Şcoala de Studii Avansate a Academiei Române

Institutul de Matematică "Simion Stoilow" al Academiei Române

REZUMATUL TEZEI DE DOCTORAT

Învățarea în video folosind paradigma student-profesor:
De la detectarea nesupervizată a obiectelor în video până la regăsirea videourilor

Coordonator științific:
Prof. Dr. Marius Leordeanu

Doctorand:
Ioana Croitoru

BUCUREȘTI
2022

Cuprins

Sinopsis

Vederea artificială (Computer Vision) și procesarea limbajului natural sunt două domenii de mare interes în inteligența artificială. Accentul acestei lucrări este distilarea generalizată, fiind studiată și aplicată pe probleme legate de vedere artificială și procesarea limbajului natural. Problemele pe care le considerăm sunt segmentarea obiectelor, detectarea obiectelor și regăsirea videourilor. Demonstrăm că distilarea poate fi utilizată pentru a rezolva diferite probleme, cum ar fi învățarea supervizată sau nesupervizată. Pentru învățarea nesupervizată, abordăm problema de detectare și segmentare a obiectele din prim-plan în imagini individuale. Pentru a atinge scopul ne folosim de două căi diferite: un *student*, constând dintr-o rețea neurală adâncă care învață să prezică, dintr-o singură imagine de intrare, iesirea celei de-a doua căi, *profesor*, care efectuează detectarea nesupervizată a obiectelor în video. Mai mult, demonstrăm că performanța poate fi îmbunătățită pe parcursul mai multor generații de *studenți* și *profesori*. Pentru învățarea supervizată considerăm problema de regasire a videourilor. Pentru aceasta, suntem primii care investighează metode ce învață din mai multe atribute de text (text embbedings) pre-antrenate și propunem o nouă metodă de distilare generalizată, numită TEACHTEXT, care folosește indicii complementare de la mai multe codificatoare de text pentru a oferi un semnal de supervizare adițional pentru modelul ce efectuază regăsirea videourilor. Mai mult, extindem metoda la modalități video și arătăm că putem reduce efectiv numărul de modalități utilizate în timpul testării fără a compromite performanța. Abordarea noastră îmbunătățește rezultatele cu o marjă semnificativă pentru problema de regăsire pe mai multe baze de date și nu adaugă nicio necesitate computațională adițională în momentul testării.

Capitolul 1

Introducere

Vederea artificială este o ramură a învățării automate (IA) care procesează imagini, videoclipuri și alte canale vizuale pentru a obține informații relevante. Informația poate însemna orice, de la detectarea marginilor până la detectarea obiectelor, segmentarea obiectelor, urmărirea unui obiect și așa mai departe. De obicei, metodele dezvoltate în vederea artificială urmăresc să reproducă capacitatea vederii umane. Există o mulțime de domenii de cercetare și o mulțime de aplicații, precum: imagistica medicală [48], captarea mișcării, siguranța auto [17] etc.

Procesarea limbajului natural (NLP) este, pe de altă parte, un domeniu al inteligenței artificiale care se concentrează pe obținerea de informații relevante din texte. În timp ce scopul vederii artificiale este de a permite computerelor să înțeleagă vizual materialul, similar cu capacitatea viziunii umane, scopul NLP este de a permite computerelor să înțeleagă cuvintele rostite în același mod în care ființele umane pot. Unele dintre domeniile de cercetare în NLP sunt: traducerea automată [55], rezumarea textului [31] și așa mai departe.

Bineînțeles că există mult interes și pentru sarcinile în care vederea artificială se intersectează cu procesarea limbajului natural, cum ar fi: regăsire text-video [10], subtitrări pentru videouri [3], subtitrări pentru imagini [57] și așa mai departe.

În zilele noastre, învățarea cu rețele neurale adânci [16] se află la baza majorității metodelor de viziune computerizată de ultimă generație, precum și a metodelor de procesare a limbajului natural. Începând cu 2014, învățarea cu rețele neurale adânci a sporit performanța diferitelor probleme în ambele domenii. Pentru ca o rețea neurală adâncă să aibă rezultate

competitive, sunt necesare cantități uriașe de date. Pentru a obține rezultate bune, datele trebuie adnotate. Dar datele etichetate sunt greu de colectat, deoarece implică un număr mare de persoane care să adnoteze imaginea, textul sau videoclipul. Deci, există, de asemenea, un interes uriaș pentru zona nesupravegheată, semi-supravegheată și auto-supravegheată atât a vederii artificiale, cât și a procesării limbajului natural. Există cantități uriașe de videoclipuri, texte și imagini pe web, ceea ce a făcut această zonă populară.

Mai mult, pentru a crește performanța unei rețele neurale adânci fără a necesita date suplimentare adnotate, apare ideea de a învăța din alte metode. Distilarea generalizată [32] este un cadru care unifică distilarea cunoștințelor [19] și informația privilegiată [54], două metode care permit mașinilor să învețe de la alte mașini. Distilarea generalizată implică că o mașină, care este de obicei numită student, este capabilă să învețe de la o altă mașină instruită, numită de obicei profesor. Lucrările recent publicate în domeniu de distilare generalizată [9, 10] au dovedit că performanța unui student utilizând rețele neurale adânci poate fi îmbunătățită prin utilizarea informațiilor furnizate de profesor. Desigur, combinația a două sau mai multe metode diferite care sunt antrenate pentru a rezolva aceeași sarcină și sunt combinate împreună într-un ansamblu s-a dovedit a obține scoruri mai mari decât un singur model [2].

1.1 Contribuții

În această lucrare ne propunem să studiem comportamentul învățării prin distilare generalizată de la un ansamblu de profesori, atât în scenariul nesupravegheat, cât și în scenariul supravegheat. Sarcinile pe care le-am folosit pentru distilarea generalizată în scenariul nesupravegheat este segmentarea și detectarea obiectelor, în timp ce sarcina în cazul supravegheat este regăsirea text-video. În acest fel, demonstrăm eficiența învățării de la un ansamblu de profesori în mai multe scenarii: în cazul nesupravegheat și în cazul supravegheat precum și într-o aplicație de viziune computerizată și într-o aplicație care se află la intersecția vederii computerizate și a procesarii limbajului natural, și anume recuperarea text-video.

Segmentarea obiectelor. Detectarea obiectelor este sarcina în care este nevoie de a localiza și clasifica fiecare obiect într-o imagine sau videoclip. Segmentarea obiectelor este sarcina în care fiecare pixel dintr-o imagine ar trebui să fie etichetat cu clasa corespunzătoare.

Pentru segmentarea obiectelor ne-am concentrat pe învățarea nesupravegheată din datele vizuale, deoarece are o valoare practică imensă, fiindcă cantități uriașe de videoclipuri neetichetate pot fi colectate la costuri reduse. Ne atingem obiectivul prin antrenarea unui căi student, constând dintr-o rețea neuronală adâncă care învață să prezică, dintr-o singură imagine de intrare, rezultatul unei căi profesor care realizează descoperirea nesupravegheată a obiectelor în video. Acest lucru are un dublu beneficiu: în primul rând, permite, în principiu, posibilități nelimitate de generalizare în timpul antrenamentului, rămânând în același timp rapid la testare. În al doilea rând, studentul nu numai că devine capabil să detecteze în imagini mult mai bine decât profesorul său nesupravegheat, dar obține și rezultate de ultimă generație pe două seturi de date actuale, YouTube Objects și Object Discovery. La momentul testării, sistemul nostru este cu două ordine de mărime mai rapid decât alte metode anterioare.

În plus, am îmbunătățit sistemul propus prin proiectarea acestuia pentru a învăța de-a lungul mai multor generații de profesori și studenți. La fiecare generație, profesorul efectuează descoperiri nesupravegheate de obiecte în videoclipuri sau colecții de imagini, iar un modul de selecție automată preia cadre segmentate corespunzător și le transmite căii student. La fiecare generație sunt instruiți mai mulți studenți, cu arhitecturi de rețele adânci diferite pentru a asigura o diversitate mai bună. Studenții de la o iterație ajută la formarea unui modul de selecție mai bun, formând împreună o cale profesor mai puternică la următoarea iterație.

Regăsirea text-video. Pentru a demonstra eficacitatea distilării generalizate, am luat în considerare și un scenariu supravegheat pentru o sarcină foarte populară care se află la intersecția vederii artificiale și procesării limbajului natural, anume regăsirea videourilor. Având la dispoziție o propoziție în limbaj natural și o colecție de videoclipuri, scopul este de a proiecta un sistem care este capabil să regasească videoclipul care este cel mai bine descris de propoziție. În ultimii ani, s-au realizat progrese considerabile în ceea ce privește regăsirea text-video, prin valorificarea pregătirii preliminare la scară largă pe seturi de date vizuale și audio pentru a construi codificatoare video puternice. Prin contrast, în ciuda simetriei naturale, proiectarea algoritmilor eficienți pentru exploatarea la scară largă a limbajului rămâne subexplorată. În această lucrare, suntem primii care investighează proiectarea unor astfel de algoritmi și propune o nouă metodă de distilare generalizată, TEACHTEXT, care folosește indicii complementare de la mai multe codificatoare de text pentru a oferi un semnal de supraveghere îmbunătățit modelului de regăsire.

