PaliGemma 2 Görsel Dil Modeli Değerlendirmesi: Boyut mu yoksa çözünürlük mü?

Doruk Özgen
Freya Research Fellows

Google, 14 Mayıs 2024 tarihinde açık ağırlıklı (open-weight) olan PaliGemma görsel dil modelini yayınladı. 5 Aralık 2024 tarihinde ise "PaliGemma 2: A Family of Versatile VLMs for Transfer" makalesiyle PaliGemma ile aynı mimariyi kullanan PaliGemma 2 görsel dil modelini yayınladı.

PaliGemma ve PaliGemma 2 benzer bir multimodal mimari yaklaşımını paylaşmaktadır. Her iki model ailesinde de SigLIP-So400m görsel encoder'ı kullanılmaktadır. Ancak PaliGemma, Gemma 1 2B dil modelini kullanırken PaliGemma 2, Gemma 2 ailesindeki 2B, 9B ve 27B dil modelleriyle sunulmaktadır.

Özetle PaliGemma 2 görsel dil modeli, orijinal modelin Gemma 2 ailesiyle entegre edilerek ölçeklendirilmiş ve kapasitesi artırılmış halidir. Ayrıca PaliGemma 2, birçok benchmark'ta PaliGemma'ya kıyasla daha yüksek performans göstermektedir. Bununla birlikte iyileşme her görevde gözlenmemektedir.

PaliGemma 2 görüntü dil modeli tam olarak nedir ve hangi görevlerde kullanılır?

PaliGemma 2 görsel dil modeli; görsel (resim) ve metinsel (yazı) girdileri beraber işleyerek metin tabanlı çıktılar üreten bir yapay zeka modelidir. Temel olarak Görsel Soru Cevaplama (VQA), Görsel Betimleme/Açıklama (Image Captioning) ve Nesne Tespiti (Object Detection) gibi karmaşık görevlerde kullanılmak üzere tasarlanmıştır.



Modelin görsel ve metin girdilerini birlikte işleyerek anlamlı bir metinsel çıktı üretebildiğini doğrulamak amacıyla örnek bir inference gerçekleştirilmiştir. Bu deneyde, modele yukarıdaki görsel ile birlikte *“Bu resimde tam olarak ne görüyorsun? Detaylı bir şekilde açıkla.”* metin girdisi sağlanmıştır. Model, görseldeki iki kediyi ve çevresindeki ip yığınını doğru biçimde tanımlayarak *“İki kedi bir yığın ipte yatıyor.”* çıktısını üretmiştir.

Bu metin girdisinde detaylı bir açıklama istenmesine rağmen PaliGemma 2 görsel dil modeli bizlere kısa bir açıklama üretmiştir. Bu durum, PaliGemma 2'nin sohbet dil modeli yerine kısa görev istemleriyle eğitilmiş bir model olmasından kaynaklanmaktadır.

Kullanılan inference ortamı ve çalıştırma adımları Google Colab notebookunda sunulmuştur.

Peki "PaliGemma 2: A Family of Versatile VLMs for Transfer" makalesi PaliGemma 2 dil modeli hakkında tam olarak ne anlatmaktadır?

Makale, bizlere PaliGemma 2 görsel dil modelinde Gemma 2 dil modelinin farklı boyutlarının kullanılmasının ve modeli eğitirken farklı çözünürlüklerin kullanılmasının nasıl avantajlar ve dezavantajlar sağladığını açıklamaktadır. Makaledeki 3B, 10B ve 28B ifadeleri PaliGemma 2'nin toplam parametre boyutlarını ifade etmektedir.

Model, farklı çözünürlüklerde ve aşamalarda eğitilerek transfer öğrenmeye uygun temel modeller oluşturmuş; ardından bu modeller farklı görevler için fine-tune edilerek benchmarklarda değerlendirilmiştir.

1. **Aşama:** Önceden eğitilmiş SigLIP-So400m ve Gemma 2 modellerinin ham versiyonları birleştirilerek 1 milyar örnek ile eğitilir. 1 milyar örnekteki resimler 224 × 224 çözünürlüğündedir.

2. **Aşama:** İlk olarak 50 milyon adet 448 × 448 çözünürlüğündeki resimler ile eğitilir. Ardından 10 milyon adet 896 × 896 çözünürlüğündeki resimler ile eğitilir.

3. **Aşama:** 1. veya 2. aşamadaki modele eğitildiği çözünürlüğe bağlı olarak belirli görevler ve amaçlar için fine-tuning (ince ayar) uygulanabilir. Sonrasında PaliGemma 2 modeli çeşitli akademik benchmarklar üzerinde değerlendirilmiştir. Bu benchmarklar arasında birden fazla görüntü ve kısa videolar içeren görevler de bulunmaktadır.

Ayrıca 1. ve 2. aşamadaki modellere, modellerin boyutlarına göre öğrenme oranı (learning rate) uygulanmıştır. Ana öğrenme oranı (learning rate) olarak 2×10^{-5} baz alınmıştır. Aşağıdaki tabloda uygulanan öğrenme oranları yer almaktadır:

Model	Stage 1	Stage 2
PaliGemma 2 3B	$2\times10^{-5} \times 0.5 = \mathbf{1\times10^{-5}}$	$2\times10^{-5} \times 0.5 = \mathbf{1\times10^{-5}}$
PaliGemma 2 10B	$2\times10^{-5} \times 0.25 = \mathbf{5\times10^{-6}}$	$2\times10^{-5} \times 0.25 = \mathbf{5\times10^{-6}}$
PaliGemma 2 28B	$2\times10^{-5} \times 0.25 = \mathbf{5\times10^{-6}}$	$2\times10^{-5} \times 0.25 = \mathbf{5\times10^{-6}}$

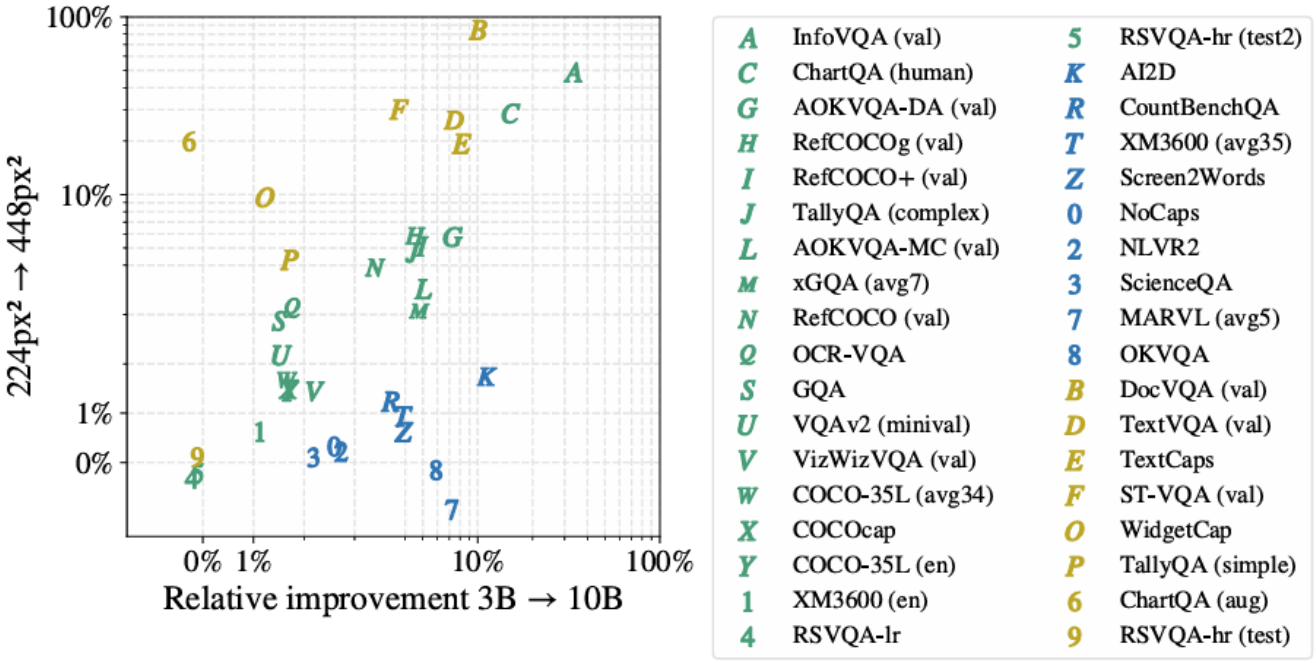
(Tablo 1)

