Raport științific și tehnic

Denumire proiect (EN)	CULTiVA: Curriculum learning in text mining and visual question answering
Denumire proiect (RO)	CULTiVA: Învățarea pe bază de curriculum în text mining și visual question answering
Acronim	CULTiVA
Cod proiect	PN-III-P1-1.1-TE-2019-0235
Număr contract	72/2020
Contractor	UNIVERSITATEA BUCUREȘTI
Tip proiect	Proiecte de cercetare pentru stimularea tinerelor echipe independente (TE)
Autoritatea contractantă	Unitatea Executivă pentru Finanțarea Învățământului Superior, a Cercetării, Dezvoltării și Inovării
Perioada de raportare	07.09.2020 - 31.12.2020
Etapa de execuție	1/2020
Director de proiect	Radu Tudor Ionescu
Adresă web	https://cultiva-proj.herokuapp.com

1. Rezumat:

În Etapa 1 – „Dezvoltarea și evaluarea unui predictor de dificultate a textelor (obiectiv 1)” am efectuat activitățile 1.1 și 1.2. Activitatea 1.1. are ca scop îndeplinirea primului obiectiv, acela de a concepe un model neuronal pentru estimarea dificultății textelor. Activitatea 1.2. are ca scop crearea și publicarea site-ului web asociat proiectului. Informațiile vor fi actualizate pe site pe parcursul desfășurării proiectului, astfel că, la momentul actual, conținutul publicat nu este complet. Pe lângă cele două activități propuse, am demarat lucrul în vederea activităților 2.1 și 2.3, ce au ca scop îndeplinirea obiectivelor 2 respectiv 4.

2. Procentaj îndeplinire obiective:

Obiectivul 1 – 100%
Obiectivul 2 – 20%
Obiectivul 3 – 0%
Obiectivul 4 – 20%
Obiectivul 5 – 20%

3. Descriere științifică și tehnică:

În conformitate cu activitățile prevăzute în Etapa 1 de raportare din Anexa II a contractului TE72/2020, am efectuat următoarele:

- **Activitatea 1.1. Task 1 – Dezvoltarea si evaluarea unui predictor de dificultate a textelor**

În vederea îndeplinirii primului nostru obiectiv, am decis să folosim un model neuronal state-of-the-art pentru estimarea dificultății textelor, anume un model tip BERT [Devlin et al., NAACL19]. Pentru a ne familiariza cu modelul BERT, am efectuat întâi experimente antrenând acest model pe problema geo-localizării textelor, aceasta fiind o problemă de

regresie asemănătoare cu cea de estimare a complexității textelor. În urma experimentelor efectuate, am ajuns la concluzia că rezultatele modelului BERT sunt mai slabe în comparație cu modele precum metode nucleu pentru șiruri de caractere. Aceste rezultate au fost publicate în lucrarea [Găman et al., VarDial20]. Totuși, am decis continuarea studiului în direcția estimării dificultății textelor pe baza modelului BERT, aceasta fiind la rândul său tot o problemă de regresie. De această dată, rezultate obținute sunt apropiate de cele raportate în lucrarea [Butnaru et al., BEA18], conform Tabelului 1 de mai jos. Avantajul modelului BERT este extinderea simplă și directă de la estimarea complexității cuvintelor la estimarea complexității propozițiilor sau textelor.

Metodă	Acuratețe	Scor F1
[Butnaru et al., BEA18]	86.78%	85.94%
BERT	86.55%	82.96%

Tabel 1. Rezultatele modelului BERT în comparație cu rezultatele modelului propus în lucrarea [Butnaru et al., BEA18] pe problema estimării complexității la nivel de cuvânt pe setul de date English News.

