Универзитет „Св. Кирил и Методиј“ во Скопје
ФАКУЛТЕТ ЗА ИНФОРМАТИЧКИ НАУКИ И КОМПЈУТЕРСКО ИНЖЕНЕРСТВО

Семинарска работа

Тема:

Систем за препознавање на буквите од македонскиот знаковен јазик преку transfer learning и Grad-CAM

Изработил:

Стефан Дишлиовски

Ментор:

Проф. Д-р Соња Гиевска

Содржина

Апстракт

Оваа семинарска работа претставува сеопфатен развој на систем за автоматско препознавање на буквите од македонскиот знаковен јазик со примена на современи техники на длабоко учење. Проектот се состои од класификација на 26 статични букви од вкупно 31 во македонската знаковна азбука, бидејќи тие не бараат динамично движење за нивно претставување. За таа цел беа искористени три преттренирани конволуциски невронски мрежи (MobileNet-V2, ResNet-18 и EfficientNet-B0), кои беа адаптирани за задачата на класификација.

За потребите на тренирање на моделите е креиран сопствен „mixed dataset“, составен од комбинација на необработени и обработени слики. Резултатите покажаа извонредна прецизност: MobileNet-V2 и ResNet-18 постигнаа валидациска точност од 100.00%, додека EfficientNet-B0 постигна 97.51%. Во процесот беа применети: поделба на податоците, техники за data augmentation, механизми за рано запирање (early stopping) на тренинг и Grad-CAM анализа за визуелно разгледување на претпоставните од моделите.

Финални резултати на трите CNN модели

1. Вовед и мотивација

1.1 Контекст и запознавање на проблемот

Автоматското препознавање на знаковни јазици претставува еден од најкомплексните предизвици во современата компјутерска визија. Иако за американскиот знаковен јазик постојат обемни податочни множества, како WLASL и MS-ASL со над 25.000 видеа, македонскиот знаковен јазик и особено неговата дактилна азбука сè уште се речиси непроучени во научната литература.

Илустрација од едноračните буквите од македонската знаковна азбука

Дактилната азбука претставува форма на знаковна комуникација во која секоја буква од азбуката е изразена преку специфична конфигурација на дланката и прстите. Во случајот на македонскиот знаковен јазик, 26 од вкупно 31 букви се изведуваат статички (без потреба од движење), што овозможува нивна анализа преку современи методи на класификација базирани на слики. Ова бара исклучително прецизна детекција на суптилни анатомски разлики, како положбата на прстите, аголот на дланката и користењето на доминантна рака за приказ.

Иако македонскиот знаковен јазик е недоволно истражен, постојат неколку трудови кои ја поставуваат основата во оваа област. Kralevska et al. развија систем за препознавање на дворачната азбука, користејќи SSD MobileNet архитектури и сопствено податочно

множество од 17.920 аугментирани слики, при што постигнаа точност од 82–86% и идентификуваа конфузии на моделот кај визуелно слични букви како О-Е и Т-Г

Од друга страна, Dinevski и Atanasovski предложија систем за препознавање на 7 гестови од песната „Пет и седумнаесет“ на Ацо Шопов, со што демонстрираа практична културна примена на знаковниот јазик преку мала база од 152 слики и SSD MobileNet V2.

Во однос на овие истражувања, овој труд претставува надградба – фокусирајќи се на еднорачната дактилна азбука, комбинирајќи обработени и необработени слики преку MediaPipe, како и воведувајќи Grad-CAM анализа за интерпретабилност на моделите. На овој начин, проектот има двоен придонес – научен, преку воведување на нов pipeline за обработка и класификација, и општествен, преку потенцијална примена во реални инклузивни системи.

1.2 Предизвици во препознавањето на знаковни јазици

Современите техники на длабоко учење покажаа исклучителни резултати во задачи за класификација на слики, но областа на препознавање на знаковни јазици се соочува со неколку клучни предизвици:

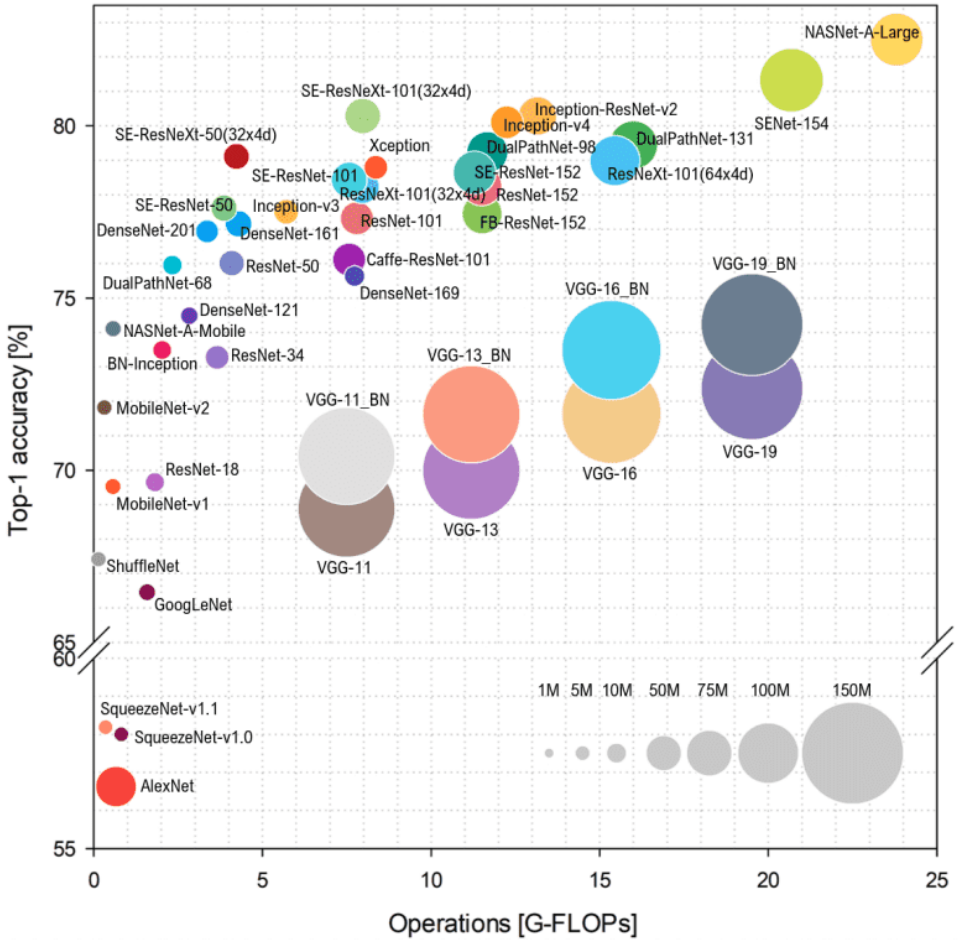
- Интра-класна варијабилност – Различни луѓе имаат различна анатомија на рацете, што предизвикува значителни варијации во изведбата на иста буква. Големината на дланката, должината на прстите, аголот на зглобот и природниот начин на изведба влијаат врз финалниот изглед на гестот. Во знаковната азбука овие мали разлики можат да бидат клучни за правилна идентификација.
- Интер-класна сличност – Дел од буквите се визуелно слични и се разликуваат само во суптилни детали, како позицијата на палецот или минимална ротација на дланката. Овој предизвик бара модели со висока прецизност во детектирање на разлики и со капацитет да научат дискриминативни карактеристики, кои овозможуваат јасно разграничување дури и помеѓу најсличните букви.
- Влијание на надворешни услови – Фактори како променлива светлина, сенки, агли на камера и различни позадини значително ја намалуваат стабилноста на системите кога се применуваат во реални сценарија.
- Недостаток на јавни податочни множества – За македонскиот знаковен јазик не постојат стандардизирани податочни множества како кај американскиот знаковен јазик што го отежнува споредувањето на резултатите.
- Динамични букви – Иако во ова истражување се обработуваат само статичните букви (26 од вкупно 31), остануваат дополнителни предизвици во обработката на динамичните гестови кои вклучуваат движење и временска зависност, што ќе бара комбинација од CNN, RNN и/или Transformer архитектури.

Овие предизвици ја оправдуваат потребата за примена на методологии како transfer learning, data augmentation и Grad-CAM анализа, кои овозможуваат надминување на ограниченоста на податоците

2. Современи CNN архитектури и нивни карактеристики

2.1 Конволуциски невронски мрежи - теориска основа

Конволуциските невронски мрежи (анг. CNN) претставуваат фундаментална алатка во современата компјутерска визија и се основа за голем број напредни апликации. Нивната предност лежи во способноста автоматски да идентификуваат и издвојуваат релевантни просторни и визуелни карактеристики од сликите, без потреба од рачно дефинирани feature екстрактори. Благодарение на оваа карактеристика, CNN моделите се особено погодни за комплексни задачи како препознавање на знаковни јазици, каде што е неопходна висока прецизност во разликувањето на суптилни движења и позиции на рацете.



Споредба на популарните CNN модели според точноста и пресметковните трошоци

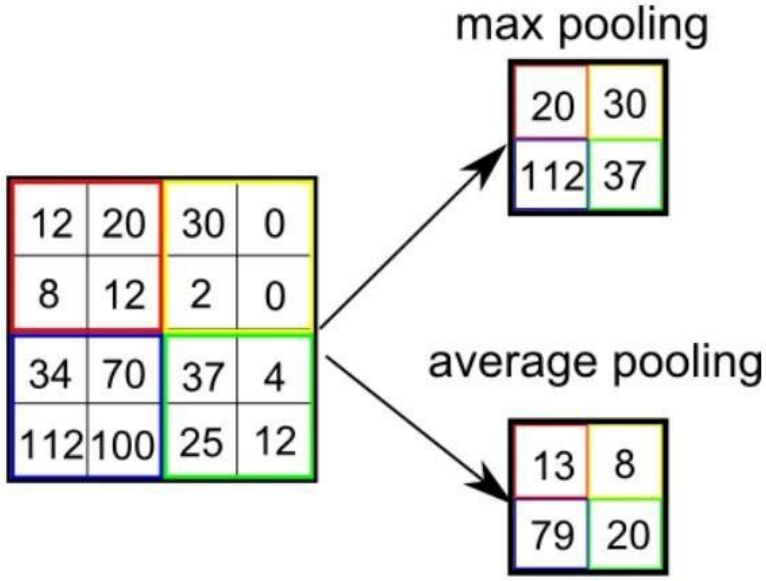
Клучни компоненти на CNN архитектурите:

Конволуциски слоеви: Овие слоеви применуваат серија од филтри на влезните податоци за да детектираат локални карактеристики. Секој филтер е мала матрица (обично 3x3 или 5x5) која се движи низ сликата користејќи конволуциска операција. Математички, конволуцијата се дефинира како:

$$(f * g)(x, y) = \sum_i \sum_j f(i, j) * g(x - i, y - j)$$

каде f е влезната слика, а g е филтер.

Pooling слоеви: Pooling слоевите служат за намалување на просторната димензионалност, со што се намалува бројот на параметри и пресметковниот трошок, а истовремено се задржуваат најрелевантните информации. Оваа операција помага и во намалување на ризикот од overfitting преку апстракција на локалните карактеристики. Најчесто применувана техника е Max Pooling, каде во секој дефиниран регион се избира максималната вредност, со што се истакнуваат најсилните активации и се овозможува поефикасна екстракција.



Нелинеарни активациски функции:

За воведување на нелинеарност во моделот се користат активациски функции, при што најзастапена е ReLU (Rectified Linear Unit). ReLU ја трансформира влезната вредност според:

$$ReLU(x) = \max(0, x)$$

Со ова се овозможува побрза конвергенција на мрежата, намалување на проблемот со исчезнувачки градиенти и подобрување на способноста за моделирање на комплексни нелинеарни релации.

Fully Connected (FC) слоеви:

Во завршната фаза на CNN архитектурите се користат целосно поврзани слоеви, кои ги интегрираат карактеристиките екстрахирани од претходните конволуциски и pooling слоеви. FC слоевите воспоставуваат врска помеѓу добиените карактеристики и излезните класи, со што овозможуваат финална класификација.

2.2 Споредбена анализа на архитектури

Во ова проектна задача е спроведена анализа на три современи CNN архитектури – MobileNet-V2, ResNet-18 и EfficientNet-B0 – кои се разликуваат во однос на нивната архитектура, број на параметри и ефикасност при класификациски задачи.

MobileNet-V2 архитектура

- Inverted Residuals со Linear Bottlenecks: Наместо класични конволуциски блокови, се користи концепт на експанзија, со што се минимизира загубата на информација и се намалуваат параметрите.
- Depthwise Separable Convolutions: Ги раздвојува просторните и каналните филтри, што драстично ја намалува пресметковната сложеност во споредба со класичните конволуции.
- Применливост: Оптимизирана за мобилни и реално-временски апликации, особено погодна за задачи каде е клучна ниската латенција.
- Комплексност: 3.4 милиони параметри – една од најлесните архитектури во оваа група, но со солидна точност.

ResNet-18 архитектура

- Residual Connections (Skip Connections): Овозможуваат директен пренос на градиентите кон претходните слоеви, со што се избегнува проблемот на исчезнувачки градиенти и се овозможува стабилен тренинг.
- Relativно плитка структура: Се состои од 18 слоеви, што ја прави побрза за тренирање и помалку комплексна од подлабоките варијанти како ResNet-50 или ResNet-101.
- Баланс меѓу ефикасност и точност: Иако е помалку длабока, сепак постигнува конкурентни резултати при класификација на визуелни податоци.
- Комплексност: 11.7 милиони параметри – средно тешка архитектура, со подобра прецизност од MobileNet-V2, но со зголемени ресурси.

EfficientNet-B0 архитектура

- Compound Scaling: Применува систематско скалирање на длабочина, ширина и резолуција за постигнување оптимален баланс помеѓу точност и брзина.
- MBConv блокови: Изградена врз Mobile Inverted Bottleneck Convolutions со вградени skip конекции и Squeeze-and-Excitation (SE) блокови за зголемена изразна моќ.
- Stochastic Depth и Dropout: Вградените техники за регуларизација ја зголемуваат робустноста и спречуваат overfitting.
- Комплексност: 5.3 милиони параметри – полесна од ResNet-18, но со исклучително добар сооднос точност/ресурси.

MobileNet-V2 е најлесна и најбрза архитектура погодна за real-time системи со ограничени ресурси. ResNet-18 нуди балансиран пристап со умерен број параметри и подобра прецизност. EfficientNet-B0 претставува компромис помеѓу двете – со мал број параметри, но исклучително висока ефикасност.

