

RAČUNARSKA INTELIGENCIJA KNOWLEDGE DISTILLATION

Seminarski rad

Andjela Živanović 4010/2023
Vuk Antov 4014/2024

ZAŠTO RADIMO KNOWLEDGE DISTILLATION?

Savremene duboke mreže su velike, spore i zahtevaju skupo hardversko okruženje

Potreba za modelima pomoću kojih bi model funkcionisao u realnom vremenu i na hardveru ograničenih resursa;

Problem: kako dobiti mali model koji ima performanse velikog?
Cilj je očuvanje sposobnosti generalizacije koje veliki modeli poseduju.

1

2

3

Knowledge Distillation se pojavljuje kao jedan od najuspešnijih načina da se na tumačenju kompresuju veliki modeli bez značajnog gubitka u tačnosti.

KNOWLEDGE DISTILLATION

Osnovni koncept Knowledge Distillation-a:

- Učitelj (složeniji, precizniji i dublji model) → proizvodi „meku“ distribuciju verovatnoća
- Student (manji, jednostavniji) → pokušava da imitira učitelja, ne samo da uči od tačnih labela. Uči od „mekših“ raspodela verovatnoća koje učitelj generiše primenom softmax funkcije sa povišenom temperaturom.

Prednosti KD pristupa:

- Mekane raspodele sadrže više informacija od tvrdih oznaka – pokazuju sličnosti između klasa.
- Student može da nauči suptilnije odnose koje tvrde oznake ne sadrže.
- Dobija se model koji je znatno lakši, ali zadržava ključne karakteristike učitelja.

TIPOVI DESTILACIJE

- **Response-based distillation:** najjednostavniji oblik, u kom se prenosi izlazna raspodela učitelja. Naglasak je na logitima na izlazu mreže.
- **Feature-based distillation:** prenos karakteristika iz međuslojeva. Učiteljeve aktivacije postaju cilj za učenikove slojeve. Ova tehnika daje dublji prenos znanja, ali je složenija za implementaciju.
- **Relation-based distillation:** student uči odnose između uzoraka, slojeva ili aktivacija. Ovaj pristup pokušava da „uhvati“ strukturu učiteljskog znanja.

TEORIJSKA POZADINA I FORMULACIJA KD LOSS

U KD-u se koristi kombinacija dve komponente:

1. **Kullback–Leibler** divergencija između logita učitelja i studenta pri povišenoj temperaturi T;
2. Klasična **Cross-Entropy** na bazi stvarnih labela.

Objašnjenje parametara:

- **Temperatura (T)**: Povećava „mekoću“ raspodele. Više vrednosti T čine da učitelj daje bogatije signale o međusobnoj sličnosti klase.
- **Koeficijent α** : Određuje koliko KD komponenta utiče na ukupan gubitak. Vrednost $\alpha=0.9$ znači da se mnogo više oslanjamo na učitelja.

Ova formula omogućava da student istovremeno uči i od stvarnih oznaka, i od znanja koje učitelj „distilira“ kroz mekše logite.

EARLY-STOPPED KNOWLEDGE DISTILLATION (ESKD)

Previše dobro obučen učitelj često generiše „previše oštре“ raspodele (veoma uverene – jedna klasa dobija skoro 1.0, ostale 0.0).

Ključna ideja ESKD-a:

- Učitelj ne treba da bude potpuno konvergovan;
- Snapshot ranijih epoha (npr. e3 ili e6) proizvodi „šire“, bogatije raspodele;
- Student uči bolje od učitelja koji nije prenaglašeno uveren u svoj izbor klase



ARHITEKTURA SISTEMA I ORGANIZACIJA KODA

- **Trening učitelja (train_teacher.py):**

Obezbeđuje treniranje ResNet modela uz snimanje snapshot-ova na odabranim iteracijama. Omogućava kontrolu hiperparametara, scheduler-a, optimizatora i logovanja.

 - **Trening studenta (train_student_kd.py):**

Implementira kompletan KD→CE proces. Omogućava izbor studenta, učitelja, checkpoint-a, broja KD iteracija i drugih parametara.
 - **Sweep sistem (run_sweep.py):**

Automatizuje pokretanje velikog broja eksperimenata, snima logove i rezultate, čime omogućava sistematsku analizu u zavisnosti od snapshot-a, a, T i drugih parametara.