Capitolul 2

Învățare nesupravegheată din videoclipuri pentru a detecta obiectele din prim-plan în imagini individuale

Învățarea nesupravegheată din datele vizuale este una dintre cele cele mai dificile provocări în vederea artificială. Este esențială pentru a înțelege cum funcționează recunoașterea vizuală. Învăţarea nesupravegheată are o valoare practică imensă, deoarece cantități uriașe de videoclipuri neetichetate pot fi colectate cu un cost scăzut. În acest capitol abordăm sarcina învățării nesupravegheate de a detecta si segmenta obiectele din prim-plan în imagini individuale. Ne atingem scopul prin formarea unei căi student, constând dintr-o rețea neurală adâncă care învață să prezică, de la o singură imagine de intrare, rezultatul unei cai profesor care descoperă nesupravegheat obiectele în video. Abordarea noastră este diferită de metodele publicate care realizează descoperiri nesupravegheate în videoclipuri sau în colecții de imagini. Mutăm faza de descoperire nesupravegheată în timpul etapei de antrenare, în timp ce la momentul testării aplicăm procesarea standard. Aceasta are un beneficiu dublu: în primul rând, permite, posibilități nelimitate de generalizare în timpul antrenamentului, rămânând în același timp rapidă la testare. În al doilea rând, studentul nu numai că devine capabil să detecteaze în imagini individuale mult mai bine decât profesorul său, dar atinge și performante înalte pe doua seturi de date actuale, YouTube Objects și Object Discovery. La momentul testării, sistemul nostru este cu două ordine de mărime mai rapid decât alte metode anterioare.

Acest capitol se bazează pe lucrarea "Unsupervised learn-ing from video to detect foreground objects in single images." Croitoru, Ioana, Simion-Vlad Bogolin, and Marius Leordeanu. International Conference on Computer Vision. 2017.

2.1 Introducere

Învățarea nesupravegheată este una dintre cele mai dificile și mai interesante probleme din vederea artificială și învățarea automată de astăzi. Cercetătorii cred că învățarea nesupravegheată din video ar putea ajuta la decodarea întrebărilor dificile cu privire la natura inteligenței și a învățării. Deoarece videoclipurile neetichetate sunt ușor de colectat la costuri reduse, rezolvarea acestei sarcini ar aduce o mare valoare practică în vederea artificială și robotică.

Sistemul nostru este prezentat în figura 1. Avem o etapă de pregătire nesupravegheată, în care o rețea neurală adâncă, calea student, învață cadru cu cadru de la o cale profesor nesupravegheată, care efectuează segmentarea obiectelor în videoclipuri. Calea profesor profită de consistența în aspect, formă și mișcare manifestată de obiectele din video. În acest fel, descoperă obiecte din videoclip și produce o segmentare a prim-planului pentru fiecare cadru individual. Apoi, calea student încearcă să imite pentru fiecare cadru segmentarea profesorului, având în același timp ca intrare doar o singură imagine - cadrul curent. Calea profesor este mult mai simplă ca structură, dar are acces la informații în timp. În schimb, studentul are o structură mult mai adâncă, dar are acces doar la o singură imagine. Astfel, informația descoperită de profesor în timp este surprinsă de student în profunzime, peste straturi neurale de abstractizare. În experimente, arătăm un fapt foarte încurajator: elevul învață cu ușurință să-și depășească profesorul și descoperă de la sine cunoștințe generale despre proprietățile de formă și aspect ale obiectelor, cu mult peste abilitățile profesorului. Deoarece există metode disponibile pentru descoperirea video cu performanță bună, sarcina de a antrena un astfel de student devine imediat fezabilă. În această lucrare am ales algoritmul VideoPCA introdus ca parte a sistemului în [51] deoarece este foarte rapid (50-100 fps). VideoPCA folosește caracteristici foarte simple (culorile pixelilor) și este nesupravegheat - fără a utiliza caracteristici pre-antrenate.

2.2 Context științific

Metodele recente nesupravegheate urmează două direcții. Prima fiind învățarea unor funcții puternice într-un mod nesupravegheat și apoi folosirea acestora într-o schemă clasică de învățare supravegheată în combinație cu diferiți clasificatori, cum ar fi SVMs sau CNNs [40]. În metoda foarte recentă [38], dezvoltată independent de a noastră, o rețea adâncă învață, de la un sistem nesupravegheat folosind indicii de mișcare în video, caracteristici de imagine care sunt aplicate mai multor sarcini de învățare prin transfer. A doua abordare a învățării nesupravegheate este de a descoperi, la momentul testării, modele comune în datele neetichetate folosind formulări de grupare sau de extragere a datelor [20]. Învățarea nesupravegheată în videoclipuri este, de asemenea, legată de co-segmentare [21] și de localizarea slab supravegheată [11]. Metodele anterioare se bazează pe potrivirea caracteristicilor locale și pe detectarea tiparelor lor de co-apariție [51], în timp ce metodele mai recente [23] descoperă tuburi de obiecte legând obiectele candidate între cadre cu sau fără rafinarea locației lor. În mod tradițional, sarcina de învățare nesupravegheată din secvențele de imagini a fost formulată ca o problemă de potrivire a caracteristicilor sau de optimizare a grupării datelor, care este foarte costisitoare din punct de vedere computațional.

2.3 Arhitectura sistemului

În experimentele noastre, elevul își depășește într-adevăr profesorul. Mai mult, obține rezultate de ultimă generație pe două seturi de date de referință. Succesul acestei paradigme de învățare nesupravegheată se datorează faptului că elevul este obligat să surprindă din aparență doar trăsături vizuale care sunt buni predictori ai prezenței obiectelor. Prezentare generală a sistemului nostru este prezentată în Figura 1.

Calea profesor: descoperirea nesupravegheată în video. Am folosit algoritmul VideoPCA, care este o parte a întregului sistem introdus în [51]. Acesta rulează la $50-100$ fps, astfel că putem produce un milion de segmentări nesupravegheate într-un timp rezonabil.

Calea student: segmentare cu o singură imagine. Constă într-o rețea convoluțională adâncă, cu zece straturi (șapte convoluționale, două de grupare și un strat complet conectat). Tratăm segmentarea obiectelor din prim-plan ca pe o problemă de regresie, în care masca

