вибірок

В цій роботі я буду аналізувати різні функції втрат, що були описані раніше, на прикладі задачі класифікації зображень. Для дослідження я обрала модель CNN з архітектурою ResNet50. Навчання буде здійснюватися на трьох навчальних вибірках.

Перша навчальна вибірка складається з фотографій погодних умов, що розділені на такі категорії: схід сонця (sunrise), хмарно (cloudy), дощ (rain) та сонячно (shine). Навчальна вибірка розділена на дві множини – train і valid. У множині train знаходиться 897 об’єктів, у множині valid – 226 об’єктів. Розподіл об’єктів по класам наведено у таблиці 1.

	sunrise	cloudy	rain	shine
train	285	240	172	202
valid	72	60	43	51

Табл. 1 Кількість об’єктів у кожній з множин навчальної вибірки Weather по класам

Друга навчальна вибірка складається з фотографій фруктів, розподілені за наступними категоріями: яблуко (apple), банан (banana), декілька фруктів (mixed) та апельсин (orange). Вона розподілена на множини train і valid – 240 та 60 об’єктів відповідно. Розподіл об’єктів по класам наведено у таблиці 2.

	apple	banana	mixed	orange
train	75	73	20	72
valid	19	18	5	18

Табл. 2 Кількість об’єктів у кожній з множин навчальної вибірки Fruits по класам

Третя навчальна вибірка складається з флюорографій легень, які розподілені за такими категоріями: флюорографія легень хворого на covid-19 (Covid-19), здорової людини (healthy) та людини, що хворіє двухстороння пневмонія (bilateral pneumonia). Вона розподілена на множини train і valid – 251 та 66 об'єктів відповідно. Розподіл об'єктів по класам наведено у таблиці 3.

	Covid-19	healthy	bilateral pneu-monia
train	111	70	70
valid	26	20	20

Табл 3. Кількість об'єктів у кожній з множин навчальної вибірки Covid19 по класам

Аналізувати у цій роботі буду наступні функції витрат: Softmax, Cosine loss та Additive Margin Softmax (AM Softmax).

Для покращення значення точності моделі я буду використовувати ваги натренованої моделі на набору даних ImageNet.

2.2 Результати навчання моделі, аналіз отриманих даних

У таблиці 4 наведено значення точності моделей після навчання з використання кожної з обраних функцій втрат.

	Weather dataset	Fruits dataset	Covid datatset
Cosine loss	**98%**	88%	91%
Softmax	97%	**92%**	**94%**
AM_Softmax	97%	90%	89%

Табл. 4. Значення точності моделі після навчання на вибірках даних при зміні функцій втрат (найкращі результати виділено жирним)

З цих результатів ми можемо побачити, що модель, навчена на наборі Weather dataset має найкращий результат параметра точності з використанням функції втрат cosine, а на набора Fruits dataset та Covid dataset – використовуючи функцію soft max.

Тепер розглянемо як змінювалось значення функції втрат в залежності від епох. Як помітно з графіків на малюнку 6 при використанні функцій Softmax величина втрат на тестовій виборці в кінці навчання є більшою, ніж на тренувальних данних, що свідчить про перенавчання моделі. При використанні функції втрат cosine ми бачимо картину досить непоганого навчання моделі, тому точність моделі саме на цій функції, в порівнянні з іншими обраними, і дала найбільшу точність, незважаючи на одну область після 8 епохи збільшення значення втрат на тестовій вибірці.