Bu görselde PaliGemma 2'nin farklı parametreli versiyonları için çözünürlüklere göre eğitim maliyetleri verilmiştir:

	Vision Encoder	LLM	Params.	Training cost / example		
				224px^2	448px^2	896px^2
PaliGemma 2 3B	SigLIP-So400m	Gemma 2 2B	3.0B	1.0	4.6	23.5
PaliGemma 2 10B		Gemma 2 9B	9.7B	3.7	18.3	67.7
PaliGemma 2 28B		Gemma 2 27B	27.7B	18.9	63.5	~155.6

(Kaynak: "PaliGemma 2: A Family of Versatile VLMs for Transfer")

Model çözünürlüğünü 224 × 224 çözünürlüğünden 448 × 448 çözünürlüğüne artırdığımız da veya model parametresini 3 milyar (3B) parametreden 10 milyar (10B) parametreye çıkardığımızda elde edilen göreceli iyileşmeleri gösteren benchmark sonuçları aşağıdaki grafikte yer almaktadır:



(Kaynak: "PaliGemma 2: A Family of Versatile VLMs for Transfer")

Model çözünürlüğünü veya boyutunu artırmanın FLOPs (işlemci gücü gereksinimi) üzerinde de etkisi bulunmaktadır.

Yaklaşım	Değişiklik	Hesaplama maliyeti
Model boyutunu artırma	Gemma 2 dil modelini 2B'den 9B'ye çıkarma; PaliGemma 2 toplam model boyutunu yaklaşık 3B'den 10B'ye yükseltme	3,7× FLOPs
Çözünürlüğü artırma	3B PaliGemma 2 modelinde çözünürlüğü 224 × 224'ten 448 × 448'e çıkarma	4,6× FLOPs

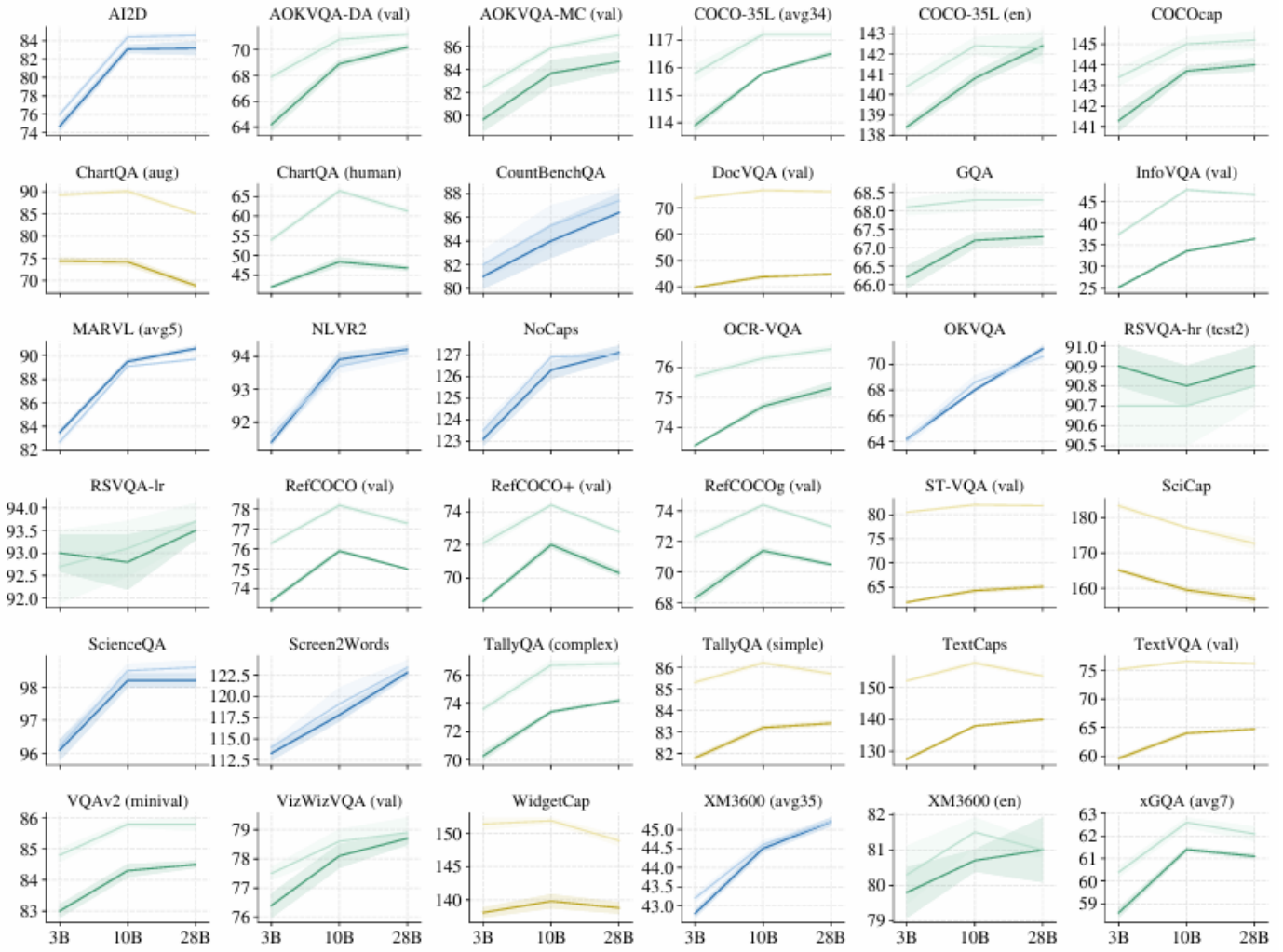
(Tablo 2)

Yeşil renk ile işaretlenen görevler, modelin hem çözünürlüğünün hem de boyutunun artırılmasından performans artışı sağlandığını göstermektedir. Sarı renk ile işaretlenen görevler yazı, döküman, ekran ve grafik anlama görevlerinde modelin çözünürlüğünün artırılmasından performans artışı sağlandığını göstermektedir. Mavi renk ile işaretlenen görevler ise model boyutunun artırılmasından daha fazla performans artışı sağlamaktadır. Bu görevlerin bazıları çok dilli veriler içeren XM3600 gibi benchmarkları kapsarken bazıları AI2D, CountBenchQA ve NLVR2 gibi ileri düzey görsel muhakeme gerektiren görevlerden oluşmaktadır.