- **Activitatea 1.2. Task 5 – Dezvoltarea unui website public pentru prezentarea proiectului si publicarea modelelor / codului aferent.**

Această activitate corespunde cu obiectivul 5 din cadrul propunerii de proiect, acela de a publica rezultatele obținute pe site-ul proiectului. În vederea îndeplinirii acestui obiectiv, am dezvoltat și publicat site-ul web al proiectului la: **https://cultiva-proj.herokuapp.com**. Conținutul publicat include un rezumat precum și informații despre membrii implicați în proiect și articolele publicate până în momentul prezent.

- **Activități premergătoare Activităților 2.1. și 2.3.**

În plus față de cele prevăzute în activitățile 1.1. și 1.2., am studiat posibilitatea aplicării învățării pe bază de curriculum pentru a antrena modele neuronale pentru problema detectării obiectelor din imagini. Rezultatele obținute demonstrează ca învățarea pe bază de curriculum conduce către performanțe mai bune. Metoda propusă și rezultatele aferente au fost sintetizate într-o lucrare trimisă spre recenzat [Soviany et al., CVIU20]. De menționat că în vederea dezvoltării unui model pentru visual question answering, una din etapele importante este tocmai detectarea obiectelor în imagini. Astfel, studiul efectuat este relevant în îndeplinirea obiectivului 4, fiind un pas premergător activității 2.3.

Pe lângă cele enumerate mai sus, am studiat problema copierii unor modele state-of-the-art de tip black-box, folosind o tehnică de generare de imagini bazată pe calcul evoluționar. Replicarea rezultatelor unor modele din literatura recentă este obiectului 2 al proiectului nostru. Totuși, în anumit cazuri, aceste modele sunt expuse printr-un API ce permite doar propagarea unor exemple și aflarea predicțiilor corespunzătoare. În acest context, tehnica propusă este de interes, având o legătură importantă cu activitatea 2.1. Metoda propusă și rezultatele aferente au fost sintetizate într-o lucrare acceptată spre publicare [Bărbălău et al., NeurIPS20].

4. **Diseminarea rezultatelor în articole științifice:**

În urma activităților de cercetare fundamentală efectuate, au rezultat 2 articole publicate în volume ale unor conferințe internaționale, dintre care 1 articol într-o conferință de categoria A* (NeurIPS 2020) și 1 articol într-un workshop de categoria B (VarDial 2020). Astfel, au fost îndeplinite cerințele minimale de diseminare a rezultatelor pe anul 2020, care prevedeau

publicarea cel puțin a unui articol într-o conferință sau workshop de categoria B. Articolele finanțate prin proiectului de cercetare sunt listate în continuare:

1. Antonio Bărbălău, Adrian Cosma, Radu Tudor Ionescu, Marius Popescu. Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. In Proceedings of NeurIPS, 2020. **(Rank A* Conference)**

2. Mihaela Găman, Radu Tudor Ionescu. Combining Deep Learning and String Kernels for the Localization of Swiss German Tweets. In Proceedings of VarDial Workshop of COLING, 2020. **(Rank B Workshop)**

5. Referințe bibliografice:

[Bărbălău et al., NeurIPS20] A. Bărbălău, A. Cosma, R.T. Ionescu, M. Popescu. Black-Box Ripper: Copying black-box models using generative evolutionary algorithms. Proceedings of NeurIPS, 2020.

[Butnaru et al., BEA18] A. Butnaru, R.T. Ionescu. UnibucKernel: A kernel-based learning method for complex word identification. Proceedings of BEA-13, pp. 175–183, 2018.

[Devlin et al., NAACL19] J. Devlin, M.W. Chang, K. Lee, K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL, pp. 4171–4186, 2019.

[Găman et al., VarDial20] M. Găman, R.T. Ionescu. Combining Deep Learning and String Kernels for the Localization of Swiss German Tweets. Proceedings of VarDial, 2020.

[Soviany et al., CVIU20] P. Soviany, R.T. Ionescu, P. Rota, N. Sebe. Curriculum Self-Paced Learning for Cross-Domain Object Detection. Computer Vision and Image Understanding, 2020 (submitted).

Data,
26.11.2020

Director proiect,
Radu Tudor Ionescu