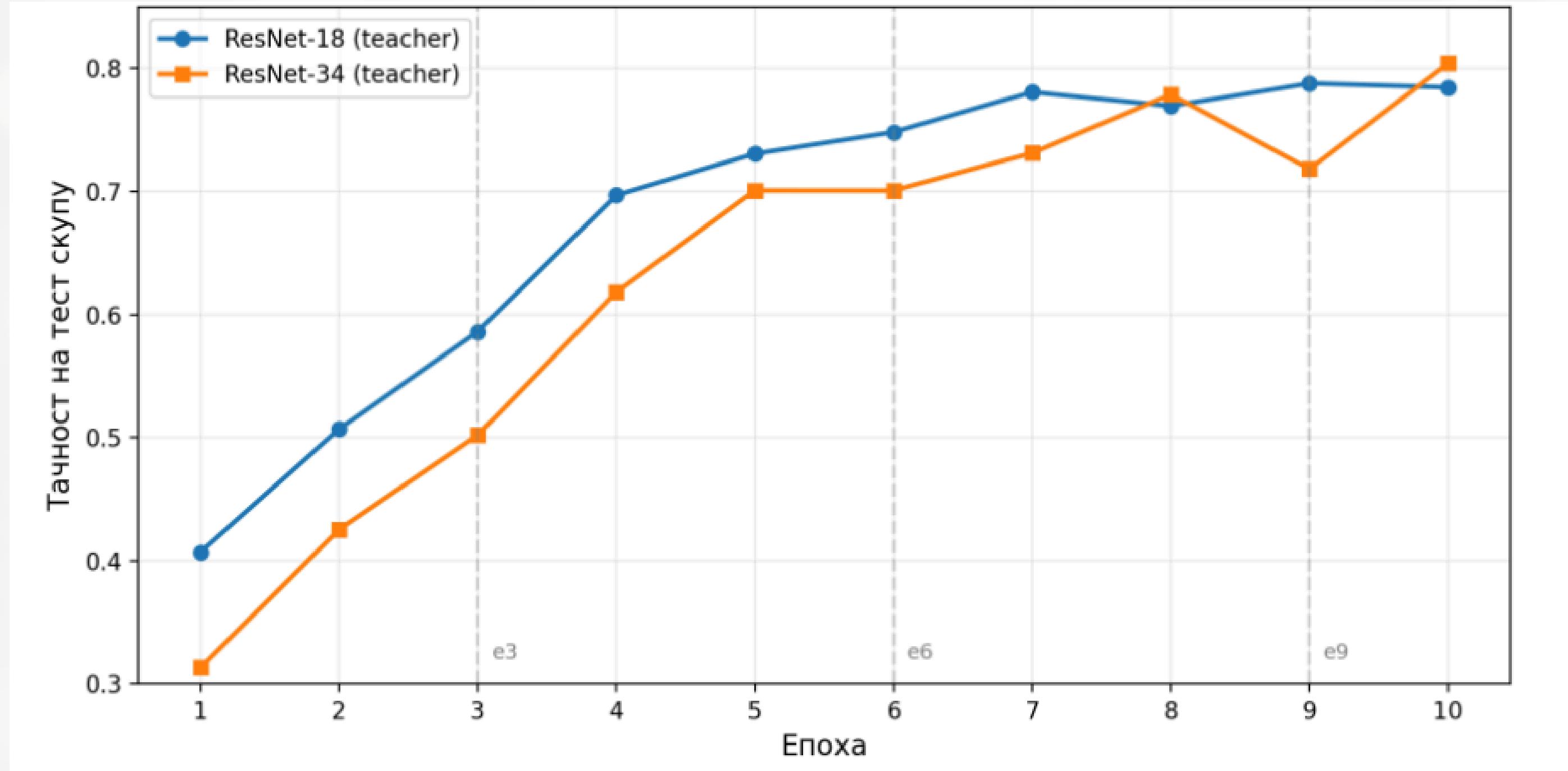
TRENING UČITELJA

Učitelji su trenirani na **CIFAR-10** datasetu, koji sadrži 10 klasa i 60.000 slika male rezolucije. Za treniranje su izabrani modeli ResNet18 i ResNet34.

Metodologija treninga:

- Optimizator: SGD sa momenatom (0.9) i regularizacijom težina (5e-4)
- Scheduler: MultiStepLR koji snižava stopu učenja u unapred definisanim epohama
- Snimanje snapshot-ova: e3, e6, e9, e10
- Poređenje konvergencije između različitih dubina modela

РЕЗУЛТАТИ УČИТЕЉА



Učitelji ResNet18 i ResNet34 pokazali su različite brzine konvergencije. ResNet34 je postigao nešto veću finalnu tačnost (~0.8041), dok je ResNet18 dostigao ~0.7848.

Ključna zapažanja su:

- Snapshot-ovi značajno variraju u tačnostima – neki, iako rani, mogu biti bolji za KD.
- Nije presudno da učitelj ima najveću moguću tačnost, već da njegove raspodele budu informativne.
- Vreme snimanja snapshot-a direktno utiče na kvalitet distilacije.

