ACADEMIA ROMÂNĂ

Şcoala de Studii Avansate a Academiei Române

Institutul de Matematică ”Simion Stoilow”

REZUMATUL TEZEI DE DOCTORAT

Descrierea activităților complexe în limbaj natural
pe baza consensului între mai multe rețele neurale și utilizarea
de reprezentări ierarhice

Coordonator științific:
Prof. Dr. Marius Leordeanu

Doctorand:
Simion-Vlad Bogolin

BUCUREȘTI
2022

Cuprins

Sinopsis

În această lucrare, ne propunem să descriem activități complexe în limbaj natural. Din punctul nostru de vedere, acesta este cel mai potrivit mod, deoarece limbajul natural este atât de versatil încât descrierea activităților nu se limitează la o listă fixă de etichete. Cu toate acestea, descrierea automată a acțiunilor descrise în videoclipuri în limbaj natural este o problemă dificilă, iar studiul ei ne poate ajuta să înțelegem mai bine legătura dintre video și limbă. Începem prin a introduce o abordare nouă, bazată pe consens, care valorifică puterea mai multor modele și folosindu-se de consens generează descrieri textuale mai diverse. Apoi, discutăm relațiile dintre conținutul vizual și limbaj și evidențiem unele dintre provocările din spatele acestei probleme. În final, propunem o abordare ierarhică, prin generarea, în primă fază a unor descrieri formate din propoziții simple, urmate la nivelul următor de o descriere mai complexă și mai fluentă în limbaj natural. În timp ce propozițiile simple descriu acțiuni simple sub forma de (subiect, verb, obiect), descrierile la nivel doi, formate dintr-un paragraf, prezintă conținutul vizual într-o manieră mai compactă, mai coerentă și mai bogată semantic. În acest scop, introducem primul set de date video din literatură care este adnotat la două niveluri de complexitate lingvistică. Efectuăm teste extinse care demonstrează că reprezentarea noastră lingvistică ierarhică, de la limbajul simplu la cel complex, ne permite să antrenăm o rețea în două etape care este capabilă să genereze paragrafe semnificativ mai complexe decât abordările actuale într-o singură etapă. În acest fel, obținem un sistem care este capabil să utilizeze schema de adnotare pe două niveluri și să descrie activități complexe în limbaj natural.

Capitolul 1

Introducere

În această lucrare, scopul nostru este de a descrie activități complexe în limbaj natural. În acest fel, ne propunem să îmbinăm două probleme diferite, dar foarte populare: recunoașterea activităților și translatarea videourilor în text.

În abordarea noastră, începem prin a transcrie mai întâi videourile în propoziții, găsind consensul între mai multe modele antrenate pentru această problemă. Fiecare model folosește o paradigmă codificator-decodor, în care codificatorul primește ca intrare secvența video și decodorul generează descrierea în limbaj natural. Aceste modele prezic de obicei propoziții valide, bine formate, dar care pot fi greșite la un nivel semantic, neexprimând întotdeanua exact ceea ce se întâmplă de fapt în videoclip. Consensul dintre mai multe modele ajută la depășirea acestei probleme și la îmbunătățirea corelației semantice.

Continuăm prin a face un studiu amplu al acestei abordări pentru a-i afla limitele. Chiar dacă propozițiile generate sunt corecte din punct de vedere gramatical și descriu conținutul vizual, ele nu sunt foarte complexe și folosesc un vocabular simplu. Ca o a doua contribuție majoră, facem un studiu amplu asupra relației dintre text și datele vizuale. Pentru a elimina necunoscutele legate de partea de generare a textului, realizăm majoritatea experimentelor pe problema de regăsire a videourilor. În acest caz, dându-se o propoziție în limbaj natural, scopul sistemului este de a găsi un videoclip (dintr-o colecție dată de videoclipuri) care este cel mai bine descris de această propoziție. Mai mult decât atât, pe baza acestei analize, suntem capabili să dezvoltăm un sistem care să obțină rezultate de top pentru această problemă.

În cele din urmă, pentru a depăși unele dintre limitări, propunem o nouă abordare ierarhică a translatării video în text, în care mai întâi generăm câteva propoziții mai simple și apoi un paragraf mai mare, mai complex, care descrie întregul conținutul vizual [4]. Pentru a face posibilă o astfel de abordare și pentru a ne atinge scopul de a descrie activități complexe prin limbaj natural, propunem un nou set de date Videos-To-Paragraphs care utilizează o schemă de adnotare pe două niveluri. În acest fel, putem proiecta un sistem care să descrie printr-un paragraf activitățile umane ce apar într-un videoclip.

Deci, scopul nostru final este să studiem dacă o descriere lingvistică duală, intermediară, la nivelul propozițiilor scurte, care descriu acțiuni simple sub formă de (subiect, verb, obiect) ar putea ajuta la generarea unui limbaj mai complex și mai fluent. Bunul simț și provocările întâlnite în cercetările actuale sugerează cu tărie că trecerea de la date vizuale la limbaj ar putea beneficia de o abordare mai graduală: ar trebui mai întâi să detectăm actorii și obiectele din scenă, apoi să înțelegem și să descriem acțiunile lor în propoziții scurte și numai după aceea, putem pune totul împreună într-o poveste mai mare, coerentă, descrisă într-un limbaj natural fluent.

Astfel, propunem o traducere a videourilor în text în două etape: în prima etapă descriem videoclipul ca o succesiune de evenimente, sub formă de propoziții simple (subiect, verb, obiect), care sunt și ele bine localizate în timp și spatiu. Cu un ușor abuz de terminologie, ne referim la astfel de propoziții care descriu acțiuni simple SVO (subiect, verb, obiect) ca propoziții SVO, chiar dacă uneori pot conține mai mult de trei cuvinte. Apoi, descrierea de nivelul doi rezumă conținutul video într-un paragraf coerent, care descrie conținutul vizual într-un mod mai elaborat, adăugând la primul nivel de informații, relații cauzale și semantice dintre actori și evenimente (fără a fi nevoie să se repete propozițiile inițiale SVO). Pentru a putea antrena modele care să folosească pe deplin aceste propoziții mai simple, cât și paragraful mai complex, introducem un set de date nou, Videos-to-Paragraphs, care conține videoclipuri filmate în interior într-un univers mai restrâns: școala.

Apoi, efectuăm experimente ample, care arată că sistemul nostru de generare a descrierii videourilor în două etape beneficiază foarte mult de această reprezentare intermediare prin propozițiilor simple, care se situează între interpretarea vizuală pură (la nivel de obiecte singulare sau acțiuni ca simple ”etichete”) și interpretările mai complexe în limbaj natural (la nivel de paragrafe).

Capitolul 2

Translatarea videourilor în text

Problema de translatare a videoclipurilor în limbaj natural este una dintre cele mai interesante și încă nerezolvate probleme ale inteligenței artificiale de astăzi. Rezolvarea acestei sarcini ar ajuta la decodarea multor întrebări importante despre cum funcționează mintea, cum percepem lumea, cum gândim și apoi comunicăm unul cu celălalt. Metodele eficiente de translatare a datelor vizuale în limbaj natural ar avea, de asemenea, o valoare practică imensă, cu aplicații în multe domenii, de la tehnologie la medicină și divertisment.

