DNN Mikroservisleri ile Makine Çevirisi Modelleri için Performans Analizi

Simla Burcu Harma[1], Mario Drumond[2], Oğuz Ergin[1] Babak Falsafi[2]

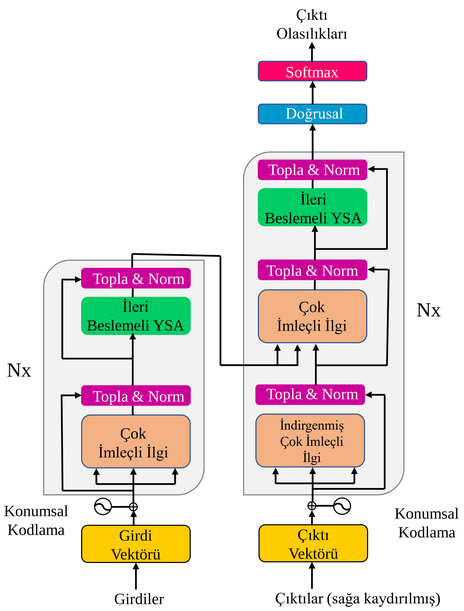
[1]TOBB Ekonomi ve Teknoloji Üniversitesi, [2]École polytechnique fédérale de Lausanne

[1]s.harma@etu.edu.tr, oergin@etu.edu.tr, [2]mario.drumond@epfl.ch, babak.falsafi@epfl.ch

Özetçe

- Derin Yapay Sinir Ağlarının (DYSA) mikroservis olarak sunulmasıyla birlikte, sunucuların DYSA gecikme kısıtlamalarını sağlamaları önemli hale gelmistir. Makine çevirisi yapan DYSA mikroservislerinde hız darboğazı, DYSA modelinin koşu süresidir. Vaswani vd. (2017)'nin önerdiği "Transformer" modeli literatürde en çok çalışılan, en gelişmiş modellerden biridir.
- Bu bildiride, makine çevirilerinin hızlandırılmasında yol gösterici olması için Transformer modelinin detaylı zaman analizi yapılmış, makine çevirisinde önemli bir yöntem olan ışın araması ayrıntılı bir şekilde incelenmiştir. Ayrıca, ışın aramasında kelime-hazinesi büyüklüğünü azaltmanın başarımı büyük ölçüde artırmasına rağmen çeviri kalitesini düşürdüğü gözlemlenmiş ve sebepleri tartışılmıştır.

Giriş



Şekil 1: "Transformer" modeli mimarisi

Metodoloji

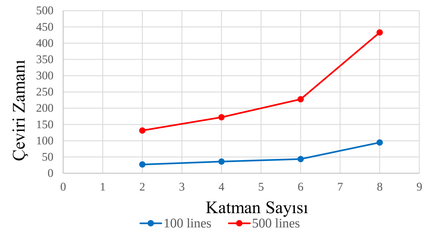
- İngilizce-Almanca çevirisi
- WMT'17 test verisi (newstest2017) (3004 cümle)
- Intel(R) Xeon(R) CPU E5-2680 v3 @ 2.50GHz
- NVIDIA GeForce GTX TITAN Xp GPU.

N	GPU'da 100 satır	GPU'da 500 satır	Param. ($\times 10^6$)	Model boyutu(KB)
2	8.105 s	43.207 s	36	137.330
4	8.590 s	46.210 s	50	190.735
6	8.710 s	55.451 s	65	247.955
8	14.156 s	72.894 s	80	305.176

Tablo 1: Katman sayısı ile model büyüklüğündeki değişim ve GPU'daki çalışma zamanları

Deneyler & Değerlendirme

- Farklı yapılandırma değerleri ile "Transformer"
- "Transformer"ın detaylı performans analizi
- Işın araması, ışın büyüklüğü ve performans/çeviri kalitesi



Şekil 2: Çeviri Zamanı ve Katman Sayısı

		N	d_{model}	d_{ff}	h	d_k	d_v	BLEU	100 satır	500 satır
Standart		6	512	2048	8	64	64	25.8	43.816 s	227.775 s
"İlgi ayracı" sayısındaki değişim					1	512	512	24.9	27.371 s	146.750 s
					4	128	128	25.5	52.202 s	272.670 s
					16	32	32	25.8	60.981 s	4315.091 s
					32	16	16	25.4	73.278 s	408.541 s
Katman sayısındaki değişim		2						23.7	26.857 s	131.478 s
		4						25.3	35.845 s	172.184 s
		8						25.5	94.582 s	433.037 s
Model boyutundaki değişim			256			32	32	24.5	37.398 s	248.761 s
			1024			128	128	26	96.560 s	667.856 s
				1024				25.4	60.178 s	412.966 s
				4096				26.2	64.702 s	564.027 s