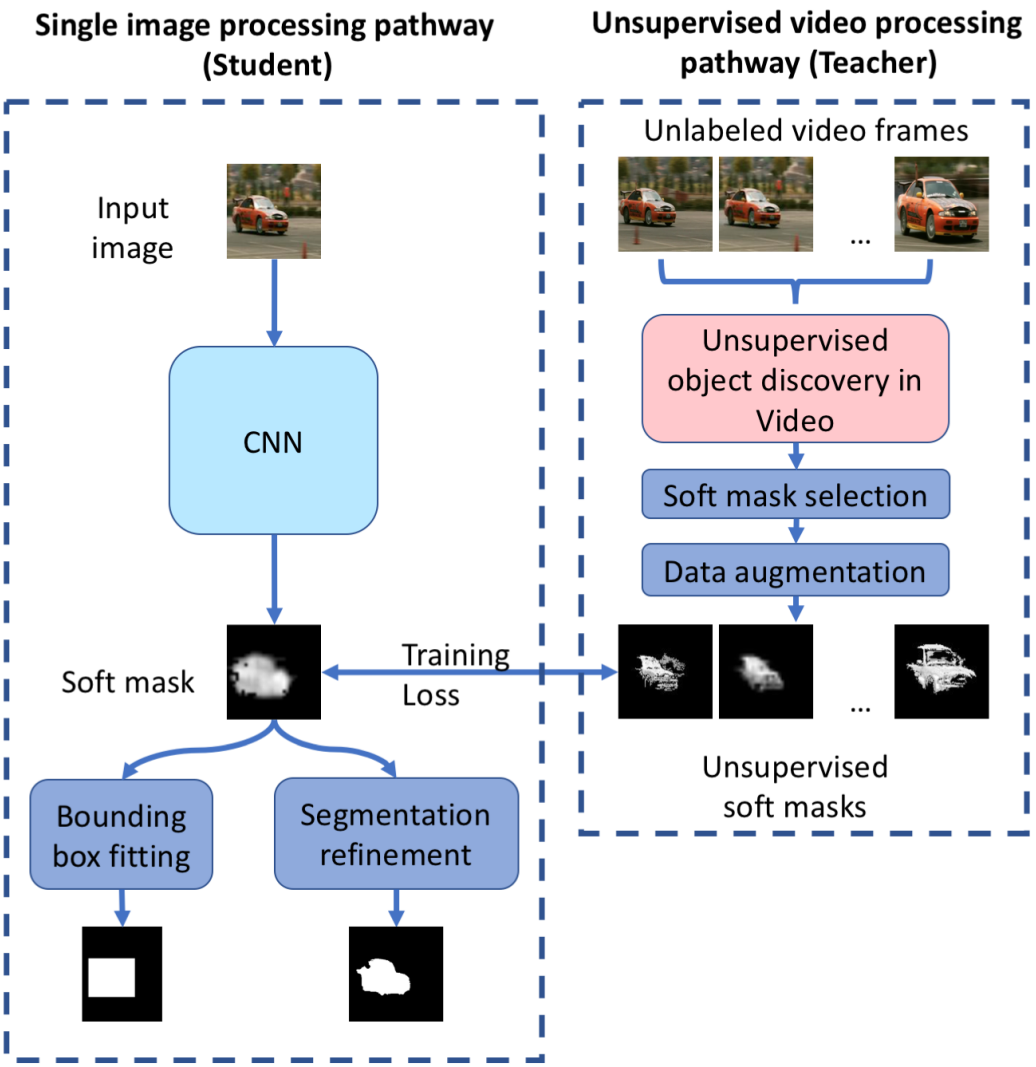

Figure 1: **Sistemul dual student-profesor propus pentru învățarea nesupravegheată pentru a detecta obiectele din prim-plan în imagini.** Are două căi: profesorul, în dreapta, descoperă într-un mod nesupravegheat obiectele din prim-plan în video. Acesta produce măști de segmentare pentru fiecare cadru. Măștile rezultate sunt apoi filtrate pe baza unei valori simple și eficiente de calitate nesupravegheată. Setul de segmentări selectate este apoi mărit într-un mod relativ simplu, automat. Setul final de perechi rezultat - imaginea de intrare (un cadru video) și masca pentru acel cadru particular care acționează ca o etichetă nesupravegheată - sunt utilizate pentru antrenamentul căii student.

	Airplane		Car		Horse	
	P	J	P	J	P	J
[26]	80.20	7.90	68.85	0.04	75.12	6.43
[21]	49.25	15.36	58.70	37.15	63.84	30.16
[22]	47.48	11.72	59.20	35.15	64.22	29.53
[46]	88.04	55.81	85.38	64.42	82.81	51.65
[8]	90.25	40.33	**87.65**	64.86	86.16	33.39
Ours_1	90.92	**62.76**	85.15	66.39	**87.11**	54.59
Ours_2	**91.41**	61.37	86.59	**70.52**	87.07	**55.09**

Table 1: **Rezultate pe baza de date Object Discovery in Internet Images [46] (metrici P, J).** Ours_1 reprezintă rețeaua noastră antrenată folosind setul de date VID (cu selecție de 10%), în timp ce Ours_2 reprezintă rețeaua noastră antrenată pe seturi de date VID și YTO (cu o selecție de 10%). Observăm că Ours_2 are rezultate mai bune (având media P **88,36** și media J **62,33** comparativ cu Ours_1 (media P: 87,73, media J: 61,25).

produsă de sistemul de segmentare video nesupravegheată este considerată ieșirea dorită.

Selecția nesupravegheată a măștilor. Am folosit o măsură simplă a calității măștilor bazată pe următoarea observație: când măștile sunt aproape de eticheta reală, media valorilor lor diferite de zero este de obicei mare. Astfel, atunci când profesorul este încrezător, este mai probabil să aibă dreptate. Valoarea medie a pixelilor diferiti de zero din mască este apoi folosită ca indicator de scor pentru fiecare cadru segmentat.

2.4 Analiză experimentală

Comparații cu alte metode. Comparăm sistemul nostru nesupravegheat cu metodele de ultimă generație concepute pentru sarcina de descoperire a obiectelor în colecții de imagini, care ar putea conține una sau câteva categorii principale de obiecte de interes. Un punct de referință actual reprezentativ în acest sens este setul de date Object Discovery in Internet Images. Spre deosebire de alte metode din literatura, nu avem nevoie de o colecție de imagini în timpul testării, deoarece fiecare imagine este procesată independent de sistemul nostru, la momentul testării. Prin urmare, performanța noastră nu este afectată de structura colecției de imagini sau de numărul de clase de interes prezente în colecție.

Am testat sistemul nostru pentru sarcina de segmentare fină a obiectelor din prim-plan și am comparat cu cele mai bune metode din literatura de specialitate pe setul de date Object

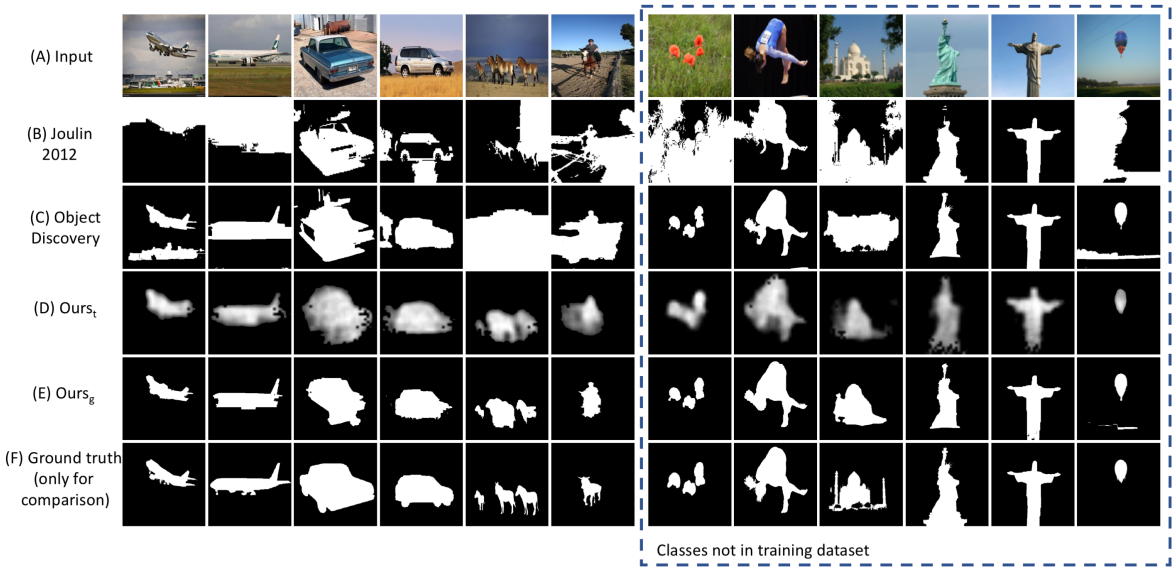

Figure 2: **Rezultate vizuale pe setul de date Object Discovery.** A: imagine de intrare, B: segmentare obținută de [22], C: segmentare obținută de [46], D: masca produsă de rețeaua noastră , E: masca de segmentare produsă după rafinarea ieșirii rețelei noastre cu GrabCut [45], F: segmentarea adevărată. Pentru mai multe detalii și rezultate consultați: https://sites.google.com/view/unsupervisedlearningfromvideo.

Discovery în Tabelul 1. Evaluăm pe baza acelorași metrici P, J așa cum este descris de Rubinstein *et al.* [46] - cu cât P și J sunt mai mari, cu atât mai bine. P se referă la precizia per pixel, în timp ce J este indexul Jaccard (intersecția peste uniunea dintre rezultat și segmentările etichetei reale). În Figura 2 prezentăm câteva exemple calitative.

2.5 Concluzii

Am arătat în experimente extinse că este posibil să se utilizeze o metodă relativ simplă pentru descoperirea nesupravegheată a obiectelor în video pentru a antrena o rețea neurală adâncă pentru detectarea și segmentarea obiectelor în imagini individuale. Rezultatul este interesant și încurajator și arată cum un sistem ar putea învăța, nesupravegheat, caracteristici vizuale generale care prezic bine prezența și forma obiectelor în imagini. Rețeaua descoperă, în esență, trăsăturile obiectelor din imagini individuale, la diferite niveluri de abstractizare, care sunt puternic corelate cu consistența spațiu-temporală a obiectelor din video.

Rețeaua student, în timpul fazei de pregătire nesupravegheată, este astfel capabilă să învețe caracteristici generale de obiecte care depășesc cu mult capacitățile profesorului său. Aceste

caracteristici includ forma bună, închiderea, contururile netede, precum și contrastul cu fundalul. Ceea ce profesorul descoperă în timp, studentul fiind mai profund și complex este capabil să învețe caracteristici ale imaginii la diferite niveluri de abstractizare. Prin urmare, modelul nostru de învățare nesupravegheată, testat în experimente extinse, aduce o contribuție valoroasă la problema învățării nesupravegheate în vederea artificială.