În această lucrare, ne propunem să descriem activitățile umane în limbaj natural. Începem prin a introduce un model general de translatare video în text, care este descris în continuare.

2.1 Arhitecturi de rețea

În această secțiune, vom descrie toate arhitecturile utilizate de sistemul nostru de translatare video în text. Începem prin a descrie una dintre cele mai utilizate paradigme pentru această problemă, secvență-la-secvență (sequence-to-sequence or Seq2Seq) și apoi introducem arhitecturile propuse de noi.

Lucrările timpurii [1] pentru traducerea video în text abordează această problemă ca una de traducere automată. Cu toate acestea, lucrurile nu sunt atât de simple. În loc să traducem cuvinte dintr-o limbă în alta, acum trebuie să înțelegem și să traducem caracteristicile vizuale în limbă. Deci, unele schimbări sunt cu siguranță necesare. Videoclipurile, fiind o secvență

Acest capitol se bazează pe metoda Duta, Nicolicioiu, Bogolin and Leordeanu [9]

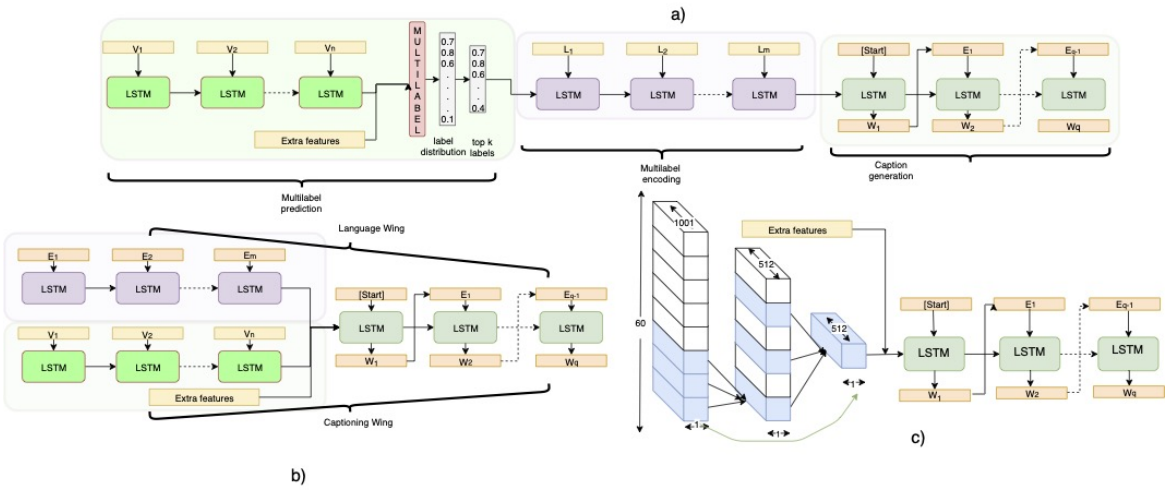


Figure 1: **Arhitecturi principale.** a) *Rețea în două etape*: date vizuale - cuvinte - propoziții, b) *Rețea cu două aripi*, c) *Rețeaua convoluțională temporală (TCN).* Arhitecturile diferă în structură în mod semnificativ și, în general, produc propoziții diferite, dar au o performanță generală similară. Imagine refolosită din [9].

de cadre, sunt în mod natural potrivite pentru a fi procesate de rețele neurale recurente. În acest fel, majoritatea modelelor secvență-la-secvență folosesc LSTM-uri pentru a produce un vector caracteristic la nivel de video. Acesta este de obicei numit codificator. Odată ce videoclipul este procesat (codat) este transmis apoi către un decodor care generează textul, de unde schema de numire secvență la secvență. Acum vom descrie arhitecturile propuse.

Rețea cu două aripi cu reconstrucție a limbajului natural. Modelul seq2seq tinde să producă, în experimente, propoziții simple cu un vocabular foarte limitat. În mod ideal, ne-am dori un decodor mai puternic, capabil să genereze propoziții mai realiste și complexe. Ne propunem să realizăm acest lucru printr-un model pe care îl numim **Rețeaua cu două aripi** (după cum se arată în Figura 1b). Cele două ramuri (aripi) sunt antrenate alternativ, decodorul având parametri comuni. Rețineți că modelul de reconstrucție a limbajului este utilizat numai în timpul antrenamentului pentru a învăța un decodor mai puternic. În timpul testării, este utilizată doar ramura video pentru a genera propoziții.

Rețea în două etape, de la video la cuvinte la propoziții. Al doilea model pe care îl propunem, Rețeaua în două etape, pune două rețele de codificator-decodor una după alta (Figura 1a). Rețeaua din prima etapă învață să genereze cuvinte din videoclipuri. În cel de-al doilea stadiu, rețeaua învață să producă propoziții din seturile de cuvinte generate în prima etapă. Astfel, etichetele cuvintelor oferă o interpretare semantică intermediară între datele video și propoziția finală.

	Grupa A		Grupa A+B		Grupa A+B+C	
Model	**Cider**	**Meteor**	**Cider**	**Meteor**	**Cider**	**Meteor**
Seq2Seq	36.0	25.5	44.0	27.4	46.1	28.3
Two-Wings	32.2	25.2	42.2	27.3	46.2	28.8
Two-Stage	34.9	25.2	43.3	27.4	45.7	28.4
TCN	36.80	25.5	43.9	27.4	46.1	28.4
Attention	41.0	26.6	44.2	27.5	46.4	28.5
Media	36.0 ±2.5	25.6 ±0.5	43.9 ±0.6	27.4 ±0.2	46.0 ±0.9	28.4 ±0.3
Cel mai bun model individual					46.2	28.8

Table 1: **Performanță vs caracteristici.** Performanța modelelor noastre folosind diferite caracteristici de imagine video și audio adăugate în timpul a trei faze experimentale: Grupa A - Caracteristici Inception; Grupa B - caracteristici audio C3D + MFCC; Grupa C - VGG audio + caracteristici de etichetare a cuvintelor Y8M. În fiecare fază raportăm rezultatele medii ale fiecărui tip de modele și media tuturor modelelor. Observați modul în care caracteristicile suplimentare pregătite în prealabil pentru diferite sarcini îmbunătățesc semnificativ performanța.

Rețea convoluțională temporală. Inspirat de [2], le adaptăm ideea unei arhitecturi de rețea de convoluție temporală (TCN) pentru a înlocui codificatorul rețelei neuronale recurente. Ideea din spatele TCN (Figura 1c) este de a surprinde modul în care caracteristicile se schimbă în timp, utilizând filtre temporale unidimensionale. Prin folosirea unei ierarhii de convoluții cu o rată de dilatare în creștere, cantitatea de informații combinată crește exponențial, la diferite scări de timp.

2.2 Analiză experimentală

Din experimentele noastre, așa cum se arată în Tabelul 1, puteți vedea modul în care fiecare arhitectură utilizată afectează rezultatele. Mai mult decât atât, se poate observa că informațiile complementare de nivel înalt sunt aduse de caracteristici diferite antrenate pentru diferite sarcini. Acest fapt sugerează cu tărie că nivelul intermediar de semantică captat de aceste caracteristici este important pentru o mai bună reducere a decalajului informațional dintre viziune și limbaj.