TRENING STUDENTA

Student se trenira u dve faze:

1. KD faza (prvih kd_epochs epoha):

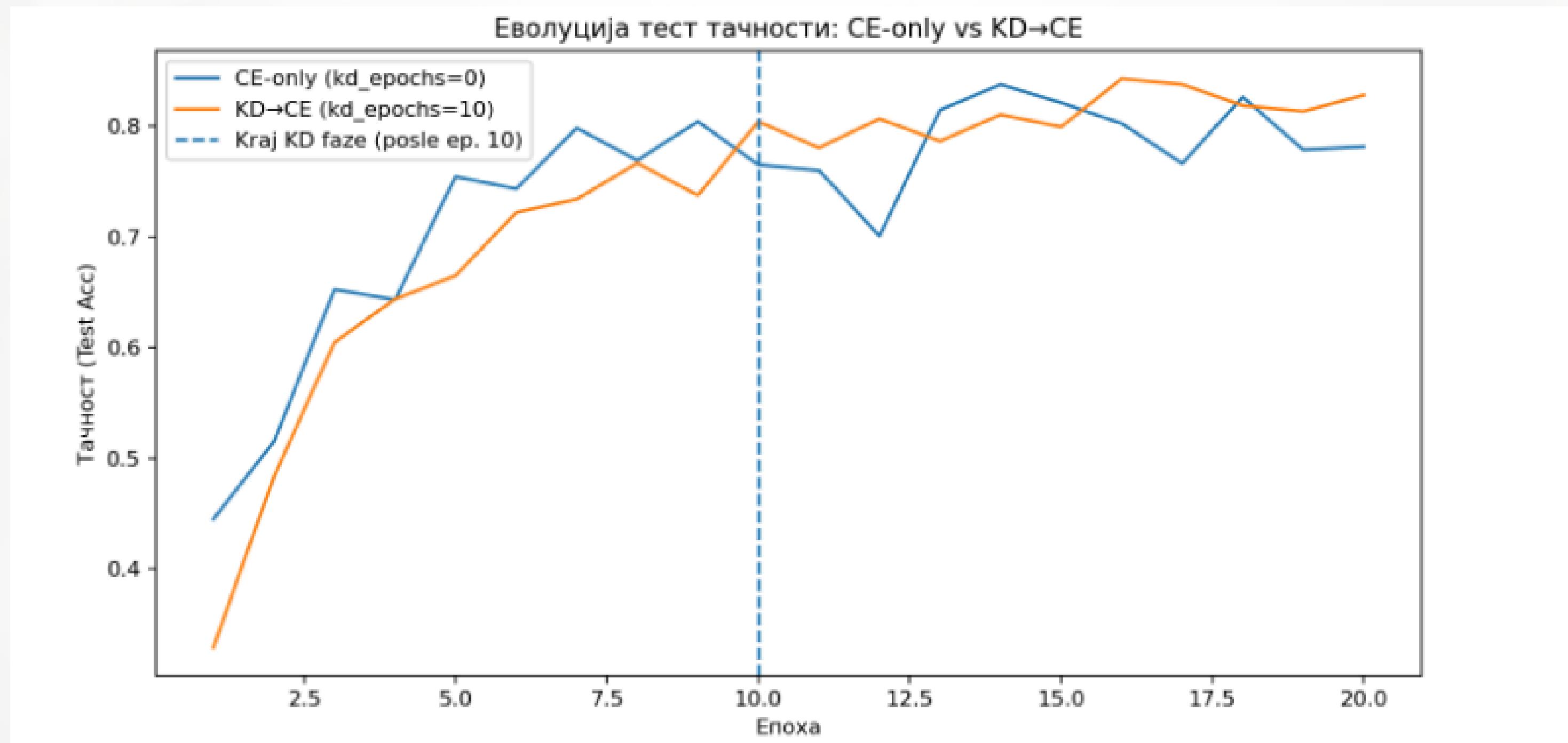
Student uči kombinacijom KD i CE gubitka. Ovde dobija najviše informacija od učitelja, učeći njegove međuklasne odnose.

2. CE-only faza:

Nakon KD faze, učenik nastavlja da trenira samostalno, koristeći samo istinske oznake. Ova faza koristi znanje stečeno u KD periodu.

POČETNI EKSPERIMENT (CE-ONLY VS KD→CE)

Poređenje pokazuje da KD→CE režim ima stabilniju krivu test tačnosti i dostiže veću konačnu vrednost tačnosti u odnosu na CE-only pristup.



- **CE-only** brzo raste, ali je neravnomerno i podložno oscilacijama.
- **KD→CE** ima sporiji početak, ali tokom CE faze postiže bolju stabilizaciju.