Capitolul 3

Învățare nesupravegheată a segmentării obiectelor din prim-plan

Învățarea nesupravegheată reprezintă una dintre cele mai interesante provocări în vederea artificială astăzi. Sarcina are o valoare practică imensă, cu multe aplicații în inteligența artificială și tehnologiile emergente, deoarece cantități mari de imagini și videoclipuri neetichetate pot fi colectate la costuri reduse. În această lucrare, abordăm problema învățării nesupravegheate în contextul segmentării principalelor obiecte din prim-plan în imagini individuale. Propunem un sistem de învățare nesupravegheat, care are două căi, profesorul și respectiv studentul. Sistemul este conceput pentru a învăța peste mai multe generații de profesori și studenți. La fiecare generație, profesorul efectuează descoperiri nesupravegheate de obiecte în videoclipuri sau colecții de imagini, iar un modul de selecție automată preia segmentările bune de cadre și le transmite caii student pentru antrenare. La fiecare generație sunt antrenați mai mulți studenți, cu arhitecturi de rețele adânci diferite pentru a asigura o diversitate mai bună. Studenții de la o iterație ajută la formarea unui modul de selecție mai bun, formând împreună o cale de profesor mai puternică la următoarea iterație. Metoda noastră obține rezultate de top pe trei seturi de date curente de descoperirea obiectelor în video, segmentarea nesupravegheată a imaginii.

Acest capitol se bazează pe lucrarea "Unsupervised learning of foreground object segmentation." Croitoru, Ioana, Simion-Vlad Bogolin, and Marius Leordeanu. International Journal of Computer Vision 127.9 (2019): 1279-1302.

3.1 Introducere

În această teză, propunem o abordare nouă a învățării nesupravegheate care tratează cu succes multe dintre provocările asociate cu această sarcină. Prezentăm un sistem care este compus din două căi principale, una care realizează descoperirea nesupravegheată a obiectelor în videoclipuri sau colecții mari de imagini - calea profesor, iar cealaltă - calea student, care învață de la profesor să segmenteze obiectele din prim-plan în imagini individuale. Procesul de învățare nesupravegheat ar putea continua pe parcursul mai multor generații de studenți și profesori. Aspectele cheie ale abordării noastre, care asigură îmbunătățirea performanței de la o generație la alta, sunt: 1) existența unui modul de selecție nesupravegheat care este capabil să selecteze măști de bună calitate generate de profesor și să le transmită pentru antrenare către studenții din generația următoare; 2) formarea mai multor studenți cu arhitecturi diferite, capabili, prin diversitatea lor, să ajute la formarea unui modul de selecție mai bun pentru următoarea iterație și să formeze, împreună cu selecția, o cale profesor mai puternică la următoarea iterație și 3) acces la cantități mai mari de date mai complexe, neetichetate, care devin mai utile pe măsură ce generaţiile devin mai puternice.

În figura 3 vă prezentăm o prezentare grafică a sistemului nostru complet. În etapa de formare nesupravegheată, rețeaua de studenți (modulul A) învață, cadru cu cadru, de la calea profesor nesupravegheat (modulele B și C) să producă măști de obiecte similare în imagini individuale. Modulul B descoperă obiecte în imagini sau videoclipuri, în timp ce modulul C selectează care măști produse de modulul B sunt suficient de bune pentru a fi trecute la modulul A pentru antrenament. Astfel, ramura student încearcă să imite ieșirea modulului B pentru cadrele selectate de modulul C, având ca intrare doar o singură imagine - cadrul curent, în timp ce profesorul poate avea acces la o întreagă secvență video.

3.2 Context științific

Literatura în învățarea nesupravegheată urmează două direcții. 1) Prima fiind învețarea unor funcții puternice într-un mod nesupravegheat și apoi folosirea lor pentru învățarea prin transfer, în cadrul unei scheme supravegheate și în combinație cu diferiți clasificatori, cum ar fi SVM-uri sau CNN-uri ([40]). 2) A doua direcție este de a descoperi, la momentul testării,

modele comune în datele neetichetate, folosind formulări de grupare, potrivire a caracteristicilor sau de extragere a datelor ([20]).

Aparținând primei categorii și strâns legată de munca noastră, abordarea din [38] propune un sistem în care o rețea neuronală profundă învață să producă măști de obiecte într-un modul nesupravegheat care folosește indicii de flux optic în video. Recent, cercetătorii au început să folosească structura naturală, spațială și temporală din imagini și videoclipuri ca semnale de supraveghere în abordări de învățare nesupravegheată care sunt considerate a urma paradigma de *învățare auto-supravegheată* ([43]). Metodele care se încadrează în această categorie includ cele care învață să estimeze pozițiile relative ale bucăților în imagini ([12]) și să prezică canalele de culoare ([28]).

A doua abordare a învățării nesupravegheate include metode pentru co-segmentarea imaginilor ([21]) și localizarea slab supravegheată ([11]). Metodele anterioare se bazează pe potrivirea caracteristicilor locale și pe detectarea tiparelor lor de co-ocurență ([51]), în timp ce cele mai recente ([23]) descoperă tuburi de obiecte legând obiectele candidate între cadre cu sau fără rafinarea locației lor. În mod tradițional, sarcina de învățare nesupravegheată din secvențele de imagini a fost formulată ca o problemă de potrivire a caracteristicilor sau de optimizare a grupării datelor, care este foarte costisitoare din punct de vedere computațional datorită naturii sale combinatorii. Mai sunt si altele lucrări ([29]) care abordează sarcini de învățare nesupravegheate, dar nu sunt complet nesupravegheate, folosind funcții puternice care sunt pre-antrenate în mod supravegheat pe seturi de date mari, cum ar fi ImageNet ([47]) sau VOC2012 ([14]).

În ceea ce privește scopul final, munca noastră este mai mult legată de a doua direcție de cercetare, anume descoperirea nesupravegheată în video. Spre deosebire de aceaste metode, noi nu descoperim obiecte în timpul testării, ci în timpul procesului de antrenare nesupravegheată, când calea elevului învață să detecteze obiectele din prim-plan.

3.3 Arhitectura sistemului

Propunem un algoritm de învățare nesupravegheat pentru segmentarea obiectelor din prim-plan care oferă posibilitatea de îmbunătățire în mai multe iterații. Metoda noastră combină în moduri complementare mai multe module care sunt potrivite pentru această sarcină.

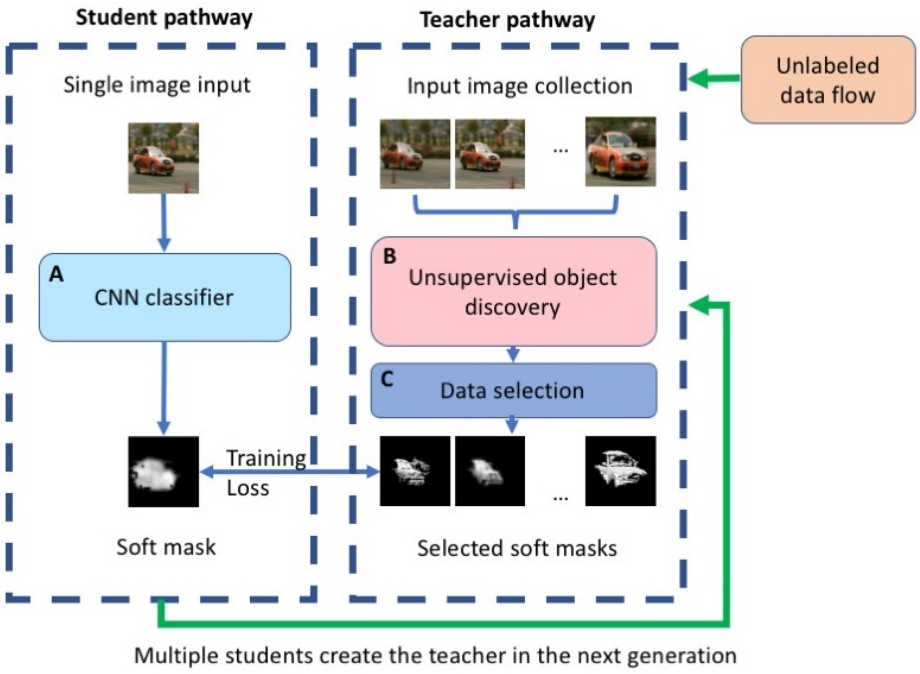

Figure 3: **Sistemul dual elev-profesor propus pentru învățarea nesupravegheată pentru a segmenta obiectele din prim plan în imagini.** Are două căi: calea profesor, care descoperă obiecte în videoclipuri sau colecțiile mari de imagini (modulul B) detectează obiectele din prim-plan. Măștile rezultate sunt apoi filtrate pe baza unei proceduri de selecție a datelor nesupravegheate (modulul C). Setul final de perechi rezultat - imaginea de intrare (sau cadru video) și masca pentru acel cadru particular (care acționează ca o etichetă nesupravegheată) - sunt folosite pentru a antrena calea student (modulul A). Întregul proces poate fi repetat de-a lungul mai multor generații. La fiecare generație sunt antrenați mai mulți studenți, care contribuie colectiv la antrenarea unui modul de selecție C mai puternic (modelat de o rețea neuronală adâncă) și formează o cale profesor mai puternică la următoarea iterație a algoritmului.