În acest capitol, am introdus trei noi arhitecturi pentru translatarea videourilor în text. În continuare, vom descrie abordarea noastră care utilizează consensul pentru a îmbunătății rezultatele.

Capitolul 3

Translatarea videourilor în text prin consensul mai multor rețele neurale

În acest capitol, prezentăm o abordare bazată pe consens pentru a face față provocărilor intrinseci care apar atunci când vorbim despre viziune și limbaj. Modelele tradiționale suferă de o pierdere a diversității în comparație cu propozițiile generate de om. Mai mult, fiecare propoziție este unică și același mesaj poate fi rostit într-un număr nelimitat de moduri. Pentru a rezolva aceste provocări, propunem o nouă abordare bazată pe găsirea descrierii lingvistice consensuale printre multiple modelele de translatare lingvistică. În timp ce fiecare model individual este capabil să genereze propoziții bine formate, care în general respectă regulile gramaticale, consensul dintre multe modele este cel care surprinde cel mai bine conținutul semantic și depășește semnificativ modelele individuale în ceea ce privește metricile de evaluare.

În timp ce modelele noastre ating un nivel de acuratețe care se opune bine literaturii publicate, există un grad relativ mare de variație în propozițiile lor generate din cauza diferitelor moduri în care codificăm conținutul vizual. Unele modele tind să aibă rezultate complexe, descriptive, cu un vocabular mai bogat, în timp ce altele generează propoziții simple, concise.

Am observat că grupul de propoziții conține foarte des propoziții corecte. Mai mult, am observat că modelele produc în general propoziții care gravitează în jurul sensului corect. Astfel, variațiile de propoziție zgomotoase ar putea fi eliminate dacă ansamblul rețelelor ar

Acest capitol se bazează pe metoda Duta, Nicolicioiu, Bogolin and Leordeanu [9]

putea funcționa în comun, în ansamblu.

Aici propunem un **algoritm de consens** eficient pentru selectarea celei mai bune propoziții din grup, compus din două etape - o primă etapă de consens folosind acorduri simple între propoziții și o a doua etapă care implică antrenarea unei rețele Oracol care alege propoziția mai bună după cum urmează:

- Pasul 1: Pentru fiecare propoziție din grup, calculează scorul CIDEr față de celelalte.
- Pasul 2: Păstrați C propozițiile cu cel mai bun scor.
- Pasul 3: Re-clasificați primele C propoziții folosind rețeaua Oracol și returnați propoziția cu cel mai bun punctaj.

3.1 Analiză experimentală

	Cider	**Meteor**	**Rouge**	**Bleu 4**
HRL [26]	48.0	28.7	61.7	41.3
dense [22]	48.9	28.3	61.1	41.4
CIDEnt-RL [17]	51.7	28.4	61.4	40.5
TGM [11]	52.9	**29.7**	-	**45.4**
Model individual	46.2	28.4	-	-
Consens	**53.8**	**29.7**	**63.0**	44.2

Table 2: **Comparație cu cele mai bune modelele din literatură pe setul de date MSR-VTT 2016.** Obținem rezultate de top pe trei metrici de evaluare.

În Tabelul 2 comparăm metoda noastră cu cele mai bune metode de la competiția MSR-VTT 2016, dar și cu modelele de top publicate după competiție pe acest set de date. Consensul dintre toate modelele îmbunătățește semnificativ performanța, obținând rezultate foarte bune.

În acest capitol, am introdus un nou algoritm bazat pe consens, care este capabil să genereze descrieri textuale mai bune. Aceasta înseamnă că suntem cu un pas mai aproape de obiectivul nostru de a descrie acțiuni prin limbaj natural. În plus, am obținut un sistem general care poate fi aplicat pe diverse tipuri de videoclipuri care obține rezultate de ultimă generație pe un set de date dificil, și anume MSR-VTT.

Capitolul 4

O discuție despre relația video-text

La o inspecție calitativă, am descoperit că, deși propozițiile generate de obicei se supun regulilor gramaticale, ele sunt adesea destul de generice. Așadar, am vrut să înțelegem mai bine relația dintre video și limbaj. În timp ce observațiile noastre sunt în concordanță cu literatura de specialitate ([20, 7]) conform căreia modelele au de obicei un vocabular mai puțin divers și generează propoziții mai scurte, am dorit, de asemenea, să studiem problema dintr-un unghi diferit. Din acest motiv, am considerat că cea mai bună opțiune pentru a înțelege mai bine relația dintre video și text este studierea acesteia pentru o altă problemă, și anume regăsirea text-video. În acest fel, putem studia relația dintre video și text fără a trebui să ținem cont de necunoscutele care apar în partea de generare a limbajului. În acest capitol, vom prezenta principalele constatări care decurg dintr-o etapă de analiză a erorilor a mai multor modele de ultimă generație pentru problema de regăsire text-video. Așadar, investigăm înglobările multimodale învățate de sistemele moderne de recuperare text-video, în care configurarea este foarte similară cu problema de traducere video în text.

Regăsirea text-video este problema în care, dându-se o propoziție în limbaj natural, scopul este de a găsi un videoclip dintr-o colecție predefinită de videoclipuri care este cel mai bine descris de acestă propoziție. Încorporarea multimodală formează un model de calcul atractiv pentru acestă problemă. Cu toate acestea, în timp ce performanța s-a îmbunătățit considerabil în ceea ce privește seturile de date, cum ar fi LSMDC [21], modelele în sine sunt inscrutabile: rămâne neclar care factori le influențează performanța și ce factori limitează progresul viitor. Un aspect interesant rezultă din analiza perturbațiilor. Scopul analizei perturbației este de a afla care cuvinte au o pondere mai mare față de o anumită regăsire și care cuvinte nu

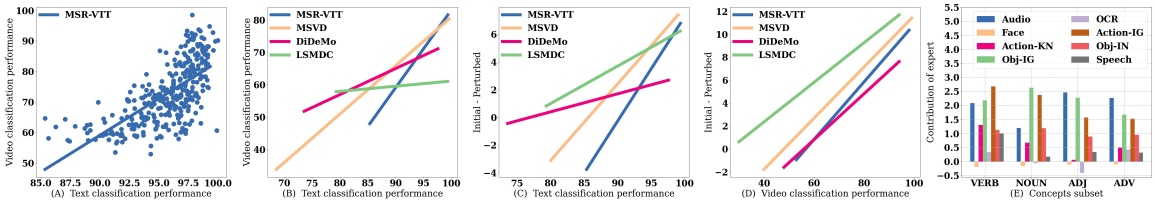


Figure 2: **Analiza clasificatoarelor de text și video.** În figura (A) este ilustrată performanța clasificatoarelor de text și video împreună cu linia potrivită pentru puncte (se poate observa cu ușurință că tendința este liniară). În figura (B), tendințele pentru toate seturile de date sunt prezentate pentru CE. Următoarele două figuri arată rezultatele perturbării (performanța originală de regăsire în medie geometrică minus performanța calculată folosind propozițiile perturbate) clasificarea textului w.r.t - (C) și clasificarea video - (D) performanța. În figura (E) este prezentată contribuția fiecărui expert pentru MSR-VTT per parte de discurs. Contribuția fiecărui expert este calculată pe baza performanței clasificatoarelor video în comparație cu linia de bază - linia de bază este expertul în scenă, așa cum este prezentat în [13].

contează atât de mult. Pentru a evalua acest lucru, am comparat performanța obținută de fiecare model înainte și după eliminarea unui anumit cuvânt din întregul set de testare. Dacă performanța se schimbă, atunci acel cuvânt este important pentru luarea deciziilor, în timp ce dacă performanța rămâne aceeași, cuvântul nu contează pentru luarea unei decizii.