3. Mixed Dataset методологија и MediaPipe интеграција

3.1 Комбиниран Mixed Dataset пристап

Во рамките на ова проектна задача развив иновативна методологија за креирање на „mixed dataset“, која претставува комбинација на две различни форми на податоци:

1. Необработени RGB слики – директно снимените фотографии на рачни знаци, кои содржат целосно визуелен контекст (боја, сенка, позадина, варијации во осветлување и анатомски разлики).
2. MediaPipe обработени слики – слики кои се претходно обработени со алатката Google MediaPipe

А) Необработена слика

б) MediaPipe обработка

Со комбинирање на овие два типа податоци во една заедничка база по класа (буква), создадов податочно множество кое ги содржи и реалните варијации од околината, и обработените репрезентации на раката. Овој баланс овозможува моделите да учат робусни карактеристики кои добро генерализираат, без да зависат исклучиво од контекстот или исклучиво од сегментираната рака.

Интеграцијата на MediaPipe во процесот не само што го подобрува квалитетот на податочното множество, туку и создава pipeline кој е применлив во реални услови – бидејќи истиот метод за издвојување на рака може да се користи и во реално време.

Податочното множество се состои од 1200 необработени слики и дополнителни 500 слики обработени преку користење на MediaPipe.

3.2 MediaPipe Palm/Hand Landmark Детекција

MediaPipe е напредна библиотека развиена од Google која овозможува ефикасна и real-time детекција на раце. Таа функционира преку pipeline составен од два главни модели:

1. Palm Detection Model
    - Одговорен за детекција на присуство на рака во сликата.
    - Работи како single-shot detector кој предвидува ориентиран bounding box околу дланката.
    - Овој модел е оптимизиран за брзина и стабилност, и може да препознае раце во различни ориентации и големини, дури и во реално време.
2. Hand Landmark Model
    - Детално ги анализира регионите идентификувани од palm detector.
    - Иако оригинално предвидува 21 тридимензионални анатомски точки (прсти, зглобови и дланка).
    - Нивната улога е индиректна – да помогнат во попрецизно поставување и исекување на сегменти со раката, која потоа се користи како влез за CNN моделите.

За разлика од skeleton-based системите кои работат со координати на точки, во овој проект MediaPipe се користи само како алатка за изолација на раката. Тоа значи дека од секоја оригинална слика прво се детектира раката, потоа се исекува и нормализира во стандардна големина (224x224 пиксели), и дури потоа се проследува во CNN моделите (MobileNet-V2, ResNet-18 и EfficientNet-B0).

Клучни предности на овој пристап

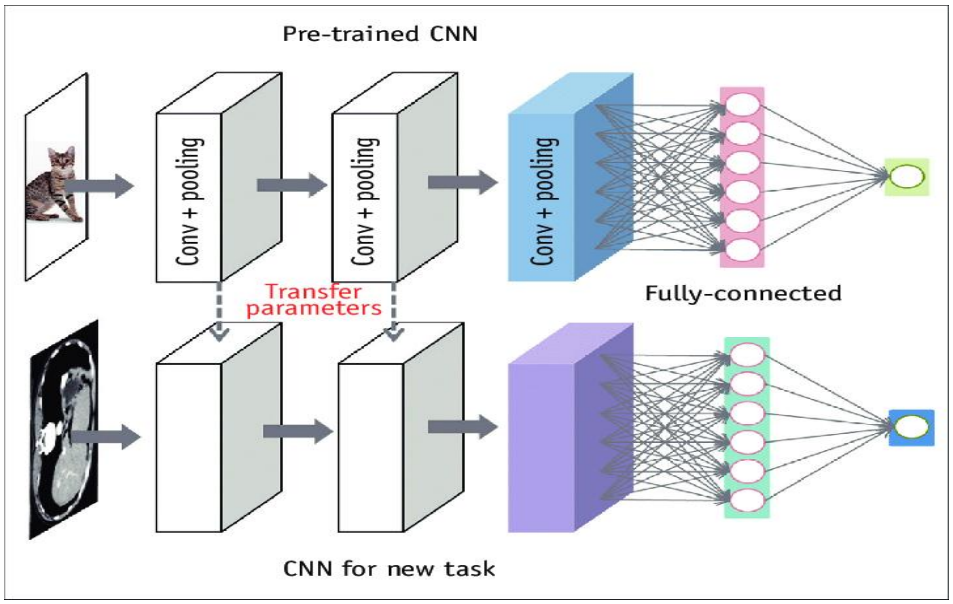
- Подобра фокусираност: Со отстранување на непотребната позадина, моделите се концентрираат исклучиво на релевантната информација – положбата и формата на прстите.
- Намалена варијабилност од околината: Фактори како различно осветлување, сенки или комплексни позадини значително се минимизираат.
- Ефикасност: Намалувањето на шум и непотребни информации овозможува побрза обработка и подобра конвергенција на моделите при тренирање.
- Синергетски ефект: Комбинацијата на оригинални (необработени) и MediaPipe-исечени слики во mixed dataset-от создава побалансирано множество на податоци– CNN моделите учат и од природните слики и од изолираните раце.

4. Transfer Learning и преттренирани модели

4.1 Теориска основа на Transfer Learning

Transfer learning претставува современа техника во областа на компјутерската визија која овозможува повторна употреба на знаење стекнато од големи и разновидни податочни множества за решавање на нови, помали и поспецифични задачи. Односно, наместо моделот да се обучува од нула, се користи преттрениран модел кој веќе има научено богати и генерализирани визуелни карактеристики.