Cale student (modulul A): segmentare cu o singură imagine. Calea student (modulul A din figura 3) constă într-o rețea convoluționala adâncă. Testăm diferite arhitecturi de rețea, dintre care unele sunt utilizate în mod obișnuit în literatura recentă privind segmentarea semantică a imaginilor. Creăm un mic grup de arhitecturi relativ diverse. În total, folosim 5 arhitecturi diferite: LowRes-Net (care produce o segmentare cu rezoluție mică), FConv (o rețea complet convoluțională) și trei variante de U-Nets [44].

Combinarea mai multor rețele de studenți. Rețelele de studenți cu arhitecturi diferite produc rezultate variate care diferă calitativ. Au puncte forte diferite, în timp ce fac diferite tipuri de greșeli. Diversitatea lor va sta la baza creării căii profesor la următoarea generație.

Am experimentat cu ideea de a folosi mai multe rețele de studenți, combinându-le pentru a forma un ansamblu sau lăsându-le să producă segmentări independente separate pentru fiecare imagine. În sistemul nostru final am preferat cea de-a doua abordare, care este mai practică, mai ușor de implementat și oferă libertatea de a lasa studenții să ruleze independent, în paralel, fără a fi nevoie să-și sincronizeze rezultatele.

Formăm un ansamblu, pe care îl numim Multi-Net, rezultatul acestuia final este cel obținut prin înmulțirea pixelilor măștilor produse de fiecare rețea student individuală. Astfel, doar pixelii pozitivi supraviețuiesc pentru a produce segmentarea finală. Studenții de la a doua iterație sunt toți antrenați direct pe rezultate de la studenții individuali de la prima iterație, filtrate cu EvalSeg-Net (metoda de selecție). Multi-Net este folosit doar pentru a antrena rețeaua de selecție nesupravegheată, EvalSeg-Net.

Profesor (modulul B): descoperirea nesupravegheată a obiectelor. Pentru modulul B din figura 3 la prima iterație, folosim algoritmul VideoPCA, care face parte din sistemul introdus în [51]. Pentru a doua iterație am luat în considerare toate rezultatele de la toți cei cinci studenți pentru a forma etichetele de antrenare.

Selecție nesupravegheată de măști (modulul C) Valoarea medie a pixelilor diferiti de zero din masca soft este folosita ca indicator de scor pentru fiecare cadru segmentat la prima iterație.

La următoarele iterații, propunem o modalitate nesupravegheată de a învăța pentru a estima calitatea segmentării, numita EvalSeg-Net. Multi-Net oferă măști de calitate superioară, deoarece anulează erorile din rețelele individuale ale studenților. Astfel, folosim asemănarea

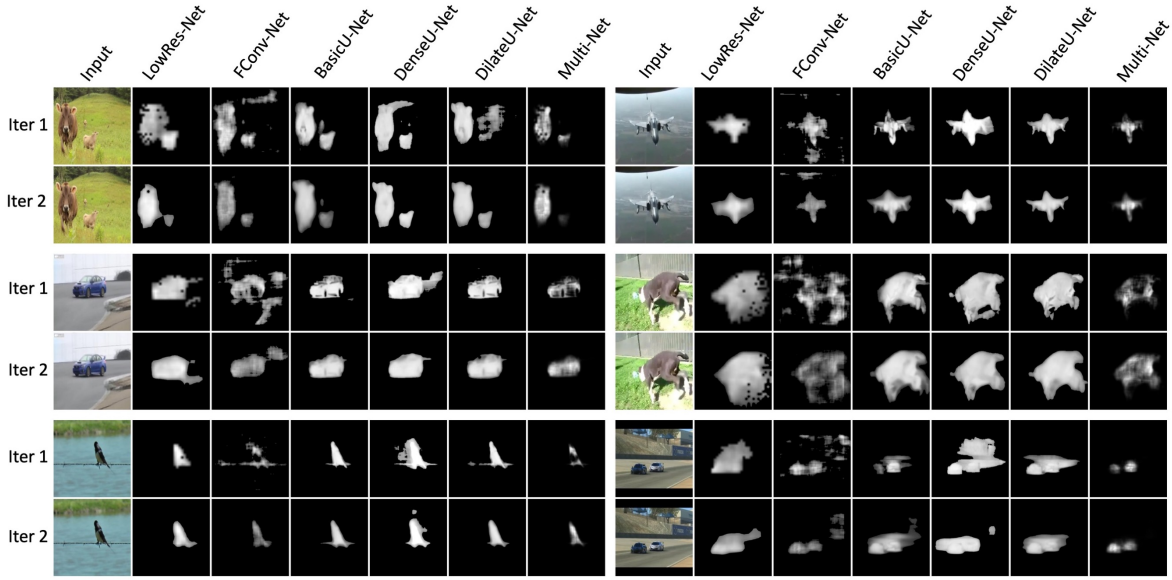

Figure 4: **Comparație vizuală între modele la fiecare iterație (generație).** Multi-Net, prezentată pentru comparație, reprezintă multiplicarea pixelilor între cele cinci modele. Vă rog să observați măștile superioare la studenții din a doua generație, cu forme mai bune, mai puține găuri și margini mai ascuțite.

dintre o anumită segmentare individuală și ansamblul Mulți-Net, pentru a produce un cost ce estimează calitatea segmentării. Folosind acest cost nesupravegheat putem antrena o rețea neurala adâncă, numita EvalSeg-Net.

În Figura 4 prezentăm o comparație vizuală între studenți și iterații. Putem observa că la a doua iterație calitatea segmentării se îmbunătățește.

3.4 Analiza experimentală

Comparație cu metode recente

Am efectuat mai întâi comparații cu metode special concepute pentru descoperirea obiectelor în video. Pentru aceasta, alegem setul de date YouTube Objects și comparăm metoda noastră cu cele mai bune metode din literatură (Tabel 2). Evaluările sunt efectuate pentru ambele versiuni ale setului de date YouTube Objects, YTOv1 ([39]) și YTOv2.2 ([25]).

Method	Aero	Bird	Boat	Car	Cat	Cow	Dog	Horse	Mbike	Train	Avg	Time	Version
[39]	51.7	17.5	34.4	34.7	22.3	17.9	13.5	26.7	41.2	25.0	28.5	N/A	v1
[36]	65.4	67.3	38.9	65.2	46.3	40.2	65.3	48.4	39.0	25.0	50.1	4s	
[24]	64.3	63.2	73.3	68.9	44.4	62.5	71.4	52.3	**78.6**	23.1	60.2	N/A	
[18]	76.3	**71.4**	65.0	58.9	**68.0**	55.9	70.6	33.3	69.7	42.4	61.1	0.35s	
LowRes-Net$_{\text{iter1}}$	77.0	67.5	**77.2**	68.4	54.5	68.3	72.0	**56.7**	44.1	34.9	62.1	0.02s	
LowRes-Net$_{\text{iter2}}$	**83.3**	**71.4**	74.3	69.6	57.4	**80.0**	77.3	**56.7**	50.0	37.2	65.7	0.02s	
DilateU-Net$_{\text{iter2}}$	**83.3**	66.2	**77.2**	**70.9**	63.4	75.0	**80.0**	53.3	50.0	**44.2**	**66.4**	0.02s	
Multi-Net$_{\text{iter2}}$(ensemble)	*87.4*	*72.7*	*77.2*	64.6	62.4	75.0	*82.7*	*56.7*	52.9	39.5	*67.1*	0.15s	
[18]	76.3	**68.5**	**54.5**	50.4	**59.8**	42.4	53.5	30.0	**53.5**	**60.7**	54.9	0.35s	v2.2
LowRes-Net$_{\text{iter1}}$	75.7	56.0	52.7	57.3	46.9	57.0	48.9	44.0	27.2	56.2	52.2	0.02s	
LowRes-Net$_{\text{iter2}}$	79.0	48.2	51.0	**62.1**	46.9	65.7	55.3	**50.6**	36.1	52.4	54.7	0.02s	
DilateU-Net$_{\text{iter2}}$	**84.3**	49.9	52.7	61.4	50.3	**68.8**	**56.4**	47.1	36.1	56.7	**56.4**	0.02s	
Multi-Net$_{\text{iter2}}$(ensemble)	83.1	53.2	54.3	*63.7*	50.6	*69.2*	*61.0*	*51.1*	37.2	48.7	*57.2*	0.15s	

Table 2: **Rezultate pe setul de date Youtube Objects, versiunile v1 ([39]) și v2.2 ([25]).** Obținem rezultate de ultimă generație pe ambele versiuni. Vă rugăm să rețineți că LowRes-Net obține deja rezultate de top pe v1, în timp ce este aproape de cel mai bun pe v2.2. Prezentăm rezultate ale modelelor individuale de top și ale ansamblului și, de asemenea, păstrăm reteaua de baza LowRes-Net la ambele iterații, pentru referință. Pentru fiecare coloana evidentiem îngroșat cel mai bun model si cu albastru italic cazurile in care ansamblul este mai bun sau egal.