Sensibilitatea conceptelor. Identificarea cuvintelor care au o pondere mare față de o anumită regăsire este importantă, dar am dorit să studiem relația dintre cuvinte și vectorul de caracteristici pentru video respectic text pentru a identifica dacă există un decalaj în ceea ce privește ce concepte pot fi derivate din care înglobări. Pentru a realiza acest lucru, am antrenat mai mulți clasificatori SVM binari liniari pentru fiecare concept, unul având ca intrare vectorul caracteristic al videourilor respectiv al textului pentru fiecare concept. Am văzut că, de obicei, clasificatoarele funcționează bine, având în vedere încorporarea textului, dar nu atât de bune pentru videoclipuri. În Figura 2 (A-D) puteți găsi performanța clasificatoarelor de text în comparație cu clasificatoarele video pentru toate cuvintele. După cum se poate observa, performanța clasificatorilor este corelată, totuși există un decalaj de performanță între clasificatorii textuali și clasificatorii video.

Concepte vs caracteristici vizuale. Mai mult, am dorit să studiem dacă și cum modalitățile utilizate pentru antrenarea modelului de regăsire contează pentru o performanță bună a clasificatoarelor pentru cuvintele care au fost prezente în datele pre-antrenate utilizate pentru fiecare modalitate si care nu conteaza. După cum se poate observa în Figura 2 (E), există o legătură clară între tipul de modalități pre-antrenate și categoria de cuvinte care pot fi clasifi-

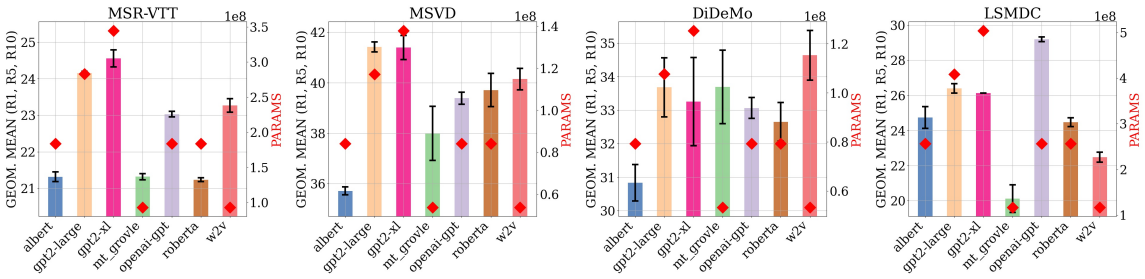


Figure 3: **Influența diverselor codificatoare de text**. Barele și axa Y din stânga indică media geometrică `t2v` a fiecărui model la R@1, R@5 și R@10. Markerii roșii și axa Y din dreapta indică numărul total de parametri utilizați de fiecare model. După cum se poate observa, modificarea caracteristicilor de text folosite are un impact major asupra performanței.

cate din înglobările video. Nu este surprinzător, dacă adăugăm caracteristici de recunoaștere a acțiunilor, atunci categoria care se îmbunătățește cel mai mult sunt verbele, în timp ce adăugarea caracteristicilor de clasificare îmbunătățește substantivele.

Apoi am vrut să vedem ce se întâmplă dacă se modifică caracteristicile textuale utilizate. Deci, pentru **studiul nostru de caracteristici textuale** puteți vedea influența modificării înglobărilor de text în Figura 3, unde axa y indică media geometrică (mai mare este mai bună) a fiecărui model la R1, R5 şi R10. Putem observa că această modificare are un impact major asupra performanței de recuperare. Așadar, chiar dacă aceste caracteristici de text sunt de obicei pre-antrenate pe o cantitate mare de date textuale, această fluctuație a performanței atât între seturi de date, cât și intra, sugerează că fiecare caracteristică de text poate avea informații suplimentare care pot fi utile pentru sarcina de regăsire. Aceasta este o perspectivă cheie care stă la baza modelului nostru de regăsire text-video pe care îl vom introduce în capitolul următor. În același timp, are și implicații puternice pentru problema de translatare video în text. Este un indiciu puternic potrivit căruia corespondența dintre video și text nu poate proveni numai din datele textuale. Aceasta înseamnă că avem nevoie de seturi de date text-video mari, la scară largă. În plus, generarea unui text mai lung poate fi problematică. Din acest motiv, introducem o nouă schemă de adnotare pe două niveluri pe care o vom descrie în Capitolul 7.

Capitolul 5

Regăsirea videourilor utilizând distilare generalizată

După cum s-a menționat anterior, scopul acestei lucrări este descrierea activităților complexe prin limbajul natural. Cu toate acestea, așa cum am discutat în capitolul anterior, pe baza sarcinii de regăsire text-video, am reușit să înțelegem mai bine relația dintre datele vizuale și limbajul natural. În acest capitol, vom prezenta pe scurt metoda noastră pentru această problemă care decurge din această analiză. Scopul este de a dezvolta un sistem care, dându-se o interogare textuală, găsește videoclipul care este cel mai bine descris de interogare, având în vedere o colecție de videoclipuri.

5.1 Algoritmul TEACHTEXT

Folosind distilare generalizată, suntem capabili să valorificăm cunoștințele din mai multe caracteristici de text și să pregătim un model student care este capabil să obțină rezultate foarte bune. Procesul începe prin formarea mai multor modele profesor, având aceeași arhitectură, dar folosind o caracteristică de text diferită. Apoi antrenăm un model student care este capabil să învețe de la profesor o matrice de similaritate agregată prin distilare generalizată. Deci, având în vedere un set de date format din perechi text-video, (x_i, t_j^k), scopul este de a antrena un model de recuperare $M = (F, Q)$ care învață o încorporare comună între video și text. F

Acest capitol se bazează pe metoda Croitoru, Bogolin, Leordeanu, Jin, Zisserman, Albanie and Liu [6]

și Q reprezintă încorporarea video pre-antrenată și, respectiv, text. Deci, dorim să atribuim o asemănare mare perechilor $F(x_i), Q(t_i)$ date la intrarea modelului nostru și o asemănare mai mică perechilor $F(x_i), Q(t_j), i \neq j$. Pentru a asigura acest lucru, folosim funcția de loss max margin ranking loss: $\mathscr{L}_r = \frac{1}{B}\sum_{i=1}^{B}\sum_{i\neq j}\left[max(0, s_{ij} - s_{ii} + m) + max(0, s_{ji} - s_{ii} + m)\right]$ unde $s_{ij} = F(x_i)^T Q(t_j)$ denotă asemănarea dintre videoclipul codificat $F(x_i)$ și textul codificat $Q(t_j)$.