Razlika u konačnim vrednostima je u korist **KD→CE režima**.
Ovo potvrđuje da KD deluje kao efikasna regularizacija.

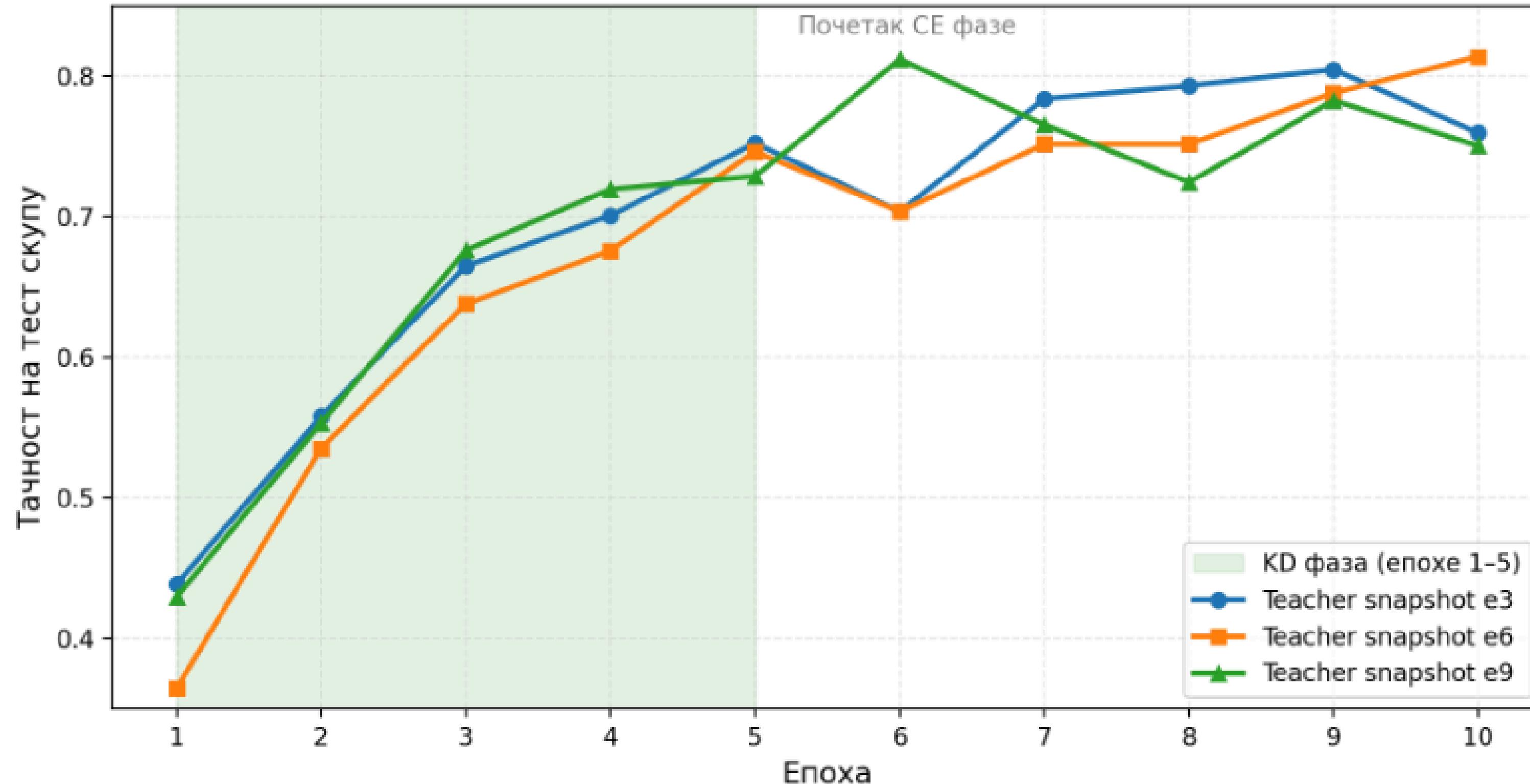
SNAPSHOT ANALIZA: UČITELJ RESNET18

Tri snapshot-a (e3, e6, e9) korišćena su za destilaciju.