3.5 Concluzii

În această teză, prezentăm o abordare nouă și eficientă a învățării din colecții mari de imagini și videoclipuri, într-un mod nesupravegheat, pentru a segmenta obiectele din prim-plan în imagini individuale. Prezentăm o metodă generală pentru această sarcină, care oferă posibilitatea de a învăța mai multe generații de studenti și profesori. Demonstrăm în practică că sistemul își îmbunătățește performanța pe parcursul a două generații. Sistemul nostru este unul dintre primele din literatură care învață să detecteze și să segmenteze obiectele din prim plan în imagini într-un mod nesupravegheat, fără a folosi caracteristici pre-antrenate sau etichetare manuală, în timp ce necesită o singură imagine la momentul testării.

Rețelele convoluționale antrenate de-a lungul căii student sunt capabile să învețe caracteristicile generale ale obiectelor, care includ formă bună, închidere, contururi netede, precum și contrastul cu fundalul. Ceea ce profesorul VideoPCA inițial mai simplu descoperă de-a lungul timpului, studentul complex este capabil să învețe la diferite niveluri de abstractizare.

Capitolul 4

TEACHTEXT: Distilare generalizată pentru regăsirea text-video

În ultimii ani, progrese considerabile în sarcina de regăsire text-video au fost realizate prin valorificarea pregătirii preliminare la scară largă pe seturi de date vizuale și audio pentru a construi codificatoare video puternice. Prin contrast, în ciuda simetriei firești, proiectarea unor algoritmi eficienți pentru exploatare pregătirii în prealabil a limbajului natural rămâne subexploatată. În această lucrare, suntem primii care investigăm proiectarea unor astfel de algoritmi și propunem o nouă metodă de distilare generalizată, TEACHTEXT, care folosește indicii complementare de la mai multe codificatoare de text pentru a furniza un semnal de supraveghere adițional modelului de regăsire. Mai mult, ne extindem metoda la modalitățile secundare video și arătăm că putem reduce efectiv numărul de modalități utilizate în timpul testării, fără a compromite performanța. Abordarea noastră avansează performanța pentru mai multe seturi de date importante și nu adaugă nicio suprasarcină de calcul la momentul testării. Nu în ultimul rând, arătăm o aplicare eficientă a metodei noastre pentru eliminarea zgomotului din seturile de date de regăsire. Codul și datele folosite pot fi găsite la https://www.robots.ox.ac.uk/˜vgg/ research/teachtext/.

Acest capitol se bazează pe lucrarea "TEACHTEXT: CrossModal Generalized Distil-lation for Text-Video Retrieval." Croitoru, Ioana, Simion-Vlad Bogolin, Marius Leordeanu, Hailin Jin, Andrew Zisserman, Samuel Albanie and Yang Liu. International Conference on Computer Vision. 2021.

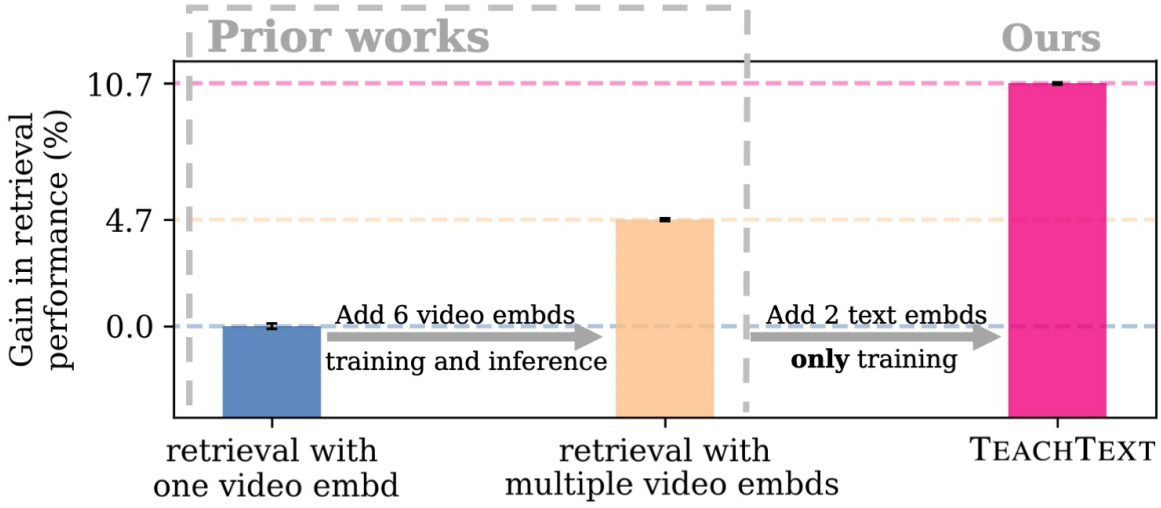

Figure 5: **Distilarea cunoștințelor din mai multe codificatoare de text pentru o performanta îmbunătățită pentru sarcina de regăsire text-video.** Lucrările anterioare [33, 30, 15] au arătat beneficiul considerabil al trecerii de la codificatoarele video care ingerează o singură modalitate (*stânga*) la codificatoare video multimodale (*centru*). În această lucrare, arătăm că performanța de regăsire poate fi îmbunătățită semnificativ prin învățarea de la mai multe codificatoare de text prin algoritmul TEACHTEXT care nu impune costuri suplimentare în timpul inferenței. Creșterea performanței de regăsire text-video (media geometrică a R1-R5-R10) este raportată pentru un model [30], precum și pentru metoda noastră pe setul de date MSR-VTT [56].

4.1 Introducere

Accentul acestei lucrări este *regăsirea text-video* — sarcina de a identifica ce videoclip dintr-un grup de candidați se potrivește cel mai bine cu o interogare în limbaj natural care descrie conținutul acestuia. Căutarea video are o gamă largă de aplicații în domenii precum monitorizarea faunei sălbatice, securitatea, monitorizarea proceselor industriale și divertisment. Mai mult, pe măsură ce omenirea continuă să producă videoclipuri la o scară din ce în ce mai mare, capacitatea de a efectua astfel de căutări în mod eficient capătă o importanță comercială critică pentru platformele de găzduire video precum YouTube.

O temă centrală a metodelor de regăsire propuse recent a fost investigarea modului de utilizare optimă a mai multor modalități video pentru a îmbunătăți performanța. În special, arhitecturile bazate pe amestecuri de experți [33, 30] și transformatoare multimodale [15] au arătat avantajul utilizării diverselor seturi de modele pre-antrenate pentru sarcini conexe (cum ar fi clasificarea imaginilor, recunoașterea acțiunii și clasificarea sunetului ambiental) ca bază pentru codificarea video în timpul antrenamentului și testării.

În această lucrare, examinăm dacă ar putea fi obținute câștiguri similare prin valorificarea mai multor codificatoare de text învățate pe corpuri textuale scrise la scară largă. Spre deosebire de codificatoarele video care utilizează mai multe modalități și sarcini de preantrenare, este mai puțin evident că există o diversitate suficientă între colecțiile de text pentru a obține o creștere semnificativă a performanței. De fapt, inspirația noastră provine dintr-o investigație atentă a performanței diferitelor codificatoare de text. În mod surprinzător, observăm nu numai că există o variație considerabilă a performanței între codificatoarele de text, ci și că *topul lor nu este consecvent intre seturile de date*, susținând cu tărie ideea de a utiliza mai multe codificatoare de text.

Motivați de această constatare, propunem un algoritm simplu, TEACHTEXT, pentru a exploata eficient cunoștințele capturate de colecții de text. Abordarea noastră necesită un model "student" pentru a învăța de la un singur sau mai multe modele de regăsire "profesor" cu acces la diferite codificatoare de text prin distilarea matricelor de similaritate text-video. După cum se observă în Fig. 5, sistemul propus TEACHTEXT este capabil să ofere un câștig semnificativ de performanță. Mai mult, acest câștig este complementar cu cel de adăugare a mai multor modalități video la codificatorul video, dar important, spre deosebire de adăugarea modalităților video, nu implică costuri de calcul suplimentare în timpul inferenței.