Tablo 2: "Transformer"ın farklı yapılandırma değerleri ile 100 ve 500 satırlık girdiler için BLEU Skorları ve çeviri zamanları

	Katman	CPU	CPU oranları	GPU	GPU oranları
Kodlayıcı	Çok İmleçli İlgi	7 s	0.56%	1.8 s	0.33%
	Normalizasyon	1.1 s	0.09%	0.2 s	0.04%
	İleri Beslemeli YSA	7.5 s	0.60%	0.8 s	0.15%
Kodçözücü	İndirgenmiş Çok İmleçli İlgi	358 s	28.70%	87.3 s	16.22%
	Normalizasyon	23.8 s	1.91%	11.9 s	2.21%
	Çok İmleçli İlgi	242.9 s	19.47%	63.3 s	11.76%
	Normalization	22.8 s	1.83%	12.1 s	2.25%
	İleri Beslemeli YSA	102.3 s	8.20%	48 s	8.92%
Generator	Generator	119.3 s	9.56%	2 s	0.37%
Işın Araması	Işın Araması	276 s	22.12%	238 s	44.22%
TOPLAM		1247.6 s		538.19 s	

Tablo 3: 3004 satırlık girdi için CPU ve GPU'da "Transformer"ın her katmanı için detaylı zaman analizi

Işın	Çeviri Zamanı	Kelime-seviyesinde BLEU	Ayraç-seviyesinde BLEU
2	665.51 s	27.58	33.77
3	856.17 s	27.88	33.93
4	1087.96 s	27.99	33.99
5	1247.60 s	28.09	33.97
10	2092.67 s	28.07	33.69
15	2931.36 s	27.83	33.37

Tablo 4: 3004 satırlık girdi için CPU'da ışın büyüklüğü ile çeviri zamanı ve BLEU Skorları değişimi

Sözde kod	*A*	*T*
Eğer cümle başı ise:		
ışın_skorları = ayraç olasılıkları	0.01 s	0.01 s
Değilse:		
Mevcut ayraç olasılıklarını, o yoldaki ayraçların olasılıkları toplamına eşit olan ışın_skorlarına ekle	4.1 s	11.3 s
Eğer <CS> (cümle sonu) ayracı gelirse, yolu burada durdur ve çocuğu olmasına izin verme	11.4 s	13.3 s
En yüksek olasılıklı k tane ayracı bul	15.8 s	184.9 s
Kelime hazinesinde bu ayraçların indislerini bul ve kaydet	13.9 s	3.4 s
Tutulan ışın araması bilgilerini güncelle	15.5 s	17.4 s
Eğer en yüksek olasılıklı k ayraçtan biri <CS> ise bu yolu bitmiş yollara ekle	12.3 s	12.8 s
TOPLAM IŞIN ARAMASI ZAMANI	82.73 s	252.78 s
BLEU SKORU	17.42	28.09

Tablo 5: 3004 satırlık girdi için CPU'da ışın araması zaman analizi. A: Alt-kelime-hazinesi, T: Tüm-kelime-hazinesi

Sonuç

Bu çalışmada, öncelikle darboğazları belirlemek için "Transformer" modelinin etrafına bir mikroservis kurduk. Küçük girdi boyutundan dolayı ağ iletişiminin darboğaz olmadığını, bu nedenle modeli hızlandırmamız gerektiğini gözlemledik. Modele etki eden bileşenleri anlayabilmek için "Transformer"ı farklı yapılandırma değerleri ile çalıştırdık ve performansın küçük değişimlerle bile kolayca etkilendiğini gördük. Modelin davranışını daha iyi anlamak için "Transformer"ın detaylı zaman analizini çıkardık. Modelin darboğazının ışın araması olduğunu gözlemledik ve bu aşamanın her bir adımı için zaman gereksinimlerini ölçtük. Ayrıca çeviri kalitesini gösteren BLEU skorunun ayraç-seviyesinde bakıldığında ışın büyüklüğü değişimine hassasiyet göstermediğini gördük. Son olarak, küçük bir kelime hazinesi kullanıldığında, 3 kat hızlanma elde edilebileceğini ancak bu alt-kelime-hazinesinin seçiminin çeviri kalitesini etkilediğini gözlemledik.