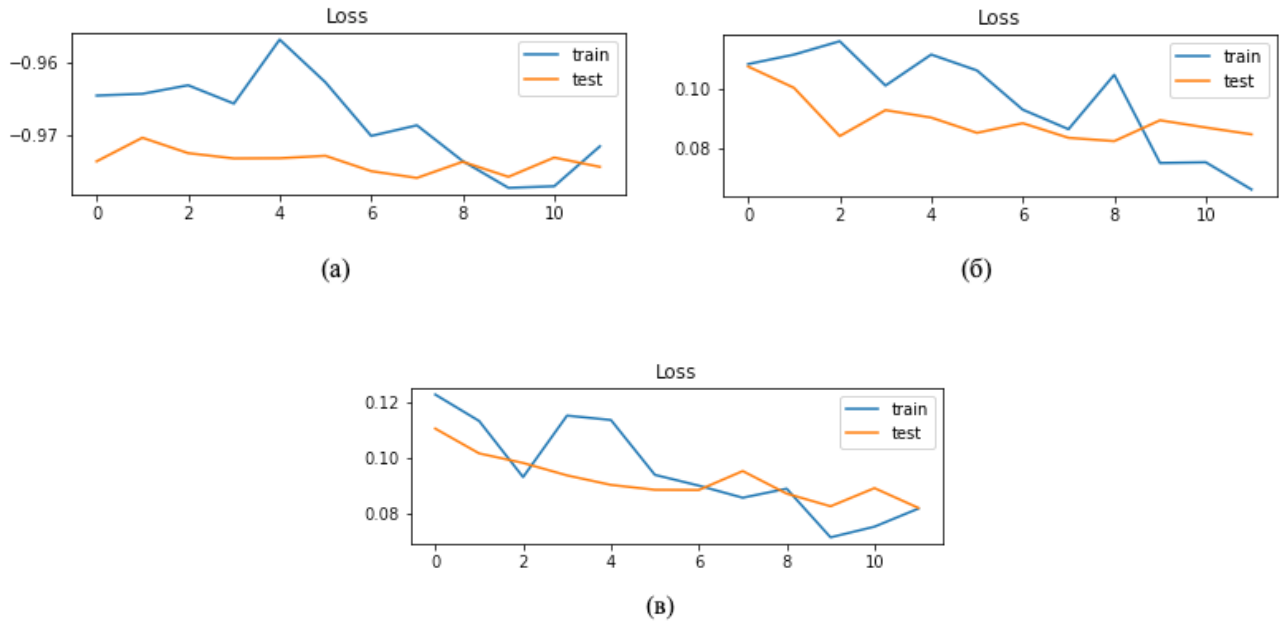

Рис. 6. Графік залежності величини втрат на кожній з епох тренування (вибірка Weather dataset)

а) функція втрат cosine; б) функція втрат Softmax; в) функція втрат (AM Softmax)

Тепер розглянемо результати тренування на вибірці Fruits dataset. У цьому розподілі, що зображений на малюнку 7, використання функції Softmax та AM Softmax показують кращу картину навчання, аніж функція втрат cosine, що є еквівалентно отриманим у ході експерименту точностям моделі.

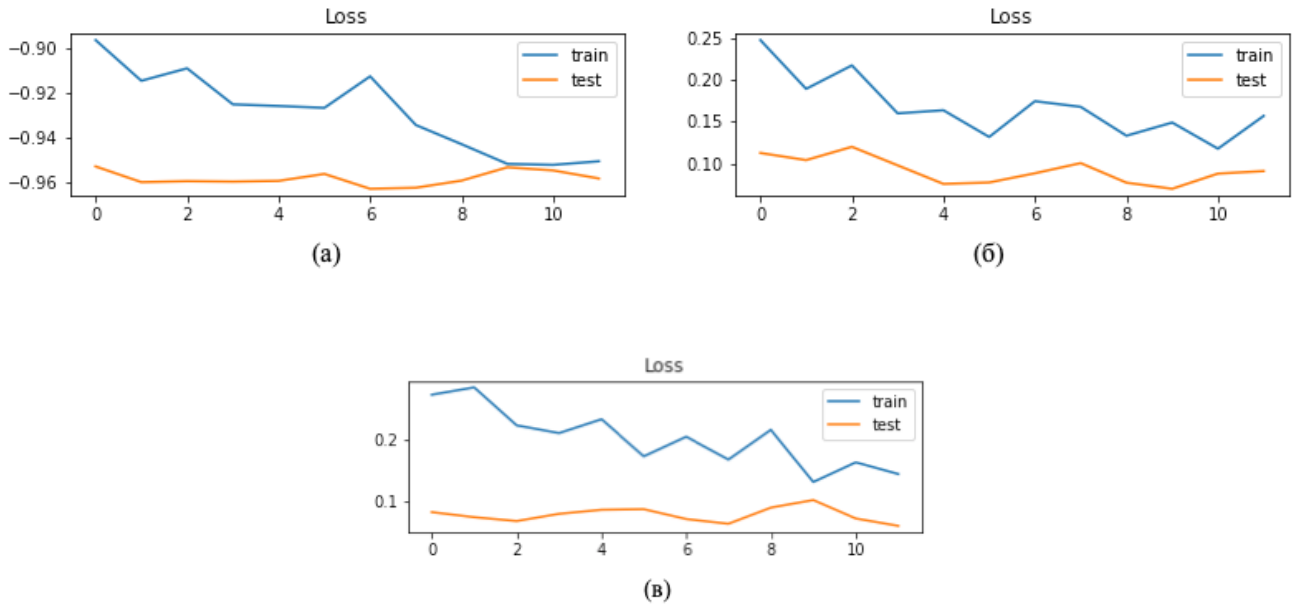

Рис. 7. Графік залежності величини втрат на кожній з епох тренування (вибірка Fruits dataset)

а) функція втрат cosine; б) функція втрат Softmax; в) функція втрат (AM Softmax)

