Производительность TensorRT (RTX 4090M 48GB)

	int8		fp16		best		fp32	
	Samples/s	Samples per watt	Samples/s	Samples per watt	Samples/s	Samples per watt	Samples/s	Samples per watt
BERT-Base (Seq=128, B=8)	4006,816	8,904	4507,832	10,017	4499,688	9,999	1972,776	4,384
BERT-Base (Seq=128, B=24)	4964,832	11,033	5055,816	11,235	5060,472	11,245	2130,158	4,734
BERT-Large (Seq=128, B=8)	1786,216	3,969	2046,704	4,548	2047,600	4,550	784,644	1,744
BERT-Large (Seq=128, B=12)	1967,987	4,373	1958,687	4,353	1955,916	4,346	750,110	1,667
Resnet50 (B=8)	22801,120	50,669	14382,800	31,962	22834,000	50,742	5176,184	11,503
Resnet50 (B=32)	38303,040	85,118	19235,328	42,745	38315,840	85,146	5806,336	12,903
MS Resnet50 1.5 (B=8)	22188,560	49,308	13860,880	30,802	22140,880	49,202	4858,312	10,796
MS Resnet50 1.5 (B=32)	36419,840	80,933	18396,192	40,880	36514,560	81,143	5482,784	12,184
Swin Base (224x224, B=1)	615,973	1,369	817,914	1,818	819,536	1,821	588,728	1,308
Swin Base (224x224, B=8)	2136,944	4,749	3540,400	7,868	3534,856	7,855	1764,368	3,921
Swin Large (224x224, B=1)	469,916	1,044	553,252	1,229	553,346	1,230	314,507	0,699
Swin Large (224x224, B=8)	1178,896	2,620	1565,936	3,480	1559,496	3,466	666,186	1,480
ViT Base (224x224, B=1)	807,533	1,795	1041,250	2,314	1038,950	2,309	696,075	1,547
ViT Base (224x224, B=8)	4850,128	10,778	6022,536	13,383	6014,304	13,365	3168,256	7,041
ViT Large (224x224, B=1)	335,664	0,746	386,880	0,860	385,721	0,857	184,035	0,409
ViT Large (224x224, B=8)	702,493	1,561	815,928	1,813	815,440	1,812	314,007	0,698
Yolo V4 (608x608, B=1)	1000,840	2,224	691,826	1,537	1000,520	2,223	247,439	0,550
Efficientnet B0 (B=1)	2324,230	5,165	2587,110	5,749	2539,030	5,642	1867,620	4,150
Efficientnet B0 (B=8)	13747,040	30,549	12896,720	28,659	13626,720	30,282	7243,952	16,098
Efficientnet B0 (B=32)	28699,424	63,776	21303,520	47,341	28693,504	63,763	8737,216	19,416
Efficientnet B4 (B=1)	1058,340	2,352	1172,640	2,606	1162,030	2,582	652,372	1,450
Efficientnet B4 (B=8)	5840,368	12,979	5345,760	11,879	5842,440	12,983	2602,120	5,782
Efficientnet B4 (B=32)	11617,792	25,817	8351,136	18,558	11634,528	25,855	3156,675	7,015

✅ RTX 4090M 48GB показывает отличные результаты как по скорости обработки, так и по энергоэффективности в популярных нейросетевых моделях.

✅ Карта идеально подходит для интенсивных вычислений и инференса, обеспечивая высокую скорость и экономичность.