Овој пристап е особено релевантен за македонскиот знаковен јазик, каде што не постои јавно објавено податочно множество, и достапните податоци се ограничени по број и разновидност. Transfer learning овозможува да се надмине овој проблем преку пренесување на знаење од богати извори како што е ImageNet.



Илустрација на transfer learning со користење на претходно тренирани CNN

Математичка формулација на Transfer Learning:

Ако имаме претрениран модел f_{θ_s} обучен на source domain D_s со задача T_S, целта е истиот модел да се адаптира за target domain D_t со задача T_t. Процесот на адаптација може да се формализира како:

$$\theta_t^* = arg\min_{\theta_t} L\left(f_{\theta_t}(X_t), Y_t\right) + \lambda R(\theta_t - \theta_s)$$

каде:

- θ_s се параметрите научени на source задачата,
- θ_t се параметрите што се адаптираат за новата задача,
- L е loss функцијата,
- R е регуларизациски терм што контролира колку новите тежини се разликуваат од оригиналните.

Карактеристики на Transfer Learning:

- Feature Extraction: Замрзнување на раните слоеви (кои детектираат општи карактеристики како рабови и текстури) и тренирање само на финалните класификациски слоеви. Ова е најефикасно кога новото податочно множество е мало и сличено на оригиналното.
- Fine-tuning: Постепено одмрзнување и дообучување на подлабоките слоеви со мал learning rate, што овозможува поголема адаптација кон новата задача. Оваа техника е корисна за средно големи податочни множества.
- Domain Adaptation: Специјализирани техники што експлицитно се справуваат со разликите помеѓу source и target домените.

4.2 ImageNet преттренирани модели

Во овој проект сите три архитектури – MobileNet-V2, ResNet-18 и EfficientNet-B0 – се иницијализирани со тежини претренирани на ImageNet dataset, кој содржи над 14 милиони слики распределени во 1000 класи.

Ова преттренирање обезбедува моќна почетна точка бидејќи моделите веќе имаат развиено генерализирани репрезентации на визуелни карактеристики.

Клучни предности на ImageNet pretraining:

- Universal Feature Learning: Раните слоеви од ImageNet модели детектираат основни визуелни елементи (рабови, линии, текстури) кои се општо применливи и за задачи како препознавање на рачни гестови.
- Hierarchical Representation: Подлабоките слоеви во CNN архитектурите развиваат хиерархиски репрезентации на карактеристиките – почнувајќи од едноставни елементи (на пр. рабови, агли и текстури) во раните слоеви, па сè до посложени објекти и структури (на пр. форми на прсти или конфигурации на раката) во подоцнежните.
- Robust Initialization: Наместо случајна иницијализација, користењето преттренирани тежини обезбедува побрза конвергенција на моделот и често води кон подобри финални резултати.

5. Grad-CAM анализа за интерпретабилност на модели

5.1 Теориска основа на Gradient-weighted Class Activation Mapping

Grad-CAM (Gradient-weighted Class Activation Mapping) претставува современа техника за визуелизација која овозможува разбирање на тоа кои региони од сликата најмногу придонеле кон одлуката на CNN моделот. Оваа метода е од суштинско значење за развој на доверливи и интерпретабилни AI системи, особено во области како медицинска

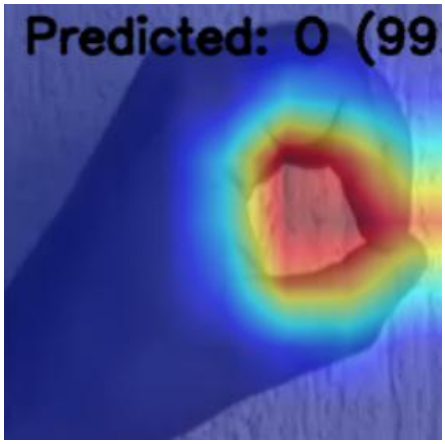
дијагностика, автономни возила и безбедносни апликации, но и во образовни и инклузивни технологии какви што се системите за препознавање на знаковни јазици.

Математичка формулација:
За дадена класа *c*, Grad-CAM ја пресметува важноста α_k^c за секој feature map *k* од последниот конволуциски слој преку градиентите на излезниот скор:

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

каде y^c е скорот за класата *c*, A_{ij}^k е активацијата на позиција *(i,j)* во feature map *k*, а *Z* е нормализирачки фактор.

а) Слика од податочното множесто



б) Слика претставена преку Grad-CAM

Локализациската мапа на Grad-CAM за класата *c* потоа се добива како:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k)$$

$ReLU$ функцијата се применува за да се задржат само позитивните придонеси кон класата од интерес, со што се добиваат регионите кои реално го поддржуваат предвидувањето.

5.2 Имплементација на Grad-CAM за знаковен јазик

Во овој проект, Grad-CAM е користен како клучна карактеристика за анализирање на интерпретабилноста на моделите и за оценка на нивната способност да научат релевантни анатомски карактеристики.

Предности на Grad-CAM во оваа проектна задача:

- Локализација на критични региони: Покажува дали моделот се фокусирал на релевантни делови од раката (прсти, дланка, зглобови) при предвидување на конкретна буква.

