

DNN Mikroservisleri ile Makine Çevirisi Modelleri için Performans Analizi

Simla Burcu Harma¹, Mario Drumond², Oğuz Ergin¹, Babak Falsafi²

¹*TOBB Ekonomi ve Teknoloji Üniversitesi*

²*École polytechnique fédérale de Lausanne*



TOBB ETÜ
Ekonomi ve Teknoloji Üniversitesi

EPFL

PAISA
PARALLEL SYSTEMS
ARCHITECTURE LAB

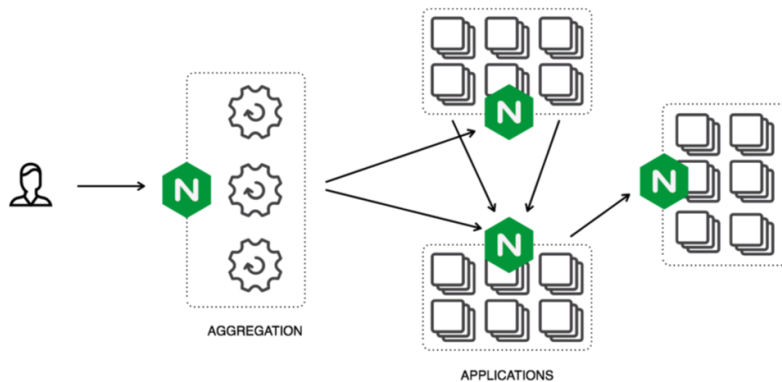

kasirga

Makine Çevirisi ve Mikroservisler

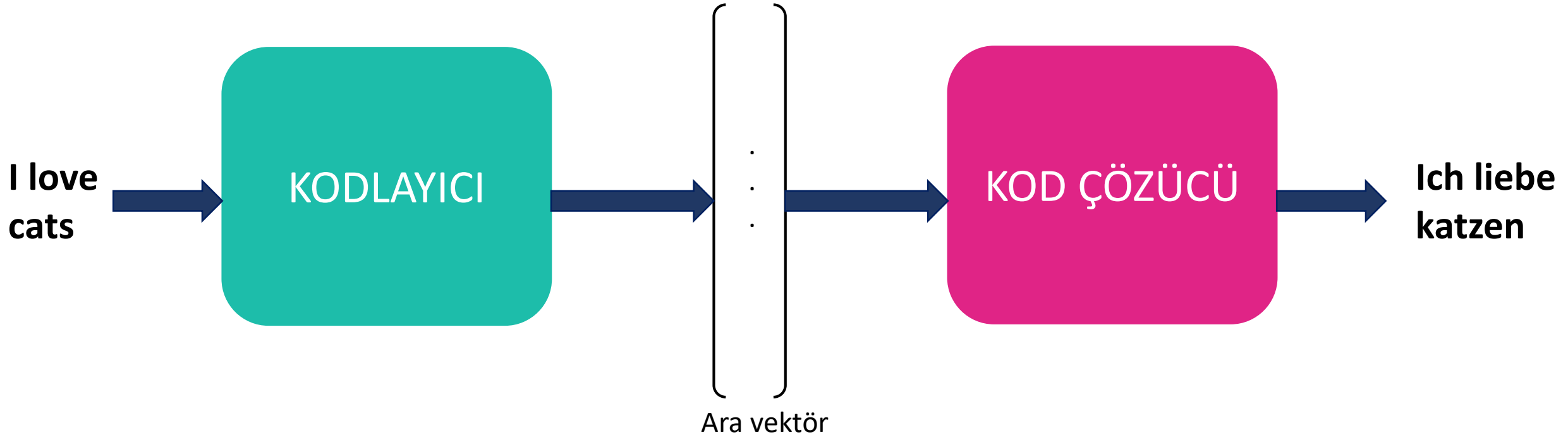
- Google Translate, Facebook, Instagram



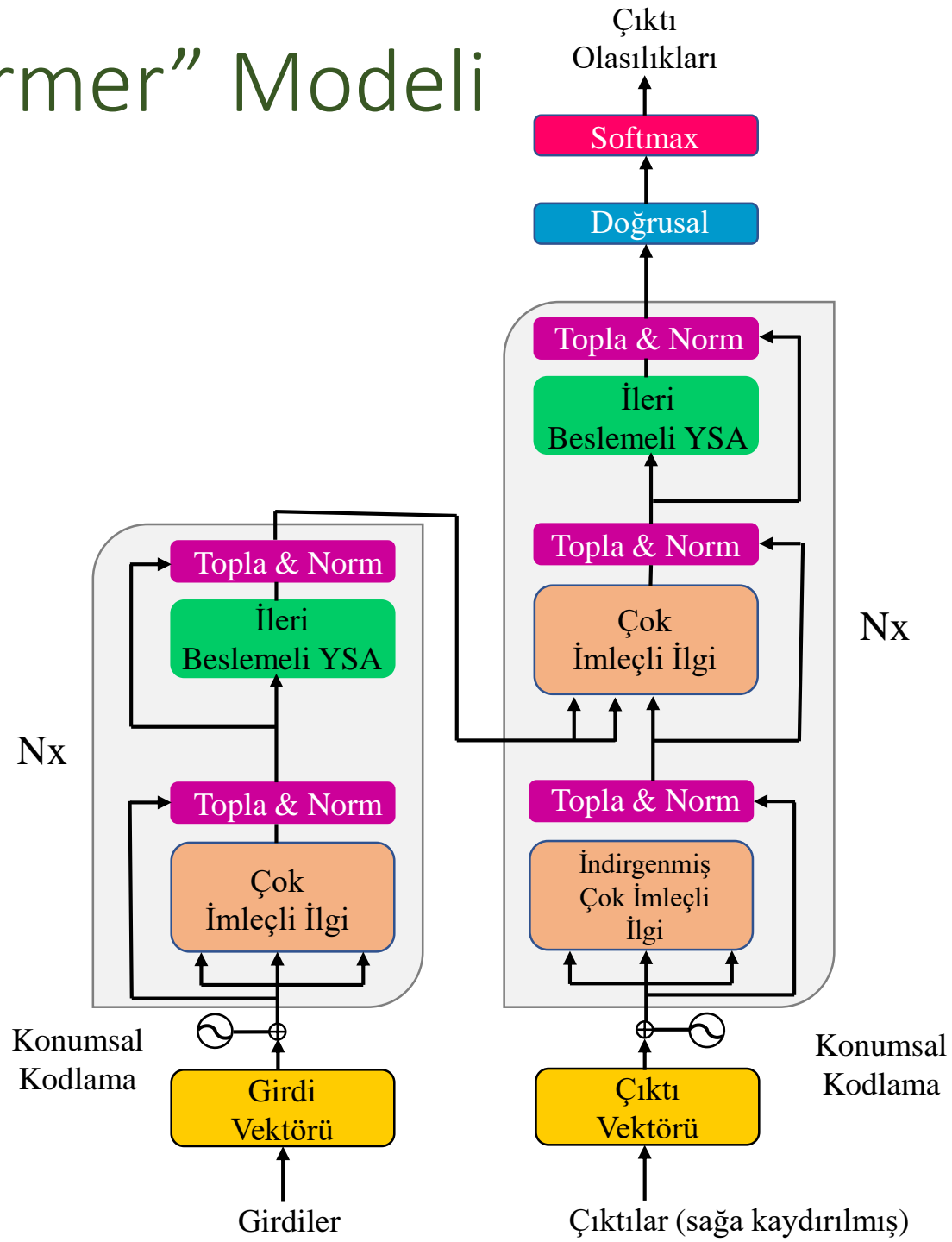
- Mikroservisler ve hızlandırıcılar



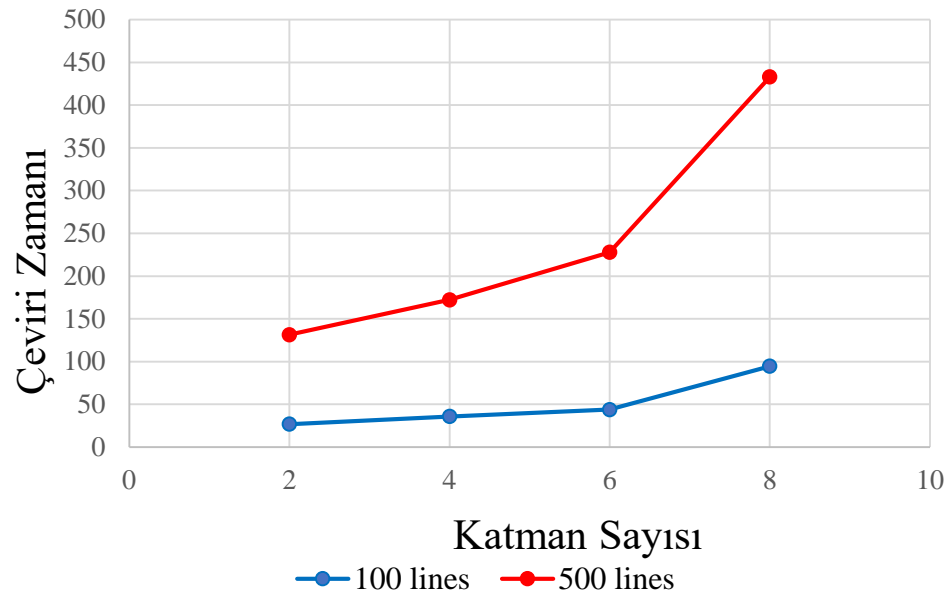
Diziden-diziye (Sequence-to-sequence) Modeller



“Transformer” Modeli



Katman Sayısının Çeviriye Etkisi



- 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.

Önbellek Boyutları:

L1: 32K

L2: 256K

L3: 30720K

N	BLEU	CPU'da 100 satır	CPU'da 500 satır	GPU'da 100 satır	GPU'da 500 satır	Parametre sayısı ($\times 10^6$)	Model boyutu (KB)