Model boyutunu 28 milyar parametreye (28B) yükseltmek ise görevlere bağlı olarak orta düzey performans artışları sağlamaktadır veya hiç performans artışı sağlamamaktadır. Bu durumda mümkün olan en iyi performansı elde etmek isteyen ve hesaplama gücü ile gecikme süresi ile ilgili herhangi bir sorunları bulunmayan kişiler en büyük PaliGemma 2 modelini (PaliGemma 2 28B) kullanmayı tercih edebilir.

Model boyutu ve çözünürlüğün transfer performansı üzerindeki etkisini gösteren sonuçlar aşağıdaki grafikte verilmiştir. Grafiklerde 3B, 10B ve 28B model boyutları karşılaştırılmıştır. Daha açık renkli çizgiler 448 × 448 çözünürlüğü temsil

etmektedir. Gölgeli alanlar standart sapmayı göstermektedir. Görevler model boyutu ve çözünürlüğe birlikte duyarlı olanlar, yalnızca model boyutuna duyarlı olanlar ve yalnızca çözünürlüğe duyarlı olanlar şeklinde gruplandırılmıştı:



(Kaynak: "PaliGemma 2: A Family of Versatile VLMs for Transfer")

İncelenen benchmark'lardaki puan farklarının ortalaması alındığında, PaliGemma 2'nin PaliGemma'ya göre 224 × 224 çözünürlükte yaklaşık 0,65 puan, 448 × 448 çözünürlükte ise yaklaşık 0,85 puanlık ortalama artış sağladığı görülmektedir. Ancak bu ortalamalar farklı ölçüm metriklerini bir araya getirdiğinden genel performans farkını tek başına kesin biçimde temsil etmez. Bu görselde yapılan benchmark testlerinin detaylı raporları bulunmaktadır:

	#par.	#char.	#sent.	NES↓
MiniGPT-4	7B	484	5.6	52.3
mPLUG-Owl2	8B	459	4.4	48.4
InstructBLIP	7B	510	4.0	42.6
LLaVA-1.5	7B	395	4.2	40.6
VILA	7B	871	8.6	28.6
PaliGemma	3B	535	8.9	34.3
PaLI-5B	5B	1065	11.3	32.9
PaliGemma 2 $448px^2$	3B	529	7.7	28.4
PaliGemma 2 $448px^2$	10B	521	7.5	20.3

Table 6 | PaliGemma 2 results for long captioning on the DOCCI data [69]. Pali* models are models fine-tuned on DOCCI at $448px^2$; the other baselines are instruction-tuned on a broad range of tasks. Average prediction length in characters and sentences, and percentage of Non-Entailment Sentences (NES), measuring factual inaccuracies.

(Kaynak: "PaliGemma 2: A Family of Versatile VLMs for Transfer")

Makalede PaliGemma 2 görsel dil modelinin farklı görsel dil görevlerindeki performansı incelenmekte ve modelin çeşitli veri setleri ve görevler üzerindeki başarısı önceki modellerle karşılaştırılmaktadır. Örneğin, uzun ve ayrıntılı görsel betimleme görevi için DOCCI veri seti kullanılarak PaliGemma 2, MiniGPT-4, mPLUG-Owl2, InstructBLIP, LLaVA-1.5 ve VILA gibi farklı görsel dil modelleriyle karşılaştırılmıştır. Bu karşılaştırmada modellerin ürettiği açıklamaların ortalama karakter ve cümle sayıları ile Non-Entailment Sentences (NES) oranı, yani görseldeki bilgilerle uyuşmayan cümlelerin oranı değerlendirilmiştir. Sonuçlara göre 448px çözünürlükteki 3B PaliGemma 2 modeli %28,4 NES değerine ulaşırken, 10B modelde bu değer %20,3'e düşmektedir. Bu sonuçlar, model boyutunun artırılmasının uzun görsel açıklamalardaki olgusal doğruluğu artırabildiğini göstermektedir.

	224px²		448px²	
Task	PG1	PG2	PG1	PG2
AI2D	72.1	74.7 (+2.6)	73.3	76.0 (+2.7)
AOKVQA-DA (val)	61.1	64.2 (+3.1)	65.7	67.9 (+2.2)
AOKVQA-MC (val)	78.5	79.7 (+1.2)	80.3	82.5 (+2.2)
ActivityNet-CAP	34.6	34.2 (−0.4)	-	-
ActivityNet-QA	50.8	51.3 (+0.5)	-	-
COCO-35L (avg34)	113.7	113.9 (+0.2)	115.8	115.8 (+0.0)
COCO-35L (en)	139.2	138.4 (−0.8)	141.2	140.4 (−0.8)
COCOcap	141.9	141.3 (−0.6)	144.6	143.4 (−1.2)
ChartQA (aug)	74.2	74.4 (+0.2)	88.5	89.2 (+0.7)
ChartQA (human)	40.0	42.0 (+2.0)	54.2	54.0 (−0.2)
CountBenchQA	81.9	81.0 (−0.9)	83.1	82.0 (−1.1)
DocVQA (val)	37.8	39.9 (+2.1)	74.1	73.6 (−0.5)
GQA	65.6	66.2 (+0.6)	67.0	68.1 (+1.1)
InfoVQA (val)	25.5	25.2 (−0.3)	37.0	37.5 (+0.5)
MARVL (avg5)	80.6	83.5 (+2.9)	76.8	82.7 (+5.9)
MSRVTT-CAP	70.5	68.5 (−2.0)	-	-
MSRVTT-QA	50.1	50.5 (+0.4)	-	-
MSVD-QA	60.2	61.1 (+0.9)	-	-
NLVR2	90.0	91.4 (+1.4)	88.9	91.6 (+2.7)
NoCaps	121.7	123.1 (+1.4)	123.6	123.5 (−0.1)
OCR-VQA	72.3	73.4 (+1.1)	74.6	75.7 (+1.1)
OKVQA	63.5	64.2 (+0.7)	63.2	64.1 (+0.9)
RSVQA-hr (test)	92.6	92.7 (+0.1)	92.8	92.8 (+0.0)
RSVQA-hr (test2)	90.6	90.9 (+0.3)	90.5	90.7 (+0.2)
RSVQA-lr	92.6	93.0 (+0.4)	93.1	92.7 (−0.4)
RefCOCO (testA)	75.7	75.7 (+0.0)	77.9	78.6 (+0.7)
RefCOCO (testB)	70.7	71.0 (+0.3)	72.4	73.5 (+1.1)
RefCOCO (val)	73.4	73.4 (+0.0)	75.6	76.3 (+0.7)
RefCOCO+ (testA)	71.9	72.7 (+0.8)	74.2	76.1 (+1.9)
RefCOCO+ (testB)	64.5	64.2 (−0.3)	64.5	67.0 (+2.5)
RefCOCO+ (val)	68.3	68.6 (+0.3)	69.8	72.1 (+2.3)
RefCOCOg (test)	68.2	69.0 (+0.8)	71.0	72.7 (+1.7)
RefCOCOg (val)	67.7	68.3 (+0.6)	70.1	72.3 (+2.2)
ST-VQA (val)	61.6	61.9 (+0.3)	79.7	80.5 (+0.8)
SciCap	162.3	165.1 (+2.8)	181.5	183.3 (+1.8)
ScienceQA	95.4	96.1 (+0.7)	95.9	96.2 (+0.3)
Screen2Words	117.6	113.3 (−4.3)	119.6	114.0 (−5.6)
TallyQA (complex)	69.6	70.3 (+0.7)	72.3	73.6 (+1.3)
TallyQA (simple)	81.7	81.8 (+0.1)	84.9	85.3 (+0.4)
TextCaps	127.5	127.5 (+0.0)	153.9	152.1 (−1.8)
TextVQA (val)	59.0	59.6 (+0.6)	74.6	75.2 (+0.6)
VATEX	79.7	80.8 (+1.1)	-	-
VQAv2 (minival)	82.1	83.0 (+0.9)	84.6	84.8 (+0.2)
VizWizVQA (val)	73.7	76.4 (+2.7)	75.5	77.5 (+2.0)
WidgetCap	136.1	138.1 (+2.0)	148.4	151.4 (+3.0)
XM3600 (avg35)	41.9	42.8 (+0.9)	42.4	43.2 (+0.8)
XM3600 (en)	78.0	79.8 (+1.8)	80.0	80.3 (+0.3)
xGQA (avg7)	57.3	58.6 (+1.3)	57.9	60.4 (+2.5)