Algoritmul TEACHTEXT începe prin a învăța mai multe modele profesor, care sunt antrenate pentru problema de regăsire T_k. Toate modelele au aceeași arhitectură, dar folosesc o caracteristică textuală diferită. O dată antrenat, calculăm o matrice de similaritate agregată $\Phi(S_1, \ldots, S_K) = \frac{1}{K}\sum_{k=1}^{K} S_k$ unde S_i este matricea de similaritate generată de profesorul $T_i, i = 1..K$. În cele din urmă, antrenăm modelul nostru student care, în plus față funcția de loss, $\mathscr{L}_r$ utilizează și o funție de loss de distilare $\mathscr{L}_d = \frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\left[l(\Phi(i,j), S_s(i,j))\right]$ unde l reprezintă Hubber-loss și este definită ca:

$$l(x,y) = \begin{cases} \frac{1}{2}(x-y)^2 & \text{if } |x-y| \leq 1, \\ |x-y| - \frac{1}{2} & altfel \end{cases} \tag{1}$$

pentru a imita matricea de similaritate agregată. În acest fel, ne atingem obiectivul de a crea un sistem de regăsire care să utilizeze mai multe încorporari de text.

Când vine vorba de arhitectura utilizată, testăm metoda noastră cu mai multe metode de ultimă generație [14, 13, 10] și propunem două noi variante ale CE [13]: CE+ care are unele îmbunătățiri tehnice cum ar fi utilizarea optimizatorului Adam și gpt2-xl ca caracteristică a textului și CE-L - o variantă ușoară care folosește w2v ca caracteristică a textului pentru a minimiza numărul de parametri.

5.2 Analiză experimentală

În Tabelul 3 facem o comparație extinsă a metodei noastre cu alte metode din literatură. Mai mult, prezentăm numărul de parametri ai fiecărei metode acolo unde este disponibil. După cum se poate vedea, algoritmul nostru TEACHTEXT aduce o îmbunătățire clară și numărul total de parametri rămâne același ca pentru arhitectura de bază. Mai mult, unele rezultate

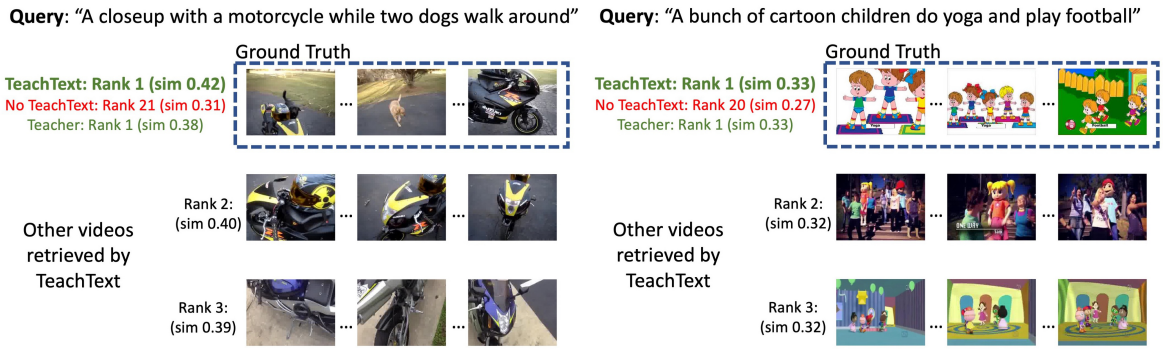


Figure 4: **Rezultate calitative.** Prezentăm primele 3 regăsiri video pentru fiecare interogare, date de metoda TEACHTEXT. Mai mult, arătăm rangul pentru profesor, precum și pentru student, fără a folosi TEACHTEXT cât și videoclipul corespunzător etichetei adevărate. Marcăm în verde cazurile în care preluarea este corectă în ceea ce privește R1 și cu roșu cazurile unde este incorectă. Pentru fiecare dintre cazurile prezentate, modelul învață de la profesor să-și corecteze predicția.

Model	Task	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$	Task	$R@1\uparrow$	$R@5\uparrow$	$R@10\uparrow$	$MdR\downarrow$	Params
Dual[8]	t2v	7.7	22.0	31.8	32.0	v2t	13.0	30.8	43.3	15.0	-
HGR[5]	t2v	9.2	26.2	36.5	24.0	v2t	15.0	36.7	48.8	11.0	-
MoEE[14]	t2v	$11.1_{\pm 0.1}$	$30.7_{\pm 0.1}$	$42.9_{\pm 0.1}$	$15.0_{\pm 0.0}$	v2t	$16.5_{\pm 0.1}$	$43.1_{\pm 0.5}$	$57.3_{\pm 0.6}$	$7.7_{\pm 0.5}$	400.41*M*
CE[13]	t2v	$11.0_{\pm 0.0}$	$30.8_{\pm 0.1}$	$43.3_{\pm 0.3}$	$15.0_{\pm 0.0}$	v2t	$17.0_{\pm 0.5}$	$43.5_{\pm 0.4}$	$57.8_{\pm 0.5}$	$7.2_{\pm 0.2}$	183.45*M*
TT-CE	t2v	$11.8_{\pm 0.1}$	$32.7_{\pm 0.1}$	$45.3_{\pm 0.1}$	$13.0_{\pm 0.0}$	v2t	$19.3_{\pm 0.4}$	$47.0_{\pm 0.7}$	$60.0_{\pm 0.4}$	$6.7_{\pm 0.5}$	183.45*M*
TT-CE-L	t2v	$13.0_{\pm 0.0}$	$34.6_{\pm 0.1}$	$47.3_{\pm 0.2}$	$12.0_{\pm 0.0}$	v2t	$22.4_{\pm 0.3}$	$50.4_{\pm 0.6}$	$63.8_{\pm 0.3}$	$5.3_{\pm 0.5}$	66.72*M*
TT-CE+	t2v	$\mathbf{15.0}_{\pm 0.1}$	$\mathbf{38.5}_{\pm 0.1}$	$\mathbf{51.7}_{\pm 0.1}$	$\mathbf{10.0}_{\pm 0.0}$	v2t	$\mathbf{25.3}_{\pm 0.1}$	$\mathbf{55.6}_{\pm 0.0}$	$\mathbf{68.6}_{\pm 0.4}$	$\mathbf{4.0}_{\pm 0.0}$	262.73*M*

Table 3: **MSR-VTT full split: Comparație cu alte modele.**

calitative pot fi văzute în Figura 4.

În acest capitol, am introdus o nouă metodă de regăsire text-video care obține rezultate de ultimă generație pe mai multe seturi de date. Cu toate acestea, poate cel mai interesant fapt care apare din această utilizare a mai multor caracteristici de text are implicații directe pentru sarcina de translatare video-text. În acest capitol, arătăm clar că, în ciuda cantității uriașe de antrenare prealabilă, există încă un decalaj de domeniu între viziune și limbaj. Folosind diferite încorporari de text pre-antrenate, putem reduce acest decalaj, cu toate acestea, implicațiile pentru sarcina de translatare rămân. Deoarece este nevoie de un pas suplimentar de generare, acesta se bazează direct pe adnotările din setul de date pe care este testat modelul. Având în vedere acest lucru, susținem că este nevoie de un nou tip de adnotări pe care le vom introduce în capitolul următor.