2	23,7	26,857 s	131,478 s	8.105 s	43.207 s	36	137.330
4	25,3	35,845 s	172,184 s	8.590 s	46.210 s	50	190.735
6	25,8	43,816 s	227,775 s	8.710 s	55.451 s	65	247.955
8	25,5	94,582 s	433,037 s	14.156 s	72.894 s	80	305.176

Modelin Detaylı Zaman Analizi

	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
	Normalizasyon	22,8 s	%1.83	12,1 s	%2.25
	İleri Beslemeli YSA	102,3 s	%8.20	48 s	%8.92
Üretici	Üretici	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	

Işın Aramasının Detaylı Analizi

	Alt-kelime- hazinesi	Tüm-kelime- hazinesi
Eğer cümle başı ise: ışın_skorları = token olasılıkları	0,01 s	0,01 s
Değilse: Mevcut token olasılıklarını, o yoldaki tokenların olasılıkları toplamına eşit olan ışın_skorlarına ekle	4,1 s	11,3 s
Eğer <CS> (cümle sonu) token'ı gelirse, yolu burada durdur ve çocuğu olmasına izin verme	11,4 s	13,3 s
En yüksek olasılıklı k tane token'ı bul	15,8 s	184,9 s
Kelime hazinesinde bu token'ların 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 token'dan biri <CS> ise bu yolu bitmiş yollara ekle	12,3 s	12,8 s