(Kaynak: "PaliGemma 2: A Family of Versatile VLMs for Transfer")

Benchmark testlerinde belirli performans düşüşleri de gözlemlenmektedir. Örneğin Screen2Words görevinde PaliGemma 1, 224 × 224 çözünürlükte 117,6 ve 448 × 448 çözünürlükte 119,6 puan alırken PaliGemma 2 aynı çözünürlüklerde sırasıyla 113,3 ve 114,0 puan almıştır. Dolayısıyla PaliGemma 2'nin puanı her iki çözünürlükte de PaliGemma 1'in altında kalmaktadır. Bununla birlikte çözünürlüğün artırılması her iki modelde de performans artışı sağlamıştır. Bu sonuç, Gemma 2 tabanlı modele geçişin her görevde performans artışı sağlamadığını göstermektedir.

Uygulanan benchmark testlerine göre hem model çözünürlüğünü hem de model boyutunu artırmanın sağladığı avantajlar ve dezavantajlar aşağıdaki tabloda yer almaktadır:

	224px²			448px²		
	3B	10B	28B	3B	10B	28B
AI2D [40]	74.7 (±0.5)	83.1 (±0.4)	83.2 (±0.7)	76.0 (±0.2)	84.4 (±0.4)	84.6 (±0.4)
AOKVQA-DA (val) [81]	64.2 (±0.5)	68.9 (±0.3)	70.2 (±0.2)	67.9 (±0.3)	70.8 (±0.5)	71.2 (±0.2)
AOKVQA-MC (val) [81]	79.7 (±1.0)	83.7 (±1.1)	84.7 (±0.8)	82.5 (±0.4)	85.9 (±0.2)	87.0 (±0.3)
ActivityNet-CAP [43]	34.2 (±0.3)	35.9 (±0.5)	-	-	-	-
ActivityNet-QA [107]	51.3 (±0.2)	53.2 (±0.4)	-	-	-	-
COCO-35L (avg34) [91]	113.9 (±0.2)	115.8 (±0.0)	116.5 (±0.1)	115.8 (±0.3)	117.2 (±0.1)	117.2 (±0.1)
COCO-35L (en) [91]	138.4 (±0.2)	140.8 (±0.3)	142.4 (±0.4)	140.4 (±0.4)	142.4 (±0.4)	142.3 (±0.8)
COCOcap[51]	141.3 (±0.5)	143.7 (±0.2)	144.0 (±0.3)	143.4 (±0.4)	145.0 (±0.3)	145.2 (±0.4)
ChartQA (aug) [63]	74.4 (±0.7)	74.2 (±0.8)	68.9 (±0.6)	89.2 (±0.4)	90.1 (±0.5)	85.1 (±0.2)
ChartQA (human) [63]	42.0 (±0.3)	48.4 (±1.1)	46.8 (±0.6)	54.0 (±0.6)	66.4 (±0.5)	61.3 (±0.6)
CountBenchQA [9]	81.0 (±1.0)	84.0 (±1.4)	86.4 (±1.6)	82.0 (±1.2)	85.3 (±1.7)	87.4 (±1.0)
DocVQA (val) [64]	39.9 (±0.3)	43.9 (±0.6)	44.9 (±0.4)	73.6 (±0.3)	76.6 (±0.5)	76.1 (±0.4)
GQA[29]	66.2 (±0.3)	67.2 (±0.2)	67.3 (±0.2)	68.1 (±0.2)	68.3 (±0.3)	68.3 (±0.1)
InfoVQA (val) [65]	25.2 (±0.2)	33.6 (±0.2)	36.4 (±0.1)	37.5 (±0.3)	47.8 (±0.2)	46.7 (±0.4)
MARVL (avg5) [52]	83.5 (±0.2)	89.5 (±0.2)	90.6 (±0.2)	82.7 (±0.3)	89.1 (±0.0)	89.7 (±0.1)
MSRVTT-CAP [101]	68.5 (±1.3)	72.1 (±0.5)	-	-	-	-
MSRVTT-QA [100]	50.5 (±0.1)	51.9 (±0.1)	-	-	-	-
MSVD-QA [12]	61.1 (±0.2)	62.5 (±0.2)	-	-	-	-
NLVR2 [87]	91.4 (±0.1)	93.9 (±0.2)	94.2 (±0.1)	91.6 (±0.2)	93.7 (±0.2)	94.1 (±0.2)
NoCaps [2]	123.1 (±0.3)	126.3 (±0.4)	127.1 (±0.3)	123.5 (±0.3)	126.9 (±0.1)	127.0 (±0.2)
OCR-VQA [67]	73.4 (±0.0)	74.7 (±0.1)	75.3 (±0.2)	75.7 (±0.1)	76.3 (±0.1)	76.6 (±0.1)
OKVQA [62]	64.2 (±0.1)	68.0 (±0.1)	71.2 (±0.2)	64.1 (±0.4)	68.6 (±0.5)	70.6 (±0.2)
RSVQA-hr (test) [55]	92.7 (±0.1)	92.6 (±0.0)	92.7 (±0.0)	92.8 (±0.0)	92.8 (±0.1)	92.8 (±0.1)
RSVQA-hr (test2) [55]	90.9 (±0.1)	90.8 (±0.1)	90.9 (±0.1)	90.7 (±0.2)	90.7 (±0.2)	90.8 (±0.1)
RSVQA-lr [55]	93.0 (±0.4)	92.8 (±0.6)	93.5 (±0.2)	92.7 (±0.8)	93.1 (±0.6)	93.7 (±0.4)
RefCOCO (testA) [106]	75.7 (±0.2)	77.2 (±0.1)	76.8 (±0.1)	78.6 (±0.3)	79.7 (±0.1)	79.3 (±0.1)
RefCOCO (testB) [106]	71.0 (±0.3)	74.2 (±0.3)	73.9 (±0.1)	73.5 (±0.1)	76.2 (±0.3)	74.8 (±0.1)
RefCOCO (val) [106]	73.4 (±0.1)	75.9 (±0.1)	75.0 (±0.0)	76.3 (±0.1)	78.2 (±0.1)	77.3 (±0.1)
RefCOCO+ (testA) [39]	72.7 (±0.2)	74.7 (±0.2)	73.6 (±0.2)	76.1 (±0.2)	77.7 (±0.2)	76.6 (±0.1)
RefCOCO+ (testB) [39]	64.2 (±0.2)	68.4 (±0.3)	67.1 (±0.1)	67.0 (±0.3)	71.1 (±0.2)	68.6 (±0.1)
RefCOCO+ (val) [39]	68.6 (±0.1)	72.0 (±0.2)	70.3 (±0.2)	72.1 (±0.3)	74.4 (±0.1)	72.8 (±0.1)
RefCOCOg (test) [61]	69.0 (±0.2)	71.9 (±0.1)	70.7 (±0.1)	72.7 (±0.1)	74.8 (±0.1)	73.7 (±0.1)
RefCOCOg (val) [61]	68.3 (±0.3)	71.4 (±0.2)	70.5 (±0.1)	72.3 (±0.2)	74.4 (±0.1)	73.0 (±0.1)
ST-VQA (val) [10]	61.9 (±0.1)	64.3 (±0.4)	65.1 (±0.4)	80.5 (±0.1)	82.0 (±0.3)	81.8 (±0.1)
SciCap [27]	165.1 (±0.5)	159.5 (±0.7)	156.9 (±1.0)	183.3 (±0.7)	177.2 (±0.3)	172.7 (±1.5)
ScienceQA [59]	96.1 (±0.3)	98.2 (±0.2)	98.2 (±0.2)	96.2 (±0.2)	98.5 (±0.2)	98.6 (±0.2)
Screen2Words [95]	113.3 (±0.8)	117.8 (±0.7)	122.8 (±0.5)	114.0 (±0.5)	119.1 (±1.9)	123.4 (±0.8)
TallyQA (complex) [1]	70.3 (±0.3)	73.4 (±0.1)	74.2 (±0.1)	73.6 (±0.2)	76.7 (±0.3)	76.8 (±0.2)
TallyQA (simple) [1]	81.8 (±0.1)	83.2 (±0.1)	83.4 (±0.1)	85.3 (±0.1)	86.2 (±0.1)	85.7 (±0.1)
TextCaps [82]	127.5 (±0.3)	137.9 (±0.3)	139.9 (±0.4)	152.1 (±0.3)	157.7 (±0.7)	153.6 (±0.5)
TextVQA (val) [83]	59.6 (±0.3)	64.0 (±0.3)	64.7 (±0.2)	75.2 (±0.2)	76.6 (±0.1)	76.2 (±0.1)
VATEX [97]	80.8 (±0.4)	82.7 (±0.5)	-	-	-	-
VQAv2 (minival) [25]	83.0 (±0.2)	84.3 (±0.2)	84.5 (±0.1)	84.8 (±0.2)	85.8 (±0.1)	85.8 (±0.2)
VizWizVQA (val) [26]	76.4 (±0.4)	78.1 (±0.4)	78.7 (±0.2)	77.5 (±0.2)	78.6 (±0.4)	78.9 (±0.5)
WidgetCap [49]	138.1 (±0.7)	139.8 (±1.0)	138.8 (±0.8)	151.4 (±0.8)	151.9 (±0.4)	148.9 (±0.7)
XM3600 (avg35) [91]	42.8 (±0.1)	44.5 (±0.1)	45.2 (±0.1)	43.2 (±0.1)	44.6 (±0.1)	45.2 (±0.1)
XM3600 (en) [91]	79.8 (±0.7)	80.7 (±0.3)	81.0 (±0.9)	80.3 (±0.8)	81.5 (±0.4)	81.0 (±0.2)
xGQA (avg7) [73]	58.6 (±0.2)	61.4 (±0.1)	61.1 (±0.1)	60.4 (±0.2)	62.6 (±0.2)	62.1 (±0.3)