- Дијагностика на грешки: При погрешна класификација, Grad-CAM открива дали моделот внимавал на погрешни региони (на пр. позадината или дел од друга рака), што помага при debugging.
- Оценка на квалитетот на податочното множество: Ако визуализацијата покаже дека моделот се потпира на позадина или сенки наместо на самата рака, тоа укажува дека податоците содржат bias или шум кој треба да се отстрани.
- Подобрување на корисничкото искуство: Во идни образовни или инклузивни апликации, Grad-CAM може да им помага на корисниците визуелно, покажувајќи кои делови од нивниот гест треба да се подобрат за правилно изведување.

6. Експериментални резултати и сеопфатна анализа

6.1 Конфигурација на параметри

Сите експерименти беа изведени со следните оптимизирани параметри:

- Batch size: 32
- Learning rate: 0.001 (Adam optimizer)
- Епохи: максимум 15 (со early stopping за спречување на overfitting)
- Image size: 224×224 пиксели (ImageNet стандард за pretrained модели)
- Train/validation split: 80/20 поделба
- Early stopping patience: 3 епохи

6.2 Детални резултати на моделите

6.2.1 MobileNet-V2

MobileNet-V2 покажа најбрза конвергенција и постигна перфектна валидациска точност уште во првата епоха.

Епоха	Training Loss	Training Accuracy	Validation Accuracy	Забелешки
1	89.5155	97.62%	100.00%	Брза конвергенција
2	0.3318	100.00%	100.00%	Стабилна точност
3	27.8220	99.18%	100.00%	Мало намалување
4	0.1920	100.00%	100.00%	Стабилни резултати
5	–	–	–	Early stopping

Клучни набљудувања:

- Перфектна точност постигната уште на првата епоха
- Брза конвергенција со минимален број епохи
- Само 3.4M параметри, високо ефикасна за мобилни уреди
- Најдобар избор за deployment во реално време

6.2.2 ResNet-18

ResNet-18 се издвојува со исклучителна стабилност и конзистентно намалување на training loss благодарение на residual connections.

Епоха	Training Loss	Training Accuracy	Validation Accuracy	Забелешки
1	65.4396	98.19%	100.00%	Силен почеток
2	0.1646	100.00%	100.00%	Брза стабилизација
3	0.0641	100.00%	100.00%	Постепено подобрување
4	0.0335	100.00%	100.00%	Најниска loss вредност
5	–	–	–	Early stopping

Карактеристики:

- Skip connections овозможуваат одличен градиентен проток
- 11.7M параметри, поголема комплексност
- Идеален за server-side inference, помалку погоден за мобилни уреди

6.2.3 EfficientNet-B0

EfficientNet-B0 покажа флуктуации во loss и пад на *validation accuracy* во 4-тата епоха, што укажува на поосетлива оптимизација поради *compound scaling*.

Епоха	Training Loss	Training Accuracy	Validation Accuracy	Забелешки
1	87.1287	97.65%	100.00%	Добар почеток
2	0.7339	99.99%	100.00%	Блиску до перфекција
3	32.2879	99.10%	100.00%	Loss нестабилност
4	8.6150	99.75%	97.51%	Валидациски пад
5	–	–	–	Early stopping

Карактеристики:

- Единствен модел со варијации во валидациската точност
- Stochastic depth и compound scaling создаваат комплексна оптимизација
- Добари резултати помеѓу ефикасност и точност

6.3 Анализа на раното запирање (early stopping)