Capitolul 6

Setul de date Videos-to-Paragraphs

După cum am arătat în cele două capitole anterioare, prin utilizarea mai multor înglobări de text pre-antrenate, obținem un câștig mare în performanță pentru problema de regăsire text-video. Acest aspect este poate și mai interesant pentru sarcina de translatare video-text. Mai exact, arătăm că caracteristicile recente de text ([15, 19]) care sunt pre-antrenate pe cantități uriașe de date textuale, nu sunt suficient de puternici și încă există un decalaj de domeniu între viziune și limbaj. Deci, pentru a aborda această problemă, propunem un nou set de date care este adnotat la două niveluri semantice pe care îl vom descrie în continuare.

Așadar, vă prezentăm setul de date Videos-to-Paragraphs (cu adnotări corespunzătoare la două niveluri lingvistice), care conține videoclipuri filmate în interior într-un univers limitat. Scenele, actorii și activitățile lor sunt centrate în jurul oamenilor, în contextul a ceea ce se întâmplă de obicei într-o sală de clasă. În acest fel, susținem că putem studia mai bine cum funcționează un sistem de translatare video-text, dar, cel mai important, suntem capabili să descriem cu exactitate acțiunile umane prin limbaj natural. Deși se limitează la o configurare mai simplă, o sală de clasă, considerăm că acesta este un prim pas către un sistem automatizat care este capabil să descrie cu exactitate activitățile umane.

Setul de date constă din 510 videoclipuri capturate într-un mediu școlar interior. Videoclipuri au fost filmate în două săli de clasă diferite și pe hol. Fiecare video are aproximativ 30 de secunde iar videoclipurile au fost filmate cu două camere diferite: una fixă și una mobilă, concentrându-se pe actorii centrali. Protagoniștii desfășoară diverse activități,

Acest capitol se bazează pe metoda Bogolin, Croitoru and Leordeanu [4]

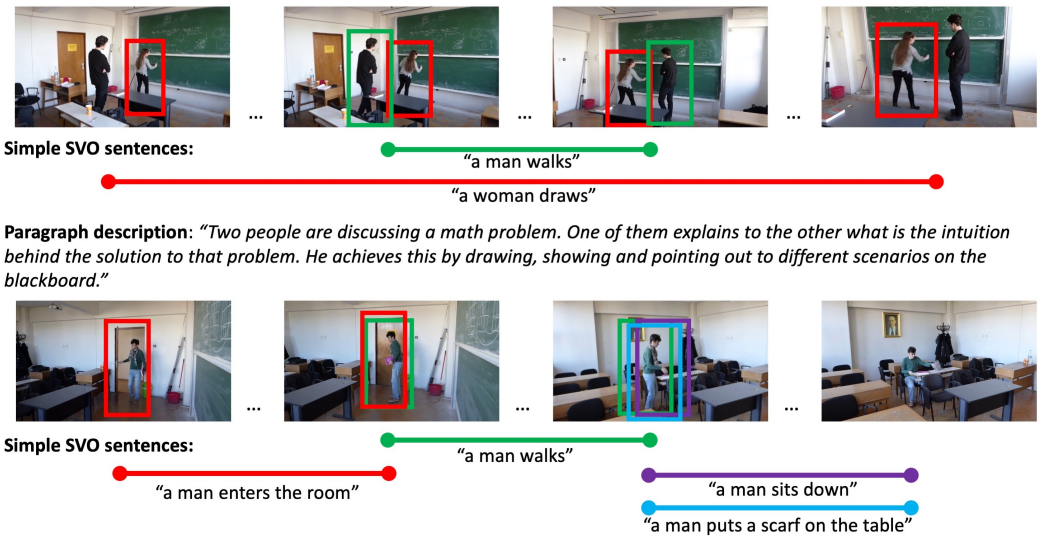


Figure 5: **Exemple din setul de date Video-to-Paragraphs.** Vă prezentăm adnotări pentru două videoclipuri diferite. Pentru fiecare, prezentăm câteva evenimente SVO (subiect, verb, obiect) împreună cu: cadrul lor de început, cadrul de sfârșit, căsuța de delimitare spațiu-timp corespunzătoare care conține evenimentul și descrierea propoziției sale simple. În partea de jos vă prezentăm descrierea video la nivel de paragraf. Paragrafele conțin propoziții mai lungi și mai complexe decât SVO-urile. Imagine refolosită din [4].

care implică de obicei interacțiunea cu diferite obiecte. Toate activitățile sunt centrate în jurul oamenilor și a interacțiunii acestora cu alți oameni și obiecte. Exemple de astfel de acțiuni sunt: intrarea/ieșirea dintr-o cameră, bea apă, așezat, ridicat, dat jos obiecte (de ex. pix, laptop, caiet, telefon mobil etc), vorbiți/îmbrățișați/scuturați mâinile cu/prezentați unei alte persoane, deschideți ușa/fereastra și altele, care de obicei au loc într-un mediu de clasă (Figura 5). Am conceput scenarii cât mai realiste posibil, astfel încât un anumit videoclip ar putea conține multe alte obiecte și acțiuni care au loc în același timp. Aceste acțiuni de bază, atomice, sunt cele descrise în propoziții simple (subiect, verb, obiect). Caracteristica principală care ne deosebește setul de date de altele din literatură este că am limitat toate scenele și scenariile la un ”univers” autonom, bine acoperit de actori, acțiuni și eveniment plauzibil, pentru a surprinde și a studia mai bine relația dintre viziune și limbaj într-un context specific. Fără acest set compact, autonom de videoclipuri, decalajul dintre viziune și limbaj este pur și simplu prea mare pentru a fi acoperit, iar sarcina este mai predispusă la overfitting - o observație care poate fi adesea făcută în literatura actuală.

Setul nostru de date este format din 510 videoclipuri (245 filmate cu o cameră fixă și 265

cu o cameră mobilă care urmărește scena), fiecare având aproximativ 30 de secunde. Din cele 510 videoclipuri, folosim 438 pentru antrenare, 20 pentru validare și 52 pentru testare. Am colectat 1048 de adnotări, astfel încât fiecare videoclip să aibă cel puțin 2 adnotări. În aceste 1048 adnotări avem 9036 SVO-uri adnotate. Astfel avem, în medie: 8,62 SVO-uri pe adnotare, cu 5,24 cuvinte pe SVO și 4,13 sec. acoperit de un SVO. Un SVO acoperă în medie, 14% din videoclip, în timp ce aproximativ 81,4% dintr-un videoclip este acoperit de toate SVO-urile adnotate pentru videoclipul respectiv. Multe SVO-uri (68% dintre ele) se suprapun cu altele temporar, deoarece diferite acțiuni pot avea loc simultan, uneori chiar făcute de aceeași persoană (de exemplu: a vorbi cu cineva, a sta jos și a pune ceva jos). O descriere la nivel de paragraf are în medie 3,66 propoziții (mai lungi decât propozițiile SVO de la nivelul inițial) și 40,03 cuvinte.

În acest capitol am descris modul în care am colectat și creat noul nostru set de date Videos-to-Paragraphs. Este primul pas necesar pentru a ne atinge obiectivul de a descrie activitățile umane prin limbajul natural. Mai mult, am introdus o nouă schemă de adnotare pe două niveluri despre care considerăm că stă la baza creării unui model de subtitrare ierarhic care este capabil să genereze propoziții mai lungi și mai complexe. Cu toate acestea, pentru a utiliza această nouă schemă de adnotare, este nevoie de o nouă abordare ierarhică pe care o vom introduce în capitolul următor.