Rezultati:

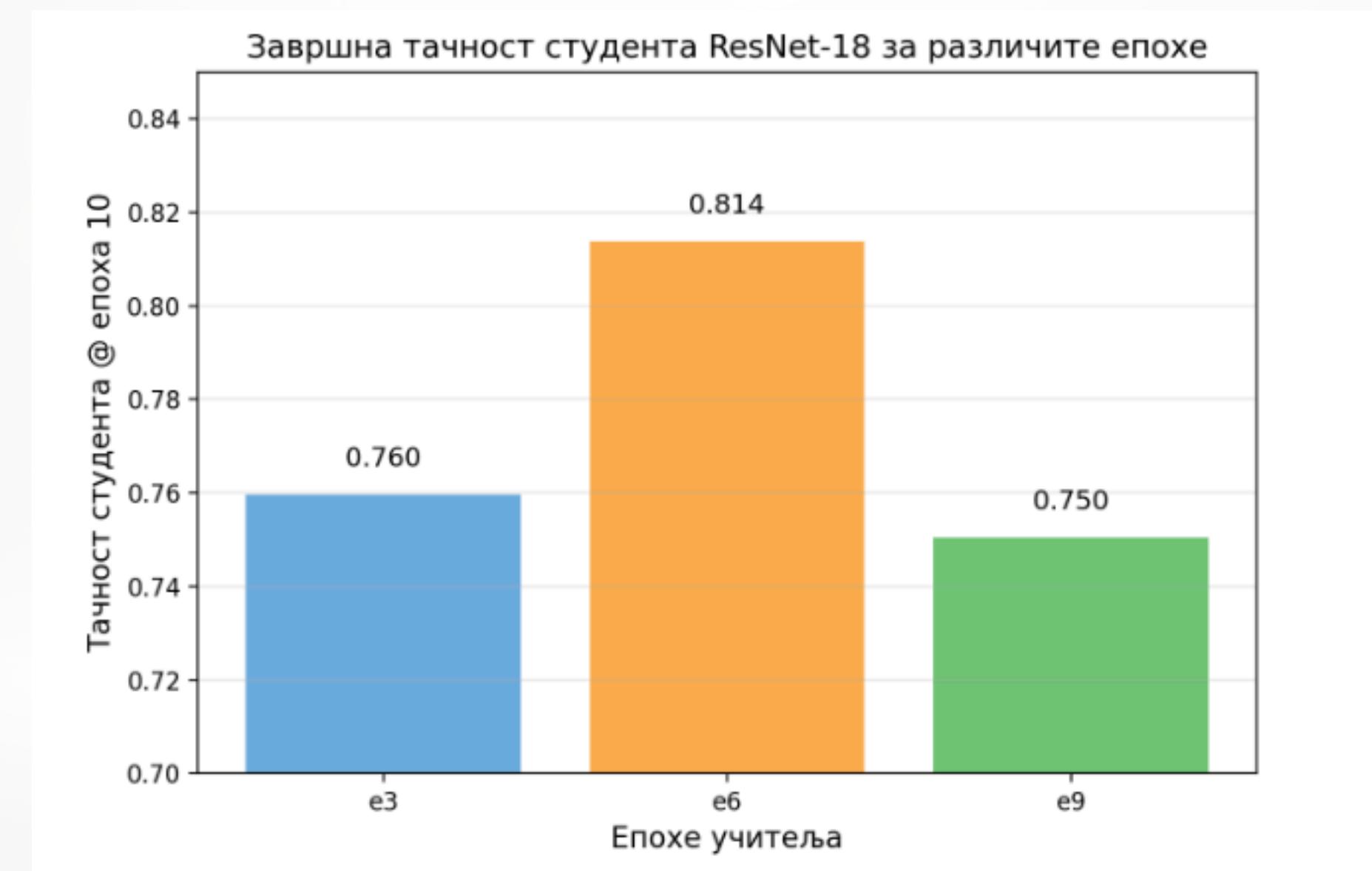
- e6 → najbolji student (0.8137)
- Umereno obučen učitelj; najbolja kombinacija informacije i generalizacije.
- e3 → nedovoljno obučen
- Daje slabije rezultate jer student dobija preosnovne, nedovoljno jasne signale.
- e9 → previše obučen učitelj
- Raspodele su oštре i student gubi mogućnost da „vidi“ srodnost klasa.

ResNet-18 студент: KD→CE тренирање за различите снапшотове учитела



Ovi rezultati potvrđuju hipotezu ESKD-a:

- Postoji optimalan trenutak obučavanja učitelja kada je njegovo znanje najkorisnije studentu.
- Ni rani snapshot, ni kasni snapshot nisu idealni.
- Umereno obučeni učitelj emituje „pravo količinu“ informacije.