(Kaynak: "PaliGemma 2: A Family of Versatile VLMs for Transfer")

Tablodaki değerleri 3B 224 × 224 temel model üzerinden karşılaştırırsak, performans artışlarının ortalaması alttaki tabloda yer almaktadır. (Eksik değer bulunan benchmarklar ortalamaya dahil edilmemiştir.)

Tablo 3. Model boyutu ve çözünürlük değişimlerinin benchmark performansına etkisi

Karşılaştırma	Benchmark sayısı	Ortalama mutlak artış (puan)	Ortalama göreceli artış (%)
PaliGemma 2 3B (224²) → 10B (224²)	42	+2,82	+4,48
PaliGemma 2 3B (224²) → 28B (224²)	42	+3,08	+4,93
PaliGemma 2 3B (224²) → 3B (448²)	42	+5,25	+8,44

(Tablo 3)

Model boyutu ve çözünürlük artışının hesaplama maliyetleri ve performansa etkisi nedir?

Ortalama skor artışları baz alındığında, incelenen benchmark'larda çözünürlüğü artırmanın model boyutunu artırmaya kıyasla daha yüksek performans artışı sağladığı görülmektedir.

- Modeli 3B boyutunda sabit tutarak çözünürlüğü 224 × 224'ten 448 × 448'e çıkarmak, hesaplama maliyetini 4,6 kat artırmaktadır (Tablo 2).

- Çözünürlüğü 224 × 224 olarak sabit tutup model boyutunu 3B'den 10B'ye çıkarmak ise hesaplama maliyetini 3,7 kat artırmaktadır (Tablo 2).

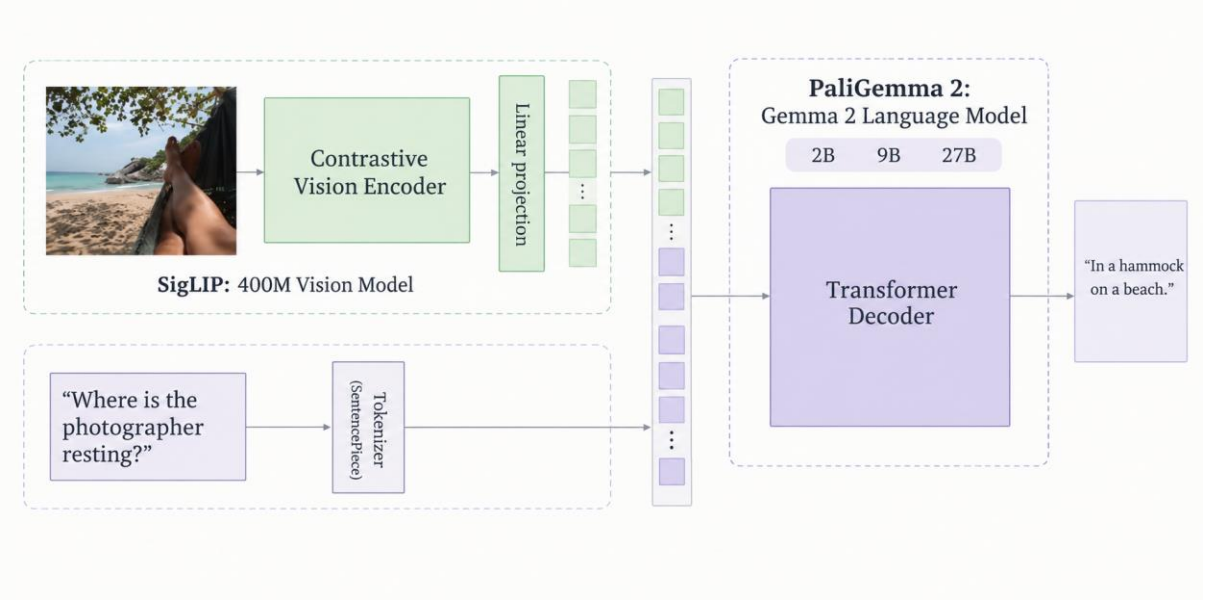
Yaklaşım	Göreceli performans artışı	Ek normalize edilmiş FLOPs maliyeti (katsayı – 1)	Artış / ek maliyet
Çözünürlük artışı	%8,44	3,6	2,34
Model boyutu artışı	%4,48	2,7	1,66