În Fig. 5 evidențiem faptul că performanța câștigată pentru un model care utilizează mai multe codificari de text (ultima bară) este comparabila cu câștigul unui model care utilizează mai multe modalități video (bara de mijloc). Prima bară reprezintă modelul CE [30] antrenat cu o singura modalitate video, și anume **Obj(IG)**. A doua bară reprezintă un model CE folosind 7 modalități video atât pentru inferență, cât și pentru antrenament. În cea de-a treia și ultima bară a diagramei prezentăm performanța utilizării a trei codificatoare de text diferite cu TEACHTEXT la antrenament, în timp ce folosim o singură codificare de text la momentul inferenței.

Principalele noastre contribuții sunt: (1) Propunem algoritmul TEACHTEXT, care folosește informațiile suplimentare date de utilizarea mai multor codificatoare de text; (2) Arătăm că învățarea directă a matricei de similaritate de regăsire între codificarea video este o tehnică eficientă de distilare generalizată pentru această sarcină (și comparăm abordarea noastră față de metode alternative, cum ar fi distilarea relației unimodale [37]); (3) Arătăm o aplicare a abordării noastre în eliminarea zgomotului din seturile de date de antrenament moderne

pentru sarcina de regăsire text-video; (4) Demonstrăm eficacitatea abordării noastre în mod empiric, obținând performanțe de ultimă generație pe șase seturi de date de regăsire text-video.

4.2 Context științific

Metode de regăsire video. Sarcina de indexare a conținutului video pentru a permite regăsirea are o istorie bogată în vederea artificială — au fost dezvoltate sisteme sofisticate pentru a găsi obiecte specifice [49], acțiuni [27] etc. În această lucrare, ne concentrăm pe sarcina de a regăsi conținut care se potrivește cu o anumită descriere a limbajului natural. Pentru această sarcină specială, există un interes considerabil în dezvoltarea metodelor intermodale care utilizează un spațiu de codificare comun pentru interogări de text și conținut video [1]. Aceste codificări comune video-text, care urmăresc să mapeze videoclipuri și descrieri de text într-un spațiu comun, astfel încât perechile video și text potrivite să fie apropiate, formează un model de calcul atractiv pentru abordarea acestei probleme, deoarece permit o indexare eficientă. Recent, două teme cheie au apărut pentru îmbunătățirea calității acestor codificări. În primul rând, metodele de preantrenare la scară largă, slab supravegheate [34], au căutat să-și extindă datele de antrenament prin exploatarea discursului conținut în videoclipuri ca un semnal de supraveghere. În al doilea rând, integrarea mai multor modalități (care a fost mult timp considerată importantă pentru indexarea semantică [50]) s-a dovedit că produce câștiguri semnificative în performanță [33, 30]. Ne concentrăm pe candidații din această ultimă categorie ca bază pentru investigarea abordării noastre.

Distilarea cunoștințelor/Informații privilegiate. Scopul distilarii cunoștințelor este de a transfera cunoștințe de la un model (profesor) la altul (student). Această idee a fost introdusă inițial în contextul simplificării arborilor de decizie [4] și a compresiei modelului [5], iar mai târziu extinsă de [19] care a formalizat acest transfer de cunoștințe ca proces parametrizat în funcție de temperatură, numindu-l *distilarea cunoștințelor*. Conceptul a fost generalizat în continuare în cadrul unificator al *distilării generalizată* [32] pentru învățarea cu informații privilegiate [53] (prin *controlul similarității* și *transferul de cunoștințe* [52]), împreună cu distilare de cunoștințe [19]. Abordarea noastră distilează cunostinte despre asemănările dintre video și text și, prin urmare, reprezintă o formă de distilare generalizată.

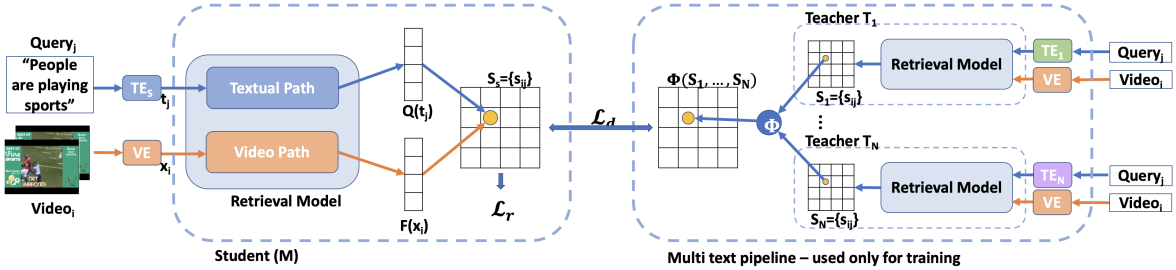

Figure 6: **TEACHTEXT prezentare generală a cadrului profesor-student.** Având în vedere un set de videoclipuri și interogări în limbaj natural în timpul antrenamentului, modelul student, M (stânga) și modelele profesor $T_1,\ldots,T_N$ (dreapta) produc fiecare matrici de similaritate. Matricea de similaritate produsă de M este încurajată să se potrivească cu matricele agregate ale profesorilor folosind un cost de distilare $\mathscr{L}_d$ în plus față de costul de regăsire $\mathscr{L}_r$. Rețineți că atât studentul, cât și profesorii folosesc aceiași codificatori video (VE), dar folosesc codificatori de text diferiti (TE$_S$ pentru student, TE$_1$,…,TE$_N$ pentru profesori). În timpul testarii, modelele de profesor sunt ignorate.

4.3 Metodă

Algoritmul TEACHTEXT. Propunem algoritmul TEACHTEXT care caută să exploateze indicii din mai multe codificări de text. O prezentare generală a abordării noastre este oferită în Fig. 6. În faza inițială a algoritmului, antrenăm o colecție de modele profesor $\{T_k : k \in \{1,\ldots,N\}\}$ pentru sarcina de regăsire text-video. Profesorii au aceeași arhitectură, dar fiecare model T_k folosește o codificare diferită de text ca intrare (folosind un codificator de text pre-antrenat TE_k). În a doua fază sunt înghețați parametrii profesorilor. Apoi procedăm prin eșantionarea unui lot de B perechi de videoclipuri și subtitrări și calculăm o matrice de similaritate corespunzătoare $S_k \in \mathbb{R}^{B\times B}$ pentru fiecare profesor T_k (Fig. 6 dreapta). Aceste matrice de similitudine N sunt apoi combinate cu o funcție de agregare, $\Phi : \mathbb{R}^{N\times B\times B} \rightarrow \mathbb{R}^{B\times B}$, pentru a formează o singură matrice de similaritate (Fig. 6, centru-dreapta). Concomitent, setul de videoclipuri și subtitrări sunt procesate de modelul student, M, care produce o altă matrice de similaritate, $S_s \in \mathbb{R}^{B\times B}$. În cele din urmă, în plus față de costul standard de regasire, un cost de distilare, $\mathscr{L}_d$, încurajează S_s să se afle aproape de agregatul $\Phi(S_1,\ldots,S_N)$. În timpul inferenței, modelele profesorului sunt eliminate, iar modelul student M necesită doar o singură codificare de text. În continuare, oferim detalii despre pierderea prin distilare utilizată pentru a conduce învățarea matricei de similaritate.

Modelul studentului. Un avantaj cheie al abordării noastre este că este agnostică raportată la arhitectura studentului și a profesorilor și, prin urmare, studentul (și profesorii) poate folosi

Model	$R@1 \uparrow$	$R@5 \uparrow$	$R@10 \uparrow$	$MdR \downarrow$
Dual[13]	7.7	22.0	31.8	32.0
HGR[7]	9.2	26.2	36.5	24.0
MoEE[33]	$11.1_{\pm 0.1}$	$30.7_{\pm 0.1}$	$42.9_{\pm 0.1}$	$15.0_{\pm 0.0}$
CE[30]	$11.0_{\pm 0.0}$	$30.8_{\pm 0.1}$	$43.3_{\pm 0.3}$	$15.0_{\pm 0.0}$
TT-CE	$11.8_{\pm 0.1}$	$32.7_{\pm 0.1}$	$45.3_{\pm 0.1}$	$13.0_{\pm 0.0}$
TT-CE+	$\mathbf{15.0}_{\pm 0.1}$	$\mathbf{38.5}_{\pm 0.1}$	$\mathbf{51.7}_{\pm 0.1}$	$\mathbf{10.0}_{\pm 0.0}$

Table 3: **MSR-VTT: Comparație cu metode recent.**

orice metodă din literatura actuală.