Capitolul 7

Translatarea videourilor în text utilizând reprezentări ierarhice

După cum s-a subliniat în capitolul anterior, scopul nostru final este să studiem dacă o descriere lingvistică duală, intermediară, la nivelul propozițiilor scurte, care să descrie acțiuni simple sub formă de (subiect, verb, obiect) ar putea ajuta la generarea unor elemente mai complexe și limbaj fluent. Bunul simț și provocările întâlnite în cercetările actuale sugerează cu tărie că trecerea de la viziune la limbaj ar putea beneficia de o abordare mai graduală: ar trebui mai întâi să detectăm actorii și obiectele din scenă, apoi să înțelegem și să descriem acțiunile lor în propoziții scurte și numai după aceea, putem pune totul împreună într-o poveste mai mare, coerentă, descrisă într-un limbaj natural fluent. Rețineți că descrierea finală nu trebuie să conțină neapărat propozițiile simple inițiale.

Astfel, propunem o translatare video-text în două etape: în prima etapă descriem videoclipul ca o succesiune de evenimente, sub formă de propoziții simple (subiect, verb, obiect), care sunt și ele bine localizate în timp și spatiu. Cu un ușor abuz de terminologie, ne referim la astfel de propoziții simple care descriu acțiuni simple SVO (subiect, verb, obiect) ca propoziții SVO, chiar dacă uneori pot conține mai mult de trei cuvinte. Apoi, descrierea de nivelul doi rezumă conținutul video într-un paragraf coerent, care descrie conținutul vizual într-un mod mai elaborat, adăugând la primul nivel informații, relații cauzale și semantice dintre actori și evenimente (fără a fi nevoie să repetă propoziții inițiale SVO).

Este timpul să prezentăm sistemul nostru complet, așa cum este descris în Figura 6. Proce-

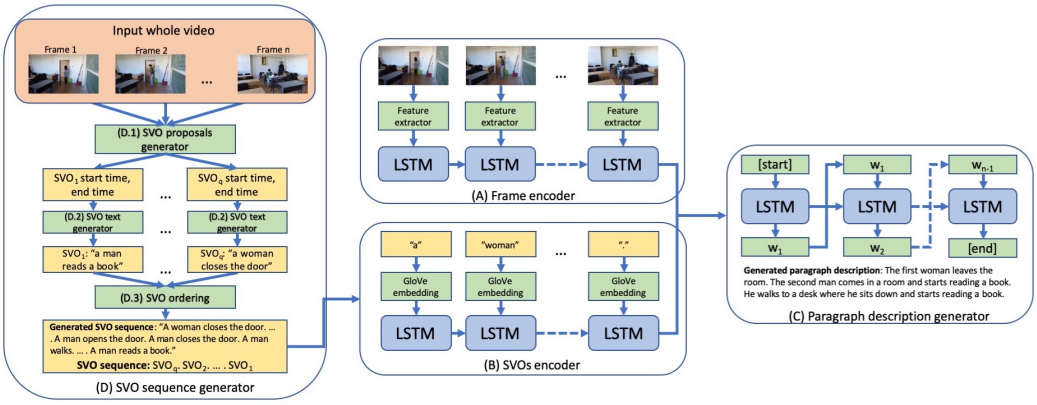


Figure 6: **Prezentare generală a sistemului.** Sistemul nostru constă din mai multe module care lucrează împreună pentru a forma două căi de codificare, una care procesează cadrele video (modulul *(A) Frame encoder*) și cealaltă care procesează SVO-urile (modulul *(B) SVOs encoder* și modulul *(D) SVOs sequence generator*). La final, un generator RNN (modulul *(C) Paragraph description generator*) emite descrierea finală. Pentru generarea unei secvențe SVO, determinăm mai întâi orele de început și de sfârșit ale potențialelor SVO (modulul *(D.1) SVO proposal generator*). Apoi, pentru fiecare fereastră de timp SVO generăm o descriere textuală (modulul *(D.2) SVO text generator*), apoi sortăm SVO-urile în ordine temporală (modulul *(D.3) SVO ordering*) și obțineți secvența SVO care este transmisă modulului *(B) SVOs encoder*. Imagine refolosită din [4].

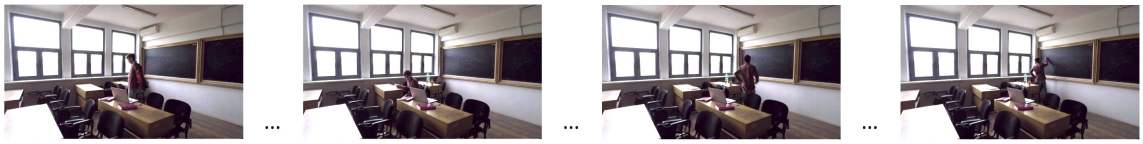
sul începe prin pre-extragerea caracteristicilor ImageNet [23] pentru toate videoclipurile din setul de date. Aceste caracteristici vor fi intrarea în *Frame encoder*. În mod similar, pentru partea de text, extragem caracteriticile GloVe [18]. Acestea vor fi introduse în codificatorul *SVOs*. Acum este timpul să antrenăm modulul *Paragraph description generator*. Acum, la inferneță, nu avem secvența SVO. Așadar, trebuie să o generăm. Așadar, antrenăm *SVO text generator*. Primește ca intrare caracteristicile video corespunzătoare fiecărui clip care are asociat un SVO. În continuare, generăm propunerile SVO și apoi folosim generatorul de text antrenat pentru a genera o propoziție pentru fiecare propunere. În cele din urmă, ordonăm SVO-urile prezise și apoi le putem folosi ca intrare pentru *Paragraph description generator* la inferență. În acest fel, ne atingem obiectivul de a genera descrieri de paragraf mai lungi.

7.1 Analiză experimentală

Validăm experimental beneficiile reprezentării textuale în două etape. Astfel, proiectăm experimente care evidențiază relevanța elementelor cheie ale metodei noastre și utilizează metricile limbajului standard din literatură: BLEU [16], ROUGE [12], CIDEr [24] și ME-

Method	B@1	B@2	B@3	B@4	M	R	C
Baseline	49.0	29.2	17.4	10.3	15.2	30.0	16.6
[9]	48.2	28.4	16.7	9.9	15.8	30.0	18.3
PG-SW-Pred-SVO	**51.2**	**33.6**	**22.2**	14.4	17.3	33.8	22.9
PG-NMS-Pred-SVO	49.3	32.5	21.7	**14.5**	**19.7**	**37.3**	**26.2**

Table 4: **Comparație cu metodele de ultimă generație.** Folosim mai multe metrici de evaluare (BLEU@N (B@N), ROUGE (R), METEOR (M) și CIDEr (C)). Metodele comparate au fost antrenate de la zero pe setul de antrenament al bazei noastre de date. La momentul testării, toate metodele au acces numai la video-ul de intrare. Abordarea noastră conduce cu o marjă semnificativă, în special față de cele mai recente valori. Atât PG-SW-Pred-SVO, cât și PG-NMS-Pred-SVO profită de adnotarea schemei cu două niveluri, diferența dintre ele este că PG-SW-Pred-SVO utilizează o modalitate mai naivă de a detecta regiunile SVO - fereastra glisantă , în timp ce PG-NMS-Pred-SVO folosește o rețea de încredere mai sofisticată în combinație cu superiune nonmaxima - modulul D.1