Üstteki bilgilere göre yapılan çıkarımlar:

- **Metin ve arayüz (UI) odaklı görevler için çözünürlük:** Optik karakter tanıma (OCR), belge okuma (DocVQA) ve grafik analizi gibi görsel ayrıntıların önemli olduğu görevlerde yüksek çözünürlük tercih edilebilir. Çözünürlüğü artırmak, hesaplama maliyetini 4,6 katına çıkarsa da küçük metinlerin, ikonların ve arayüz bileşenlerinin daha ayrıntılı işlenmesine olanak tanır. Örneğin, masaüstü uygulamalarında işlem yapan veya ürün demoları kaydeden otonom yapay zekâ ajanları için ekrandaki küçük metinlerin ve bileşenlerin doğru algılanması önemlidir. WidgetCap ve Screen2Words gibi arayüz odaklı benchmark'larda çözünürlük artışının sağladığı kazanımlar bu açıdan değerlendirilebilir. Bununla birlikte, çözünürlük artışının her metin veya arayüz görevinde model boyutunu artırmaktan daha iyi sonuç vereceği varsayılmamalıdır; bu tercih ilgili benchmark sonuçlarına göre yapılmalıdır.

- **Genel bilgi ve mantıksal çıkarım gerektiren görevler için model boyutu:** Modelin genel dünya bilgisinden yararlanması veya karmaşık ilişkileri değerlendirmesi gereken görevlerde model boyutunu artırmak da performans kazancı sağlayabilir. Örneğin, ScienceQA ve OKVQA gibi benchmark'larda model boyutunu artırmanın etkisi, çözünürlük artışının etkisiyle karşılaştırılarak değerlendirilmelidir. Buna karşın GQA sonuçlarında 3B modelin çözünürlüğünü 224 × 224'ten 448 × 448'e çıkarmak skoru 66,2'den 68,1'e yükseltirken model boyutunu 3B'den 10B'ye çıkarmak skoru 67,2'ye yükseltmektedir. Bu da GQA özelinde çözünürlük artışının daha yüksek bir kazanç sağladığını göstermektedir. Ancak bu tek örnekten hareketle tüm mantıksal çıkarım ve genel bilgi görevlerinde çözünürlük artışının daha avantajlı olduğu sonucuna varılmamalıdır. Model boyutu ile çözünürlük arasındaki tercih, görev bazındaki performans kazanımları ve hesaplama maliyetleri birlikte dikkate alınarak yapılmalıdır.

PaliGemma 2 mimarisi ve teknik detayları:



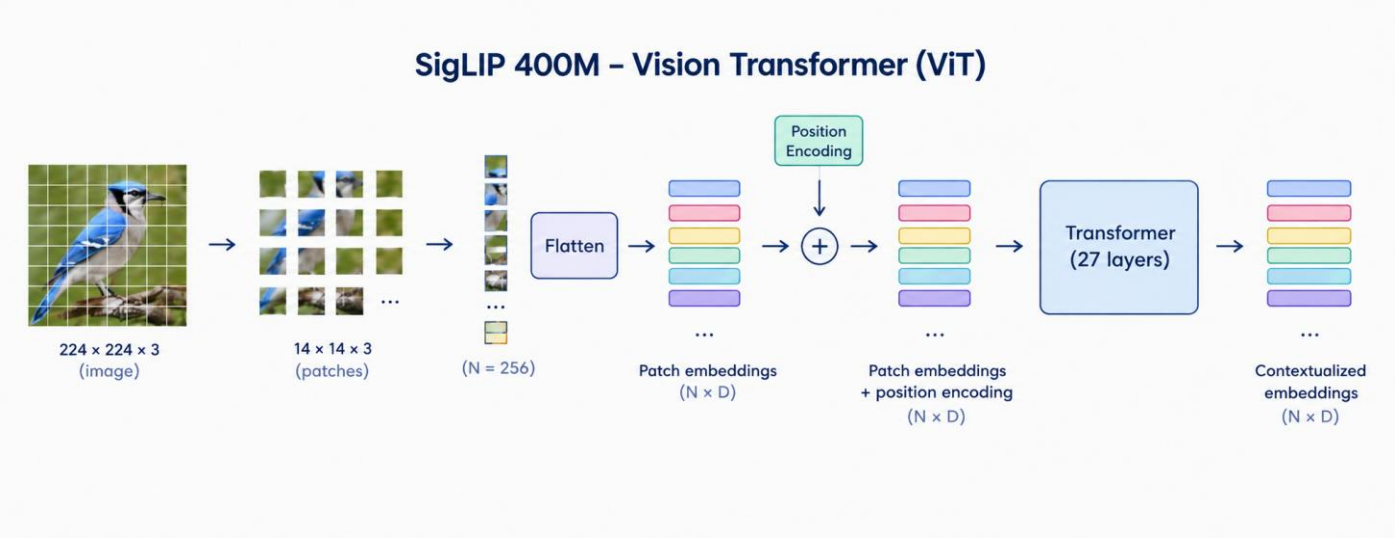
(Şekil: Yazar tarafından oluşturuldu.)

PaliGemma 2 dil modelinde kullanılan mimari yaklaşımı, blokların üst üste dizilimi şeklinde düşünülebilir. PaliGemma 2, bir vision encoder ve bir language model bileşeninden oluşan multimodal bir mimariye sahiptir:

- SigLIP 400M Görüntü Modeli (Vision Encoder)
- Gemma 2 (2B, 9B, 27B) Dil Modeli (Language Model)

Model, girilen metinsel (yazı) ve görsel girdileri işlemektedir. Görsel girdiye SigLIP 400M Vision Encoder ile gömme (embedding) işlemi uygulanır. SigLIP 400M Vision Encoder bir Vision Transformer modelidir ve modele karşılaştırmalı eğitim (contrastive training) işlemi uygulanır.