SNAPSHOT ANALIZA: UČITELJ RESNET34

Kada student uči od dubljeg učitelja, pojavljuje se sličan obrazac:

Rezultati:

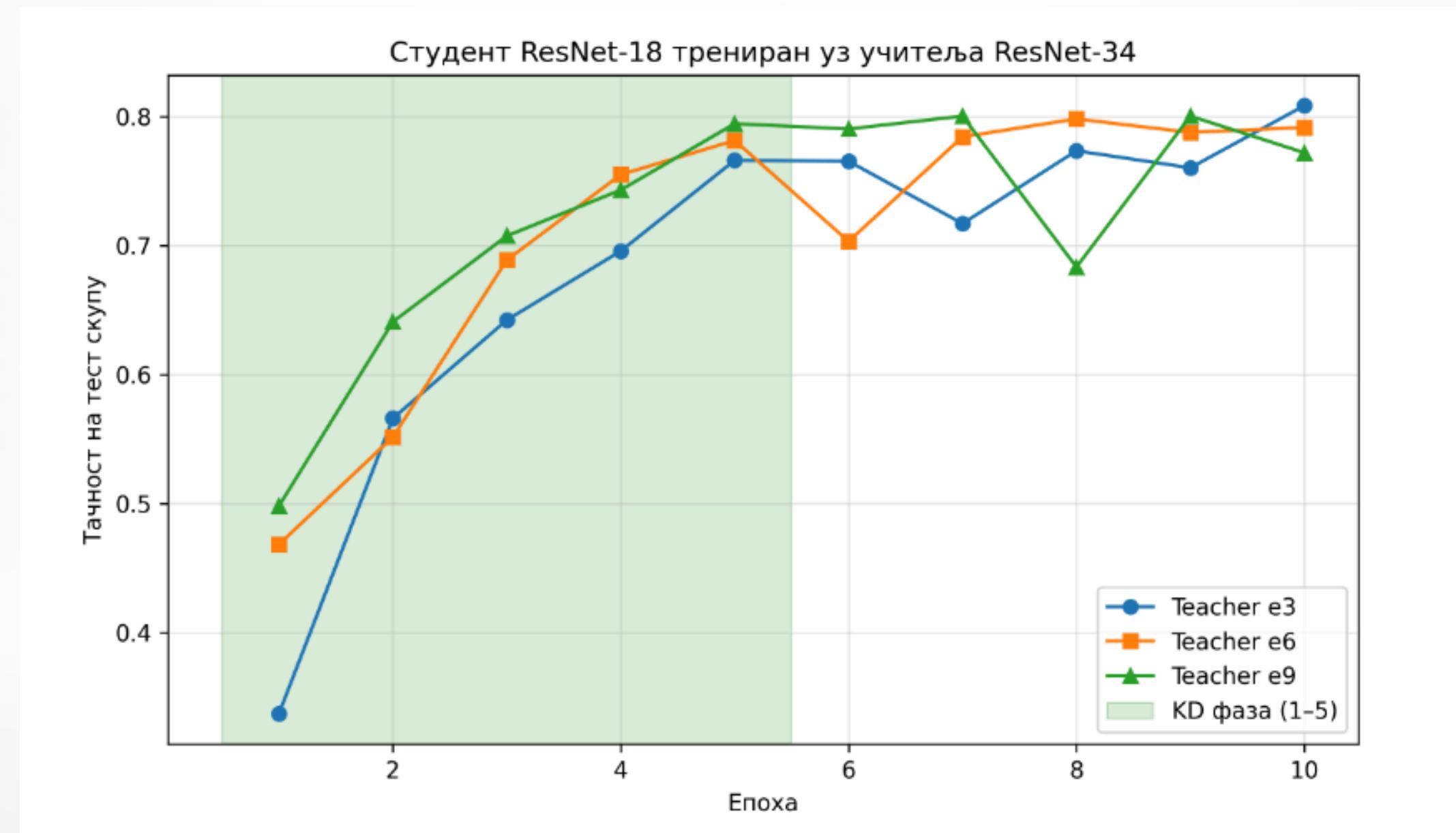
- e3 → najbolji rezultat: 0.8085
- e6 → stabilniji rezultat, ali niži
- e9 → najlošiji, zbog oštih raspodela

Iako ResNet34 ima bolju finalnu tačnost kao učitelj, to ne znači da su njegovi kasni snapshot-ovi dobri za destilaciju.

ПОРЕДЕНJE УЧИTELJA RAZLIČITIH DUBINA

Ključni zaključak: Dublji učitelj nije automatski bolji za destilaciju.