Modelul profesorului. Modelele profesorului folosesc aceeași arhitectură ca modelul studentului. Concret, creăm un grup de mai mulți profesori, fiecare folosind o codificare diferită de text pre-antrenată ca intrare. Codificările de text candidat pe care le considerăm în această lucrare sunt: mt_grovle [6], openai-gpt [41], gpt2-large [42], gpt2-xl [**?**] , w2v [35]. În acest fel, obținem un set de până la cinci modele care formează profesorii T_k, $k = 1..5$ folosite de TEACHTEXT.

4.4 Setare experimentală

Comparație cu metodele anterioare. După cum se poate vedea în Tab.3 abordarea noastră este eficientă și obține rezultate de top. Toate metodele sunt antrenate pentru sarcina de regăsire folosind numai date din seturile de date țintă. Mai mult, pentru a fi cât mai corecti posibil, în fiecare comparație am inclus rezultatele TEACHTEXT (abreviat TT în tabel) aplicate și celei mai bune metode existente pentru acel set de date.

4.5 Concluzii

În această lucrare, prezentăm un nou algoritm TEACHTEXT pentru sarcina de regăsire text-video. Folosim o paradigmă profesor-student în care un student învață să folosească informațiile suplimentare oferite de unul sau mai mulți profesori, având aceiași arhitectura, dar fiecare folosind o codificare diferită de text. În acest fel, obținem rezultate de top pe șase seturi de date.

Capitolul 5

Concluzii

În această teză am arătat cum distilarea generalizată poate fi aplicată în diferite sarcini, cum ar fi segmentarea obiectelor, detectarea obiectelor și regăsirea text-video. Pentru scenariul nesupravegheat am luat în considerare sarcinile de segmentare a obiectelor și de detectare a obiectelor și am arătat că tehnica de distilare poate fi aplicată cu succes pentru a învăța de la un profesor nesupravegheat care generează automat adnotări. Mai mult, am demonstrat că printr-un simplu proces de selecție adnotările zgomotoase pot fi filtrate, ceea ce a îmbunătățit rezultatele. În plus, am arătat că putem forma ansambluri de mai mulți studenți pentru a îmbunătăți rezultatele. Aceste ansambluri pot forma un nou profesor în a doua iterație pentru a spori performanța studenților noi pregătiți. Demonstrăm în practică că sistemul își îmbunătățește performanța pe parcursul a două generații. Sistemul prezentat este unul dintre primele din literatură care învață să segmenteze obiectele din prim-plan în imagini într-un mod nesupravegheat.

Pentru scenariul supravegheat, am luat în considerare sarcina de regasire text-video și am arătat că, folosind distilarea generalizată, putem învăța de la profesori antrenati cu diferite codificatoare de text pre-antrenate și putem îmbunătăți performanța. Folosind diferite codificatoare de text pre-antrenate, performanța variază drastic, sugerând prezența unor informații complementare. Am profitat de această informație complementară folosind distilarea generalizată.

Bibliografie

[1] Aytar, Y., Shah, M., and Luo, J. (2008). Utilizing semantic word similarity measures for video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] Beluch, W. H., Genewein, T., Nürnberger, A., and Köhler, J. M. (2018). The power of ensembles for active learning in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[3] Bogolin, S.-V., Croitoru, I., and Leordeanu, M. (2020). A hierarchical approach to vision-based language generation: from simple sentences to complex natural language. In *Proceedings of the International Conference on Computational Linguistics (COLING)*.

[4] Breiman, L. and Shang, N. (1996). Born again trees. *University of California, Berkeley, Berkeley, CA, Technical Report*, 1:2.

[5] Buciluǎ, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[6] Burns, A., Tan, R., Saenko, K., Sclaroff, S., and Plummer, B. A. (2019). Language features matter: Effective language representations for vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[7] Chen, S., Zhao, Y., Jin, Q., and Wu, Q. (2020). Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Chen, X., Shrivastava, A., and Gupta, A. (2014). Enriching visual knowledge bases

via object discovery and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Croitoru, I., Bogolin, S.-V., and Leordeanu, M. (2017). Unsupervised learning from video to detect foreground objects in single images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[10] Croitoru, I., Bogolin, S.-V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., and Liu, Y. (2021). Teachtext: Crossmodal generalized distillation for text-video retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[11] Deselaers, T., Alexe, B., and Ferrari, V. (2012). Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision (IJCV)*, 100(3).

[12] Doersch, C., Gupta, A., and Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[13] Dong, J., Li, X., Xu, C., Ji, S., and Wang, X. (2019). Dual dense encoding for zero-example video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision (IJCV)*, 111(1):98–136.

[15] Gabeur, V., Sun, C., Alahari, K., and Schmid, C. (2020). Multi-modal transformer for video retrieval. *Proceedings of the European Conference on Computer Vision (ECCV)*.

[16] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[17] Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386.

[18] Haller, E. and Leordeanu, M. (2017). Unsupervised object segmentation in video by efficient selection of highly probable positive features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[19] Hinton, G., Vinyals, O., and Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

[20] Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *Proceedings of the ACM Computing Surveys*, 31(3):264–323.

[21] Joulin, A., Bach, F., and Ponce, J. (2010). Discriminative clustering for image co-segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Joulin, A., Bach, F., and Ponce, J. (2012). Multi-class cosegmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[23] Joulin, A., Tang, K., and Fei-Fei, L. (2014). Efficient image and video co-localization with Frank-Wolfe algorithm. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[24] Jun Koh, Y., Jang, W.-D., and Kim, C.-S. (2016). Pod: Discovering primary objects in videos based on evolutionary refinement of object recurrence, background, and primary object models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Kalogeiton, V., Ferrari, V., and Schmid, C. (2016). Analysing domain shift factors between videos and images for object detection. *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 38(11).

[26] Kim, G., Xing, E., Fei-Fei, L., and Kanade, T. (2011). Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[27] Laptev, I. and Pérez, P. (2007). Retrieving actions in movies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[28] Larsson, G., Maire, M., and Shakhnarovich, G. (2016). Learning representations for automatic colorization. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[29] Lee, Y. J., Kim, J., and Grauman, K. (2011). Key-segments for video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[30] Liu, Y., Albanie, S., Nagrani, A., and Zisserman, A. (2019). Use what you have: Video retrieval using representations from collaborative experts. *The British Machine Vision Conference (BMVC)*.

[31] Liu, Y. and Lapata, M. (2019). Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

[32] Lopez-Paz, D., Bottou, L., Schölkopf, B., and Vapnik, V. (2016). Unifying distillation and privileged information. *Proceedings of the International Conference on Learning Representations (ICLR)*.

[33] Miech, A., Laptev, I., and Sivic, J. (2018). Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*.

[34] Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[35] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[36] Papazoglou, A. and Ferrari, V. (2013). Fast object segmentation in unconstrained video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[37] Park, W., Kim, D., Lu, Y., and Cho, M. (2019). Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[38] Pathak, D., Girshick, R., Dollar, P., Darrell, T., and Hariharan, B. (2017). Learning features by watching objects move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] Prest, A., Leistner, C., Civera, J., Schmid, C., and Ferrari, V. (2012). Learning object class detectors from weakly annotated video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[40] Radenović, F., Tolias, G., and Chum, O. (2016). Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[41] Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*.

[42] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *preprint*, 1(8):9.

[43] Raina, R., Battle, A., Lee, H., Packer, B., and Ng, A. Y. (2007). Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

[44] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

[45] Rother, C., Kolmogorov, V., and Blake, A. (2004). Grabcut: Interactive foreground extraction using iterated graph cuts. In *Proceedings of the ACM Transactions on Graphics*, volume 23, pages 309–314.

[46] Rubinstein, M., Joulin, A., Kopf, J., and Liu, C. (2013). Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[47] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3).

[48] Shen, D., Wu, G., and Suk, H.-I. (2017). Deep learning in medical image analysis. *Annual Review of Biomedical Engineering*, 19:221–248.

[49] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[50] Snoek, C. G. and Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35.

[51] Stretcu, O. and Leordeanu, M. (2015). Multiple frames matching for object discovery in video. In *The British Machine Vision Conference (BMVC)*.

[52] Vapnik, V. and Izmailov, R. (2015). Learning using privileged information: similarity control and knowledge transfer. *Journal of Machine Learning Research*, 16(1):2023–2049.

[53] Vapnik, V. and Vashist, A. (2009). A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557.

[54] Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE Transactions on Neural Networks*, 10(5):988–999.

[55] Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A. N., Gouws, S., Jones, L., Kaiser, Ł., Kalchbrenner, N., Parmar, N., et al. (2018). Tensor2tensor for neural machine translation. *arXiv preprint arXiv:1803.07416*.

[56] Xu, J., Mei, T., Yao, T., and Rui, Y. (2016). Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[57] You, Q., Jin, H., Wang, Z., Fang, C., and Luo, J. (2016). Image captioning with semantic attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.