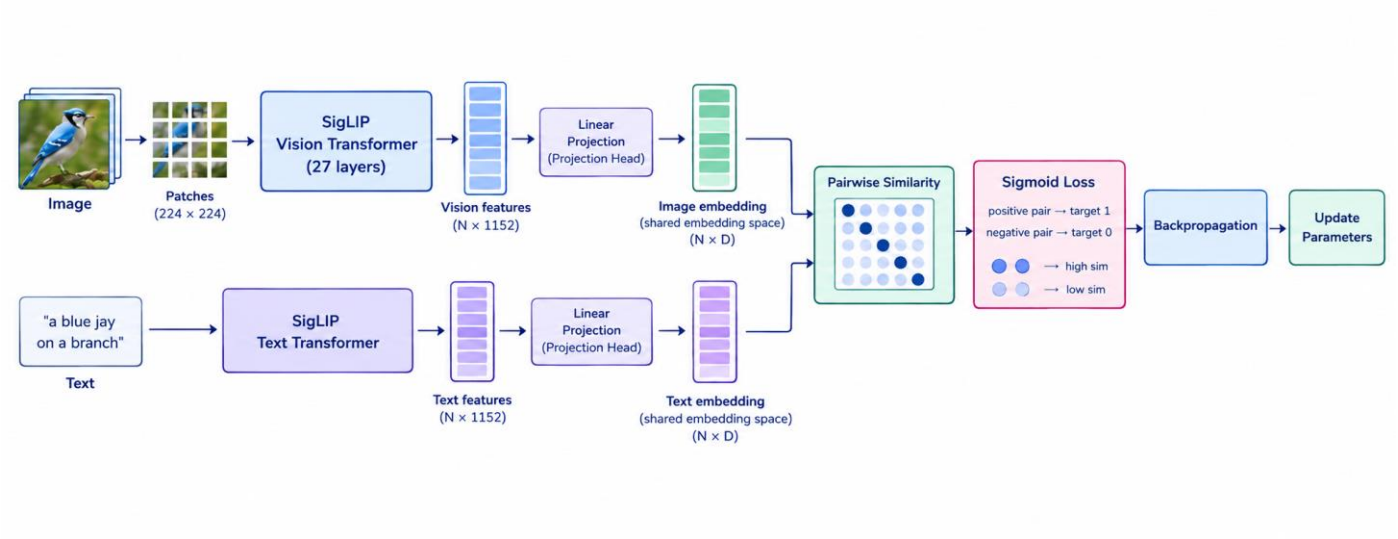
Öncelikle SigLIP 400M modelinin nasıl bir mimaride çalıştığına göz atalım:



(Şekil: Yazar tarafından oluşturuldu.)

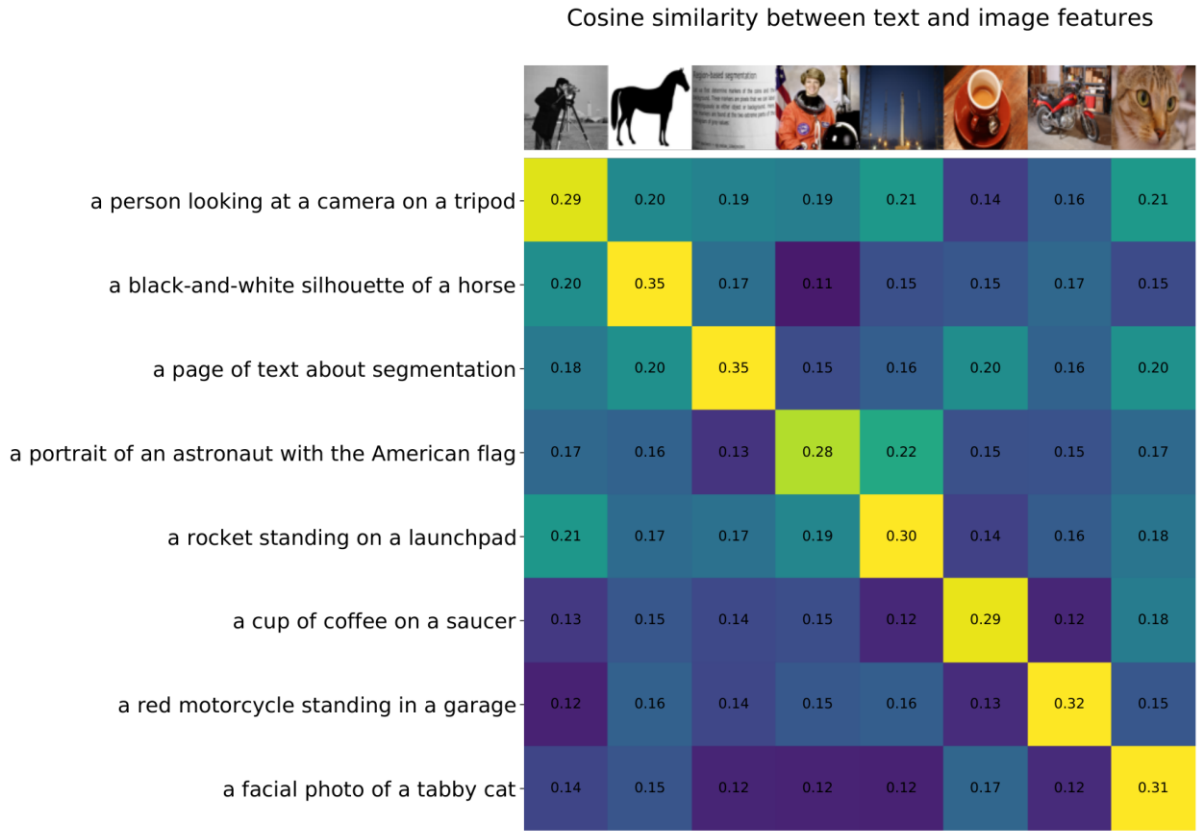
SigLIP 400M Vision Encoder modelinde öncelikle görsel girdi, boyutu önceden belirlenmiş kernel'ler ile patch'lere ayrılır. Örneğin, 224 × 224 boyutundaki görsel girdi 14 × 14 boyutundaki kernel'ler ile işlendiğinde 16 × 16 olmak üzere toplam 256 patch elde edilir. Patch'lere ayırma işlemi convolution (evrişim) işlemi ile gerçekleştirilir. Ardından flatten (düzleştirme) işlemi ile her patch kendi embedding vektörüne dönüştürülür ve toplam 256 patch embedding vektörü elde edilir. Position encoding vektörleri bu patch embedding vektörleri ile toplanarak her patch'e konum bilgisi eklenir. Ardından elde edilen vektörler 27 Transformer katmanından geçirilerek bağlamsal embedding vektörleri elde edilir. Bu sayede görsel veriler, dil modelinin işleyebileceği anlamsal vektörlere dönüştürülmüş olur.

SigLIP 400M Vision Encoder nasıl eğitilir?



(Şekil: Yazar tarafından oluşturuldu.)