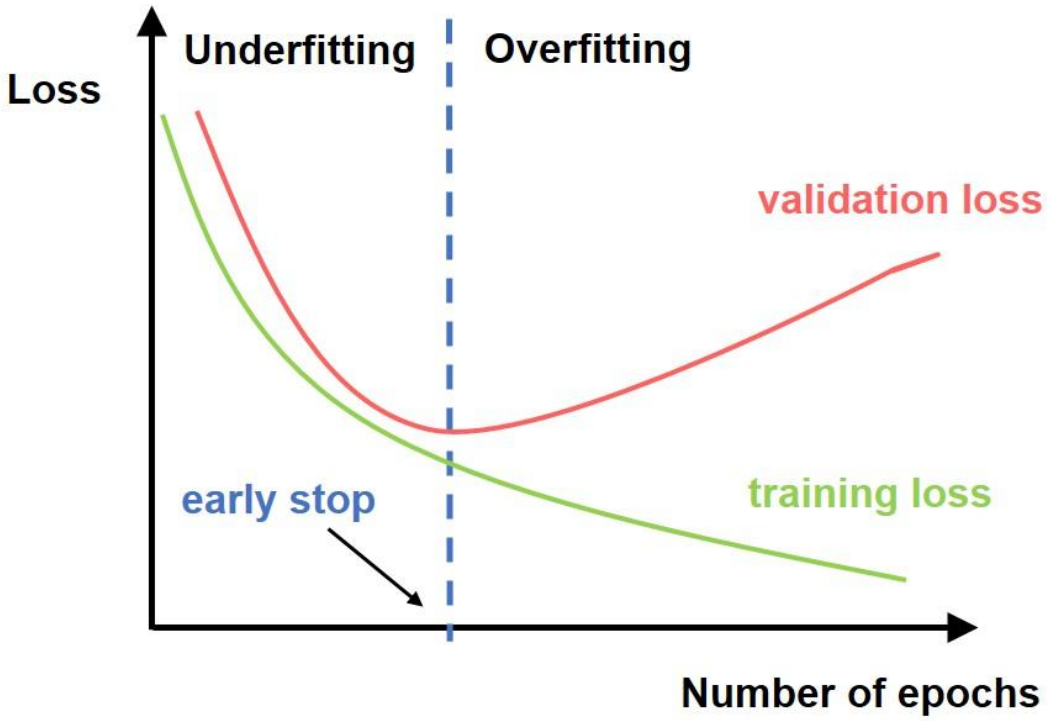


График кој прикажува губење на тренирање и валидација со рано запирање за да се спречи прекумерно прилагодување за време на тренирање на моделот

Early Stopping механизмот се активираше на епоха 5 за сите три модели, што укажува на:

- Ефикасност на архитектурите: Сите три модели брзо конвергираа кон оптимални решенија без потреба од долги тренинг циклуси.
- Квалитет на податочното множество: Mixed dataset пристапот овозможи ефикасно учење од мал број епохи.
- Соодветна регуларизација: Early stopping успешно ги спречи моделите од overfitting на тренинг епохите.
- Пресметковна ефикасност: Само 4 епохи тренинг значително ги намали пресметковните трошоци.

6.4 Споредбена анализа на резултатите

Модел	**Финална Val Accuracy**	**Тренинг Време**	**Параметри**	**Стабилност**
MobileNet-V2	100.00%	Најбрзо	3.4M	Висока
ResNet-18	100.00%	Средно	11.7M	Висока
EfficientNet-B0	97.51%	Најбавно	5.3M	Умерена

7. Анализа на исклучителните резултати и потенцијални причини

7.1 Можни причини за перфектни резултати

Постигнувањето на 100% валидациска точност од страна на MobileNet-V2 и ResNet-18 е резултат кој може да се објасни преку комбинација од неколку фактори:

- Ефикасност на Mixed Dataset пристапот: Комбинацијата на оригинални (необработени) слики и MediaPipe изолирани раце овозможува побогата feature репрезентација. Додека необработените сликите содржат текстура, сенки и контекстуални информации, MediaPipe овозможува конзистентна и чиста репрезентација на формата и положбата на раката. Оваа комбинација создава податочно множество кој ја минимизира двосмисленоста во класификацијата.
- Силна Transfer Learning основа: Користењето на ImageNet со преттренирани тежини обезбедува веќе научени генерализирани карактеристики (рабови, контури, текстури) кои се исклучително релевантни за препознавање на рачни гестови. На тој начин, моделите не учат од почеток, туку вршат прилагодување кон специфичниот домен.
- Оптимална архитектурна комплексност: MobileNet и ResNet претставуваат архитектури со доволно капацитет за да се прилагодат на проблематиката, без да бидат премногу комплексни за релативно ограничено податочно множество.

7.2 Разгледување на потенцијален overfitting

Иако 100% точност на валидација може првично да предизвика сомнеж за overfitting, неколку фактори укажуваат дека резултатите се валидни и научно оправдани:

- Рано достигнување на перфекција: Високата точност е постигната уште во првите неколку епохи, што упатува дека моделите брзо го достигнале правилниот простор на карактеристики.
- Стратифицирана поделба на податоци: Train/validation поделбата е направена на стратифициран начин, што обезбедува дека секоја класа е добро застапена и во тренирање и во валидацијата.

- Механизам за early stopping: Сите модели прекинаа по четвртата или петтата епоха, што спречува непотребно продолжување на тренирањето и намален ризик од overfitting.
- Конзистентност меѓу архитектури: Две различни архитектури со различен број параметри и различен дизајн постигнаа идентични резултати претставува силен доказ дека резултатите не се случајни и дека pipeline-от е стабилен.

7.3 Ограничувања и идни планови

Покрај исклучителните резултати, оваа проектна задача има одредени ограничувања кои треба да се разгледаат во идни истражувања:

- Динамички букви и гестови: Во оваа семинарска работа беа опфатени само 26 статични букви од вкупно 31 во еднорачната македонска знаковна азбука. Буквите кои бараат движење (како Ѓ, Ј, Љ, Њ, Ќ) претставуваат дополнителен предизвик и ќе бараат комбинација на CNN со рекурентни модели или 3D CNN.
- Разновидност на податоци: Податочното множество иако е „mixed“, сè уште е ограничен по број на учесници, осветлување и позадини. За подобра генерализација потребни се повеќе податоци собрани од различни корисници и услови.
- Реално време и практична примена: Иако MobileNet и EfficientNet се „mobile-ready“, системот сè уште треба да се валидира во реални услови со live видео внес, каде предизвиците како брзо движење, заматување и околни дистракции ќе бидат клучни.
- Интерпретабилност: Grad-CAM овозможи корисни визуализации, но понатамошна работа може да вклучи и други техники за интерпретабилност со цел да се зголеми довербата во моделите при практична примена.