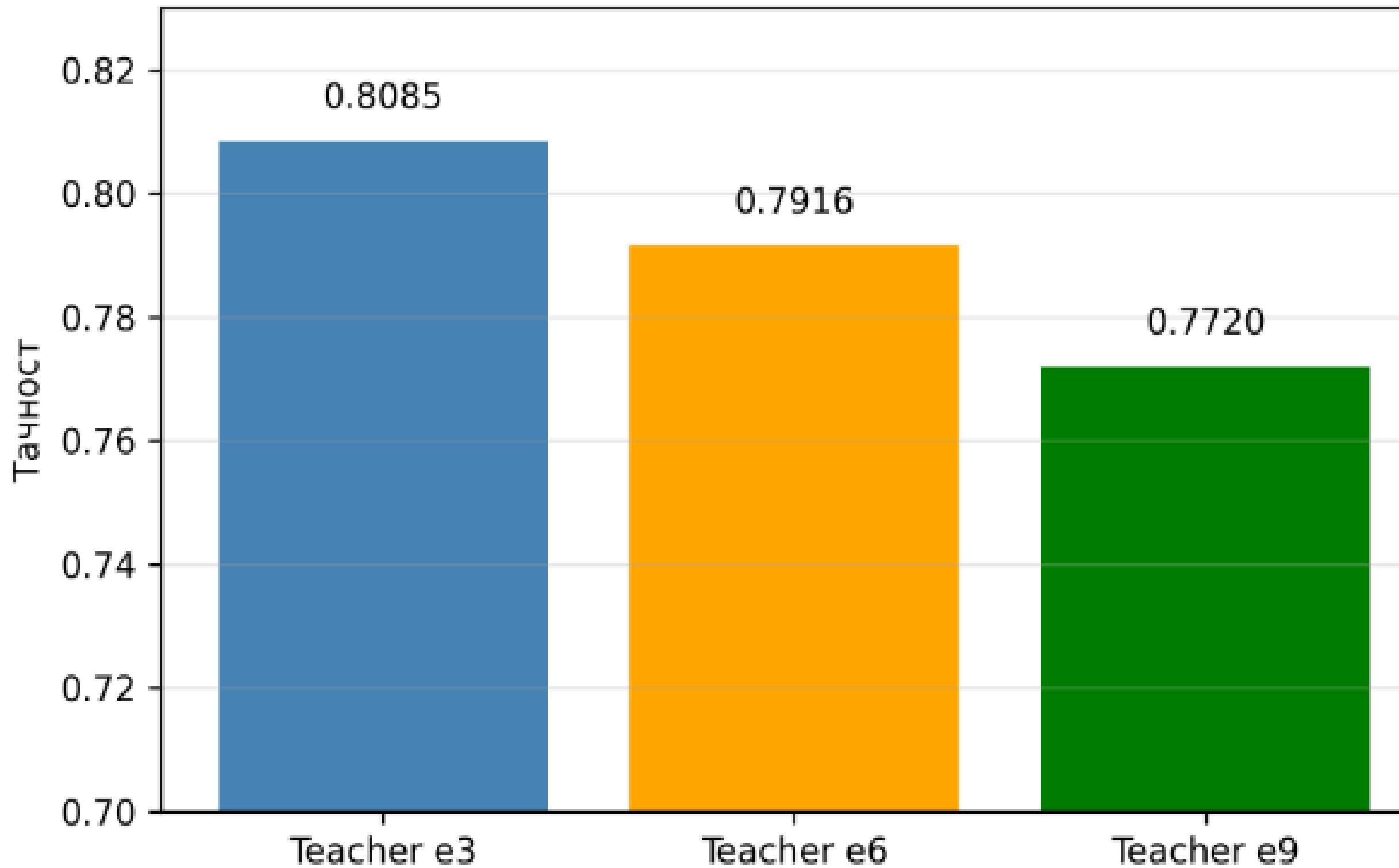
Kao i kod ResNet18, snapshot e3 ResNet34 učitelja bio je najinformativniji za studenta.
Ovo pokazuje da je stepen obučenosti važniji od arhitektonske složenosti učitelja.