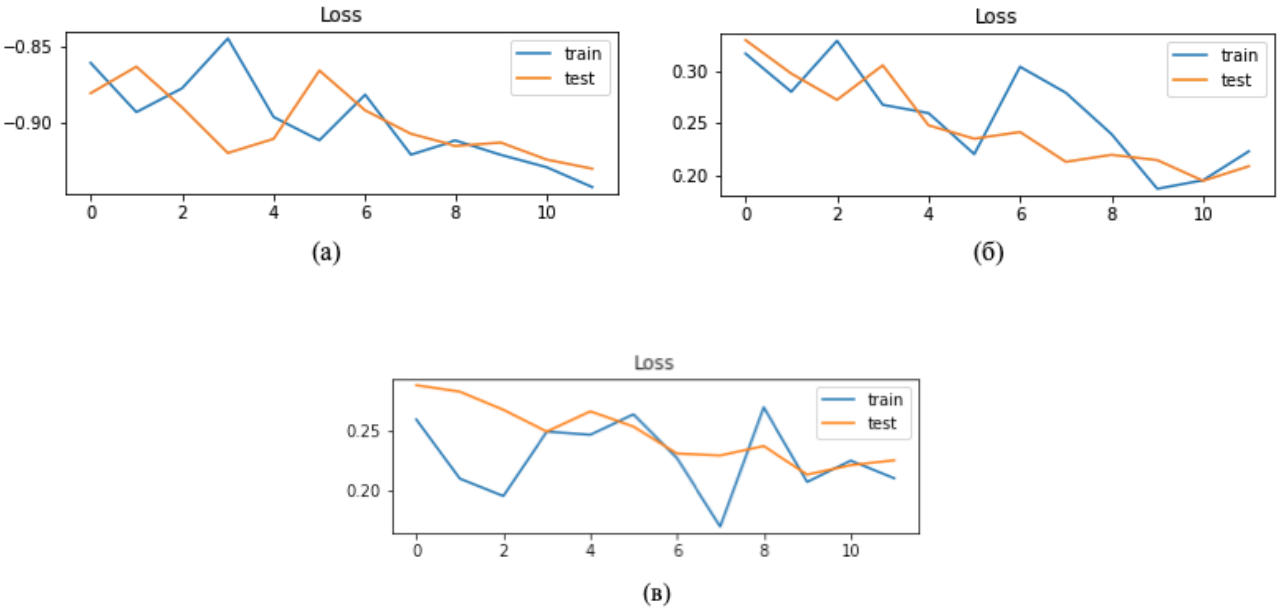

Рис. 7. Графік залежності величини втрат на кожній з епох тренування (вибірка Covid datatset)

а) функція втрат cosine; б) функція втрат Softmax; в) функція втрат (AM Softmax)

На кінець, поглянемо на картину розподілу величини втрат на навчальній вибірці Covid datatset. Знову ми бачимо, що модель, в якій використовується функція втрат Softmax має найкращий розподіл в порівнянні з іншими інтерпретаціями моделі.

Зважаючи на результати цього аналізу функції втрат Cosine та SoftMax показали найкращі результати у задачі класифікації зображень з малою навчальною вибіркою.

ВИСНОВОК

Дана робота складається з двох частин. Перша частина роботи - дослідження існуючих функцій витрат для задач класифікації зображень. У цій частині були розглянуті деякі з існуючих функцій на предмет їх впливу на оцінку якості моделі розпізнавання.

У другій частині роботи на практичному прикладі було проведено детальний аналіз впливу обраних loss функцій на оцінку моделей, що навчені на різних малих навчальних вибірках.

Після виконання цієї роботи стало очевидним, що точність моделі напряму залежить від вибору функції витрат, а задача її вибору має бути однією з основних підзадач вирішення задачі класифікації зображень. Стало зрозумілим, що є висока необхідність імплементації нових винайдених підходів обчислення втрат у загальновідомі бібліотеки для машинного навчання. Зараз найчастіше вживаний підхід покращення точності моделі є збільшення навчальної вибірки та покращення даних у ній. Проте у результаті проведеного експерименту стає зрозумілим, що підхід вибору loss функції під конкретну задачу є досить ефективним.

ЛИТЕРАТУРА

Характеристика джерела	Назва
Електронні ресурси	1. Bjorn Barz, Joachim Denzler. Deep Learning on Small Datasets without Pre-Training using Cosine Loss. In *IEEE Winter Conference on Applications of Computer Vision (WACV) 2020* – доступ до статті: https://arxiv.org/pdf/1901.09054.pdf 2. Katarzyna Janocha, Wojciech Marian Czarnecki. On Loss Functions for Deep Neural Networks in Classification – доступ до статті: https://arxiv.org/pdf/1702.05659.pdf 3. Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A Discriminative Feature Learning Approach for Deep Face Recognition – доступ до статті: https://ydwen.github.io/papers/WenECCV16.pdf 4. Weiyang Liu, Yandong , Zhiding , Meng Yang. Large-Margin Softmax Loss for Convolutional Neural Networks - доступ до статті: https://arxiv.org/pdf/1612.02295.pdf 5. Feng Wang, Weiyang Liu, Haijun Liu, Jian Cheng. Additive Margin Softmax for Face Verification – доступ до статті: https://arxiv.org/pdf/1801.05599.pdf