TOPLAM IŞIN ARAMASI ZAMANI	82,7275 s	252,779 s
BLEU SKORU	17,42	28,09

TEŞEKKÜRLER



TOBB ETÜ
Ekonomi ve Teknoloji Üniversitesi

EPFL

PAISA
PARALLEL SYSTEMS
ARCHITECTURE LAB


kasirga

1		N	d_{model}	d_{ff}	h	d_k	d_v	BLEU	100 lines	500 lines
2	Standart	6	512	2048	8	64	64	25,8	43,816 s	227,775 s
3	“Attention head” sayısındaki değişim				1	512	512	24,9	27,371 s	146,750 s
4					4	128	128	25,5	52,202 s	272,670 s
5					16	32	32	25,8	60,981 s	315,091 s
6					32	16	16	25,4	73,278 s	408,541 s
7	Katman sayısındaki değişim	2						23,7	26,857 s	131,478 s
8		4						25,3	35,845 s	172,184 s
9		8						25,5	94,582 s	433,037 s
10	Model büyüklüğündeki değişim		256			32	32	24,5	37,398 s	248,761 s
11			1024			128	128	26	96,560 s	667,856 s
12				1024				25,4	60,178 s	412,966 s
13					4096				26,2	64,702 s

Işın	Çeviri Zamanı	Kelime-seviyesinde BLEU	token-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