GLAVNI NALAZI ISTRAŽIVANJA

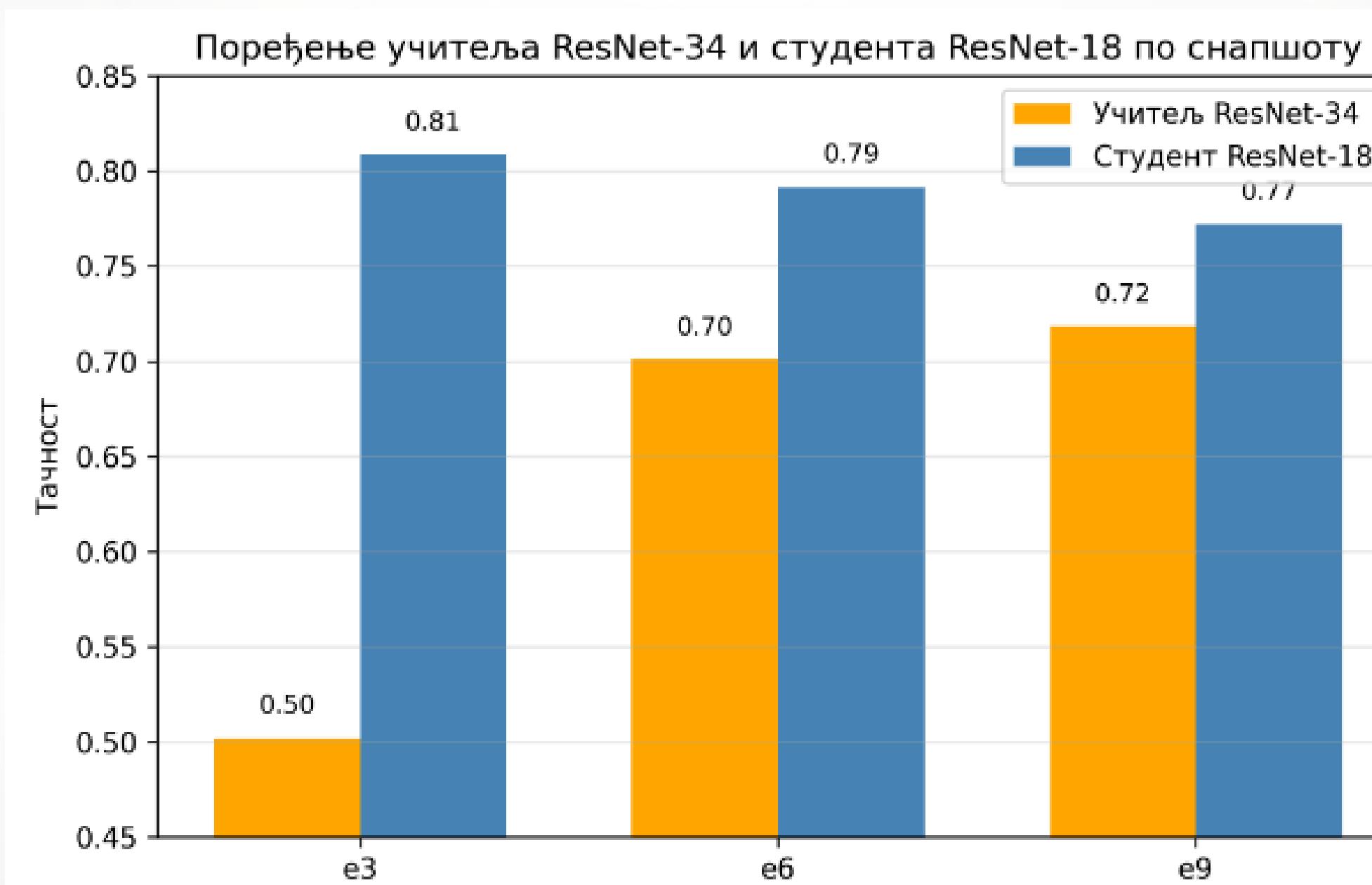
- Knowledge Distillation dosledno poboljšava rezultate studenta u odnosu na klasični CE-only trening.
- Early-Stopped učitelji su najbolji izvor informacije – ni prerano, ni prekasno obučeni.
- Dublji učitelji ne garantuju bolji prenos; informativnost raspodele je ključna.
- KD stabilizuje krivu treninga i smanjuje oscilacije.
- Rezultati su u skladu sa vodećim naučnim radovima (Hinton, Cho & Hariharan, Gou et al.).

Точность студента ResNet-18 на эпохи 10 (учитель ResNet-34)



PRAKTIČNE IMPLIKACIJE

- Moguće je napraviti mali, efikasan model koji zadržava glavne performanse velikog učitelja.
- Ovo je izuzetno važno za sisteme ograničenih resursa: mobilni telefoni, roboti, edge-AI uređaji.
- Dobijeni modeli imaju manju latenciju i manju memorijsku potrošnju.



OGRANIČENJA RADA

- Ispitivan je samo CIFAR-10 dataset.
- Korišćene su samo ResNet arhitekture.
- Razmatran je samo response-based KD, dok složeniji pristupi mogu dati drugačije rezultate.
- Trening izvođen na CPU-u → ograničen broj iteracija.

ZAKLJUČAK

Knowledge Distillation je pokazan kao izuzetno efikasan metod za prenos znanja sa složenih modela na lakše.

Rezultati ovog rada ukazuju na važnost izbora odgovarajućeg snapshot-a učitelja, kao i na značaj umerene obučenosti u postizanju najbolje generalizacije studenta.

Projekat potvrđuje relevantnost ESKD pristupa i otvara prostor za dalje naprednije tehnike destilacije.