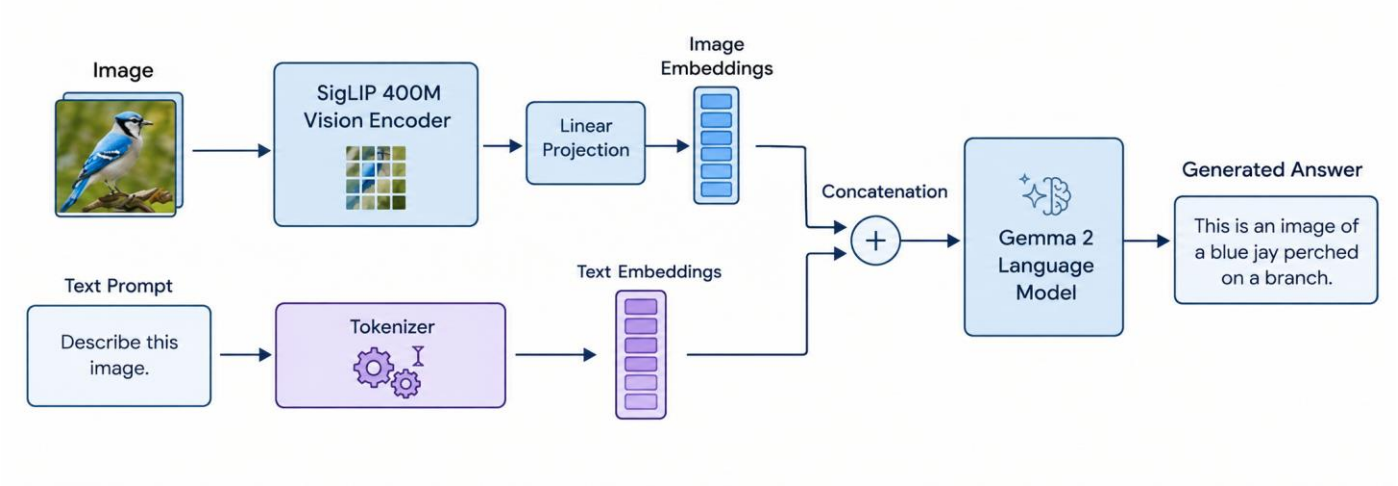
Görsel ve metinsel (yazı) girdiler embedding vektörlerine çevrilir ve tüm metinlerin tüm resimlerle eşleşme oranlarını gösteren bir tablo oluşturulur. Sonrasında tablodaki tüm görsel ve metin çiftleri (eşleşen ve eşleşmeyen her bir kombinasyon) üzerinde bağımsız sigmoid hesaplamaları gerçekleştirilir. Ardından hesaplanan kayıp (loss) değerlerinden elde edilen toplam kayıp kullanılarak geriye yayılım (backpropagation) işlemi uygulanır ve model eğitilir. Görsel ve metinsel (yazı) girdilerin karşılaştırılmasına dayalı bu eğitim işlemi karşılaştırmalı eğitim (contrastive training) olarak adlandırılmaktadır.



(Kaynak: [PDF] Multimodal datasets: misogyny, pornography, and malignant stereotypes | Semantic Scholar)

Yukarıdaki tabloda örnek görsellerin ve örnek metinlerin eşleşme oranları yer almaktadır.

SigLIP 400M Vision Encoder modelinden bağlamsal embedding vektörü elde ettik. Peki PaliGemma 2 görüntü dil modelinde resimden cevaba giden yol tam olarak nedir? Modele girilen metin (prompt), tokenizer ile embedding vektörüne çevrilmektedir. Sonrasında girilen metin (prompt) embedding vektörü ile SigLIP 400M Vision Encoder modelinin ürettiği görsel girdinin embedding vektörü birleştirilmekte ve Gemma 2 dil modeline transfer edilmektedir.



(Şekil: Yazar tarafından oluşturuldu.)

Yukarıdaki mimaride, PaliGemma 2 görsel dil modelinde görüntüden cevaba giden süreç gösterilmektedir. Görsel girdi, SigLIP 400M Vision Encoder tarafından işlenerek bir dizi görsel embedding vektörüne dönüştürülür. Bu vektörler, lineer projeksiyon katmanından geçirilerek Gemma 2 dil modelinin giriş boyutuyla uyumlu hâle getirilir. Metinsel girdi ise tokenizer aracılığıyla token dizisine dönüştürülür. Görsel ve metinsel tokenlar birleştirilerek Gemma 2 dil modeline aktarılır. Son olarak Gemma 2, bu girdileri kullanarak otoregresif biçimde metinsel bir çıktı üretir.

Yapılan deneylere ve bulgulara ilişkin:

PaliGemma 2 görüntü dil modelinde, PaliGemma 1 görüntü dil modelinde daha önce derinlemesine uygulanmamış olan birtakım deneyler ve yeni görevler uygulanmıştır. Örneğin; metin algılama ve tanıma, tablo yapısı tanıma, moleküler yapı tanıma, optik müzik notası tanıma, uzun betimleme (caption) üretimi, uzamsal akıl yürütme ve radyografi raporu üretimi gibi alanlar incelenmiştir. PaliGemma 2, bu görevlerin birçoğunda en üst düzey (state-of-the-art) sonuçlar elde etmiştir. Son olarak, modelin doğrudan cihaz üzerinde (CPU'da) çalışabilmesi amacıyla düşük hassasiyetli (low-precision) varyantlarının benchmark testleri ve analizleri de gerçekleştirilmiştir.

Modelin uzun ve ayrıntılı görsel betimleme başarısı, insan değerlendirmesi ve “Tersine Kanıt Cümleleri” (Non-Entailment Sentences, NES) metriği üzerinden incelenmiştir. Ayrıca Tablo 6’da LLaVA-1.5, VILA ve InstructBLIP gibi farklı güncel görüntü dil modelleriyle karşılaştırma yapılması, modelin bu alandaki performansının değerlendirilmesini güçlendirmektedir. Bununla birlikte, PaliGemma 2’nin DOCCI veri seti üzerinde fine-tune edildiği, karşılaştırılan diğer modellerin ise geniş kapsamlı görevler üzerinde instruction-tuning ile eğitildiği belirtilmektedir. Bu nedenle, modeller arasındaki performans farkının ne ölçüde model mimarisinden, ne ölçüde kullanılan eğitim ve fine-tuning koşullarından kaynaklandığı tam olarak ayrıştırılamamaktadır. Daha kontrollü bir karşılaştırmada, rakip modellerin de DOCCI üzerinde benzer fine-tuning koşullarında değerlendirilmesi, PaliGemma 2’nin uzun ve ayrıntılı görsel betimleme konusundaki gerçek üstünlüğünü daha net ortaya koyabilirdi.

Dil modeli tarafında yapılan kapsamlı ölçeklendirme analizinin aksine, görüntü işleme ayağında sadece SigLIP 400M modeline bağlı kalınmıştır. Sisteme entegre edilecek alternatif ve güncel vision encoder modelleriyle yapılacak kıyaslama testleri, mimarinin gerçek potansiyelini çok daha net ortaya koyabilirdi.

Makalede model boyutu ve görüntü çözünürlüğünün transfer performansı üzerindeki etkileri incelenmiş, farklı görevlerin bu değişkenlere duyarlılıkları karşılaştırılmıştır. Ayrıca Tablo 1'de ve Bölüm 4.1.1'de iki ölçeklendirme seçeneğinin FLOPs maliyetleri belirtilmiş, Şekil 3'te ise hangi görevin hangi seçenekten kazandığı gruplandırılarak gösterilmiştir. Bununla birlikte makale, farklı benchmark'lardaki göreceli performans kazanımlarını ortak bir maliyet-performans göstergesinde birleştirmemektedir. Bu çalışmada Tablo 3'teki ortalama göreceli performans artışları ve hesaplama maliyeti katsayıları

birlikte değerlendirilerek iki ölçeklendirme seçeneği için yaklaşık bir karşılaştırma yapılmaktadır. Sonuçlar, kullanılan ortalama hesaplama yöntemine ve benchmark kümesine bağlı olarak yorumlanmalıdır.