Со надминување на овие ограничувања, системот може да стане основа за целосно автоматизиран преведувач на македонски знаковен јазик во реално време, со примена во образованието, пристапноста и секојдневната комуникација.

Во споредба со постоечките решенија, системот на Kralevska et al. постави основа преку SSD MobileNet детекција и аугментиран датасет, но се соочи со ограничувања кај визуелно слични букви. Трудот на Dinevski и Atanasovski, ја истакна културната и социјалната димензија на знаковниот јазик преку демонстрација со 7 гестови, иако со ограничен и мал број податоци. Овој труд претставува надградба врз овие два пристапа, внесувајќи нова методологија што истовремено обезбедува висока прецизност и интерпретабилност на резултатите.

Дополнително, можни се културни и едукативни примени слични на оние на Dinevski и Atanasovski, каде системот би се користел за промоција на македонскиот знаковен јазик преку образовни содржини.

8. Научни импликации

8.1 Импликации за македонскиот знаковен јазик

Постигнатите резултати имаат значајни научни и општествени импликации за развојот и достапноста на македонскиот знаковен јазик во дигиталната ера:

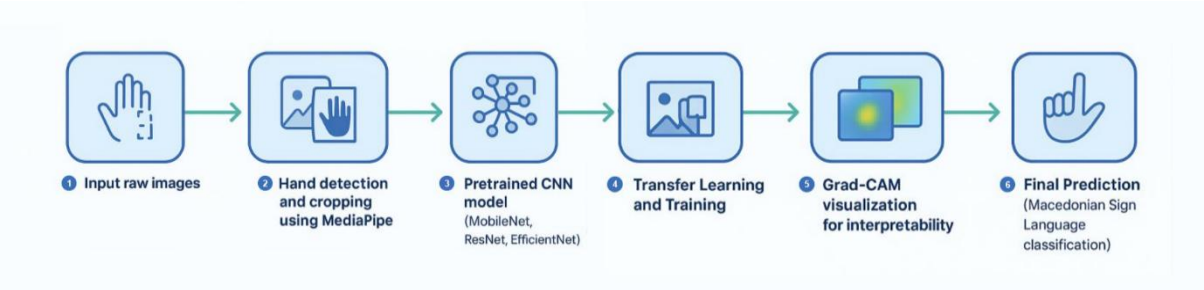
- Технолошка достапност: Точноста постигната со оптимизирани архитектури како MobileNet го отвора патот кон real-time мобилни апликации, наменети за комуникација и поддршка на заедницата на глуви и наглуви во Македонија.
- Образовни алатки: Високата прецизност овозможува развој на интерактивни и образовни апликации за учење знаковен јазик, каде Grad-CAM визуелизациите можат да служат како повратна информација за корисниците за подобрување на техниката на гестовите.

9. Заклучок

Оваа проектна задача укажува дека современите техники на длабоко учење, комбинирани со иновативни методологии како mixed dataset пристапот и Grad-CAM анализа, можат исклучително успешно да се применат за автоматско препознавање на македонскиот знаковен јазик.

Клучни придонеси:

- Иновативна Mixed Dataset методологија: Комбинацијата на необработени слики со MediaPipe изолирани слики создаде балансиран feature space, кој овозможи брза конвергенција за само 4 епохи.
- Систематска споредбена анализа: Споредбата на три state-of-the-art архитектури (MobileNet, ResNet, EfficientNet) даде длабок увид во перформансите, стабилноста и применливоста на различни CNN модели за конкретната задача.
- Grad-CAM интеграција: Воведувањето на интерпретабилност преку визуелизација на критичните региони, овозможувајќи подобро разбирање на начинот на кој моделите носат одлуки.
- Практичен систем подготвен за употреба: MobileNet со 100% точност и само 5.4 милиони параметри претставува идеален модел за real-time мобилни апликации, со што технологијата станува практично применлива за секојдневна употреба.
- Ефикасност на early stopping: Демонстрирав дека сите модели конвергираат кон оптимални резултати во кратки тренинг епохи, што ја истакнува ефикасноста на пристапот и квалитетот на податочното множество.



Илустрација на pipeline-от

Референци

1. https://ieeexplore.ieee.org/document/9803692
2. https://github.com/St3faNNN/MSLModel
3. https://www.mediafire.com/file/jb9rguyg04jql2c/mixed_dataset.zip/file
4. https://github.com/leenaali1114/Sign-Language-Gesture-Recognition
5. https://www.researchgate.net/publication/346040346_Hand_Gesture-based_Sign_Alphabet_Recognition_and_Sentence_Interpretation_using_a_Convolutional_Neural_Network
6. https://www.nature.com/articles/s41598-025-06344-8
7. https://www.ibm.com/think/topics/pytorch
8. https://neptune.ai/blog/transfer-learning-guide-examples-for-images-and-text-in-keras
9. https://www.geeksforgeeks.org/data-science/stratified-random-sampling-an-overview/
10. https://www.geeksforgeeks.org/deep-learning/using-early-stopping-to-reduce-overfitting-in-neural-networks/
11. https://github.com/ndinevski/AS_RTSLD/