GT: *"A man sits down on a chair in order to read a book. After a while he looks at the blackboard, sits up and takes a chalk. He is probably going to write something on the blackboard."*
Seq-GT-SVO: *"A man takes off his jacket. A man puts his jacket on the chair. A man sits down. A man reads a book. A man stands up. A man writes on the blackboard."*
PG-GT-SVO: *"A man takes off his jacket and puts the jacket on the table. Then he puts down the book and starts reading and browsing it."*
(Duta et al. 2018): *"A man closes the window and leaves the room then closes the window and sits down."*
PG-NMS-Pred-SVO: *"A man takes off his backpack and puts it on the desk while sitting down and starts reading. After a while he sits up and starts writing something on the blackboard."*

Figure 7: **Exemple calitative.** Observați diferențele de calitate dintre descrierile la nivelul secvențelor SVO și cele de la nivelul paragrafelor. Paragrafele generate sunt mai concise și mai coerente.

TEOR [3].

Chiar dacă introducem o nouă schemă de adnotare, este esențial să vedem cum ne situăm comparativ cu alte metode publicate. Deci, ne comparăm cu un model de bază puternic (sistemul S2VT [25]) și cu o metodă recentă [9] (vezi Tabel 4). Vă rugăm să rețineți că, pentru o comparație corectă, ne comparăm cu un model individual, care utilizează o arhitectură secvență-la-secvență și un mecanism de atenție. Mai mult, în Figura 7 puteți găsi câteva exemple calitative. După cum se poate observa, paragraful generat descrie cu acuratețe ceea ce se întâmplă în scenă.

În acest fel, introducem un model ierarhic nou, care este capabil să profite de avantajul oferit de schema de adnotare pe două niveluri introdusă cu noul nostru set de date Videos-to-

Paragraph. Începem prin a genera propoziții simple ordonate în timp, urmate de generarea, într-o a doua etapă, a unei descrieri mai lungi și mai complexe, care relaționează semantic evenimentele mai simple în spațiu și timp folosind un limbaj mai coerent.

Aceasta este prima abordare de acest fel din literatură, cu două etape explicabile pentru translatarea video în text. În acest fel, suntem capabili să descriem activități umane prin limbajul natural.

Capitolul 8

Concluzii

În această lucrare, ne propunem să descriem activități complexe prin limbaj naturale. Susținem că translatarea video în text este cea mai bună modalitate de a descrie astfel de activități, deoarece limbajul natural este atât de versatil încât se poate surprinde și descrie îndeaproape ceea ce se întâmplă fără a fi limitat de un set fix de etichete. Așadar, prezentăm mai întâi câteva abordări arhitecturale diferite pentru a codifica conținutul unui videoclip și introducem o metodă bazată pe consens pentru a îmbunătății rezultatele. La o examinare atentă, am constatat că modificări foarte mici în structura propoziției sau la nivelul cuvintelor, fără a modifica sensul general, pot modifica puternic scorurile. Mai mult, pentru a înțelege mai bine relația dintre video și text, facem o analiză amănunțită pentru o problemă diferită, dar strâns legată: regăsirea text-video. Din această analiză putem trage următoarea concluzie: există încă un decalaj de domeniu între datele vizuale și text. În cele din urmă, am introdus o abordare nouă, ierarhică, a generării de limbaj din videoclipuri care utilizează o schemă de adnotare pe două niveluri. Se începe cu generarea de propoziții simple ordonate în timp, urmată de generarea, într-o a doua etapă, a unei descrieri mai lungi și mai complexe. Aceasta este prima abordare de acest fel din literatură, cu două etape explicabile pentru generarea textului din video. Experimentele noastre extinse sugerează cu tărie că ideea generării limbajului explicit în mai multe faze, trecând de la propoziții mai simple, la paragrafe mai lungi și apoi la povești complexe, ar putea oferi o direcție interesantă.

Bibliografie

[1] Bahdanau, D., Cho, K. and Bengio, Y. [2014], 'Neural machine translation by jointly learning to align and translate', *arXiv preprint arXiv:1409.0473* .

[2] Bai, S., Kolter, J. Z. and Koltun, V. [2018], 'An empirical evaluation of generic convolutional and recurrent networks for sequence modeling', *arXiv preprint arXiv:1803.01271* .

[3] Banerjee, S. and Lavie, A. [2005], Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, *in* 'Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization', Vol. 29.

[4] 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)'.

[5] 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)'.

[6] 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', *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* .

[7] Cui, Y., Yang, G., Veit, A., Huang, X. and Belongie, S. [2018], Learning to evaluate image captioning, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[8] 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)'.

[9] Duta, I., Nicolicioiu, A. L., Bogolin, S.-V. and Leordeanu, M. [2018], 'Mining for meaning: from vision to language through multiple networks consensus', *The British Machine Vision Conference (BMVC)* p. 275.

[10] 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)* .

[11] Jin, Q., Chen, S., Chen, J. and Hauptmann, A. [Proceedings of the International Conference on Multimedia], 'Knowing yourself: Improving video caption via in-depth recap'.

[12] Lin, C.-Y. [2004], Rouge: A package for automatic evaluation of summaries, *in* 'Proceedings of ACL Workshop on Text Summarization Branches Out', p. 10.

[13] 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)* .

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

[15] Mikolov, T., Chen, K., Corrado, G. and Dean, J. [2013], 'Efficient estimation of word representations in vector space', *arXiv preprint arXiv:1301.3781* .

[16] Papineni, K., Roukos, S., Ward, T. and Zhu, W.-J. [2002], Bleu: a method for automatic evaluation of machine translation, *in* 'Proceedings of the Association for Computational Linguistics (ACL)'.

[17] Pasunuru, R. and Bansal, M. [2017], Reinforced video captioning with entailment rewards, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, pp. 979–985.

[18] Pennington, J., Socher, R. and Manning, C. [2014], Glove: Global vectors for word representation, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)'.

[19] 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* .

[20] Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T. and Saenko, K. [2018], 'Object hallucination in image captioning', *arXiv preprint arXiv:1809.02156* .

[21] Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A. and Schiele, B. [2017], 'Movie description', *International Journal of Computer Vision (IJCV)* **123**(1), 94–120.

[22] Shen, Z., Li, J., Su, Z., Li, M., Chen, Y., Jiang, Y.-G. and Xue, X. [2017], Weakly supervised dense video captioning, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[23] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. [2016], Rethinking the inception architecture for computer vision, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[24] Vedantam, R., Lawrence Zitnick, C. and Parikh, D. [2015], Cider: Consensus-based image description evaluation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)'.

[25] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K. [2015], Sequence to sequence-video to text, *in* 'Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)'.

[26] Wang, X., Chen, W., Wu, J., Wang, Y.-F. and Wang, W. Y. [2018], 'Video captioning via hierarchical reinforcement learning', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* .