GPU – TPU – CPU hız karşılaştırması

CPU: Bilgisayarın veri işleme ve yazılım komutlarını gerçekleştiren bölümüdür.

GPU: Bilgisayarda grafik yaratma, işleme ve göstermek için kullanılan aygıttır.

TPU: Google tarafından özel olarak geliştirilmiş makine öğrenimi hızlandırıcı işlem ünitesidir.

CPU	GPU	TPU
Several core	Thousands of Cores	Matrix based workload
Low latency	High data throughput	High latency
Serial processing	Massive parallel computing	High data throughput
Limited simultaneous operations	Limited multitasking	Suited for large batch sizes
Large memory capacity	Low memory	Complex neural network models

Bu karşılaştırmada Kaggle Notebook kullanıldı.

GPU: Tesla P100

CPU: Intel Xeon

TPU: v3-8

Veri: CIFAR-10

Farklı batch boyutlarında cpu, gpu ve tpu için geçen süreler hesaplandı.

Süreler ilk epoch, ilk beş epoch ve tüm epoch değerleri olarak üçe ayrılıyor.

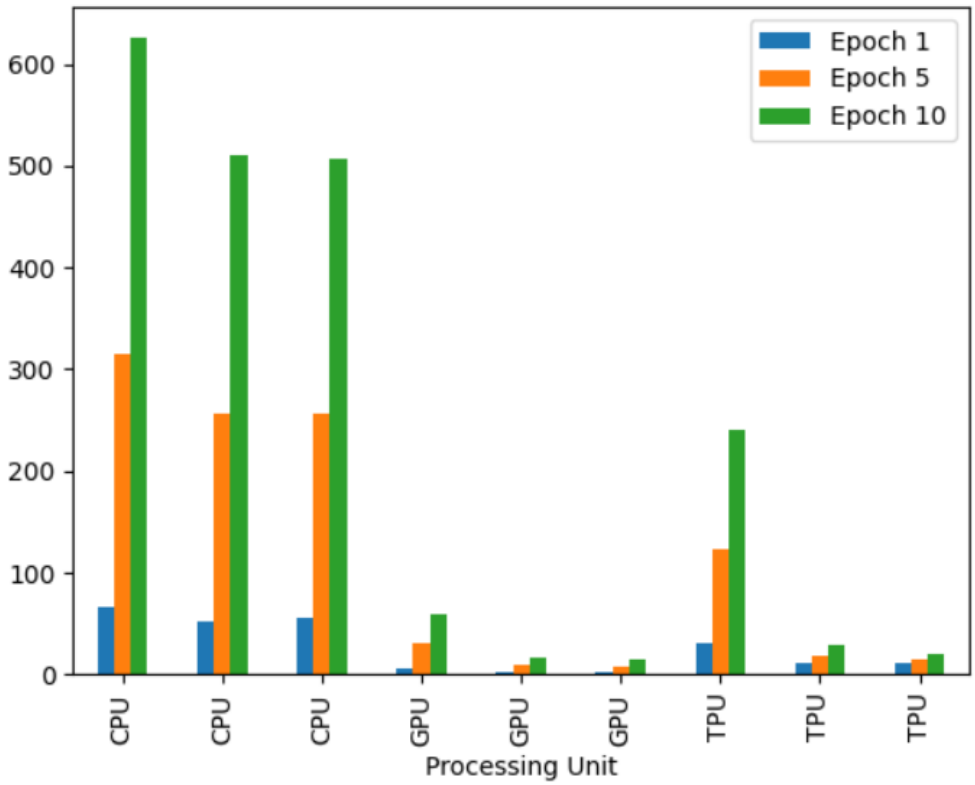
Kullanılan CNN modeli:

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 30, 30, 64)        1792
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 15, 15, 64)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 13, 13, 128)       73856
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 6, 6, 128)         0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 4, 4, 128)         147584
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 2, 2, 128)         0
_________________________________________________________________
flatten (Flatten)            (None, 512)               0
_________________________________________________________________
dense (Dense)                (None, 64)                32832
_________________________________________________________________
dense_1 (Dense)              (None, 10)                650
=================================================================
Total params: 256,714
Trainable params: 256,714
Non-trainable params: 0
_________________________________________________________________
```

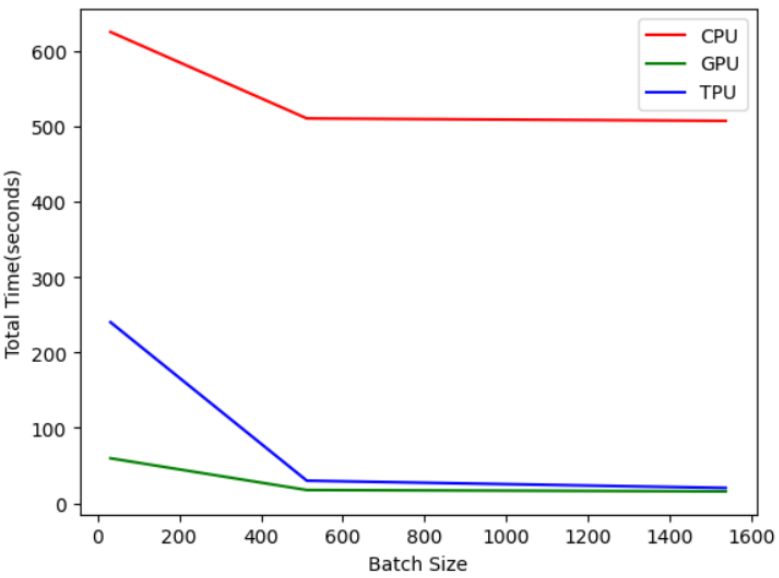
Model eğitimi sonuçları:

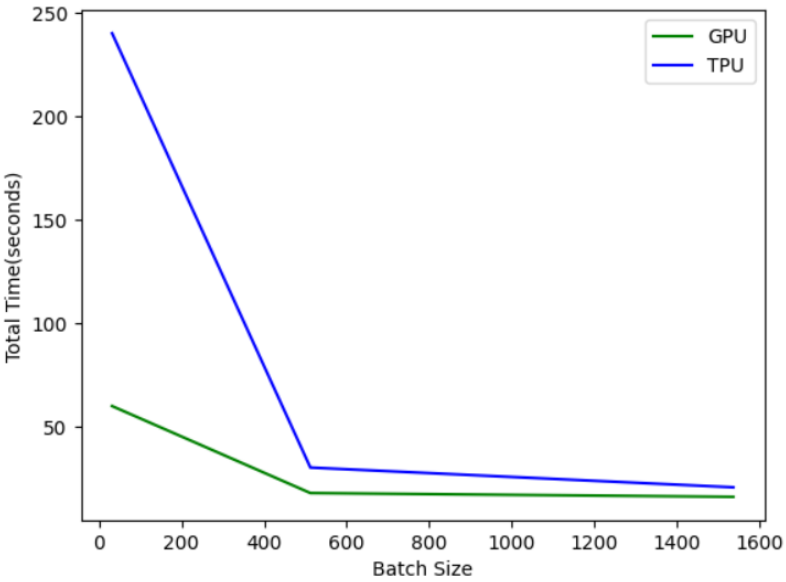
	Processing Unit	Batch Size	Epoch 1	Epoch 5	Epoch 10
0	CPU	32	66.643250	314.888119	624.910108
1	CPU	512	52.286886	255.621997	510.266024
2	CPU	1536	55.097071	256.407487	507.027361
3	GPU	32	6.785292	30.494279	59.663104
4	GPU	512	2.839296	9.744983	17.577993
5	GPU	1536	3.125681	8.768707	15.744841
6	TPU	32	30.762377	123.293698	240.057454
7	TPU	512	10.957684	19.423637	29.915033
8	TPU	1536	10.661381	15.059212	20.359209

İlk bakışta CPU ile eğitim sürelerinin GPU ve TPU değerlerine göre oldukça fazla olduğunu gözlemlemekteyiz. GPU en düşük işlem süresine sahip görünüyor.



Yukarıdaki grafikte süreler arasındaki farkı daha net şekilde görebiliyoruz. Bu sonuçlardan yola çıkarak CPU kullanımının verimsiz olacağı çıkarımını yapabiliriz. GPU genel olarak hızlı olmasına karşın TPU değerlerinde artan batch size ile sürede büyük bir azalma yaşanıyor





TPU ve GPU Süre Değişimi:

	Processing Unit	Batch Size	Epoch 1	Epoch 5	Epoch 10
6	TPU	32	30.762377	123.293698	240.057454
7	TPU	512	10.957684	19.423637	29.915033
8	TPU	1536	10.661381	15.059212	20.359209

	Processing Unit	Batch Size	Epoch 1	Epoch 5	Epoch 10
3	GPU	32	6.785292	30.494279	59.663104
4	GPU	512	2.839296	9.744983	17.577993
5	GPU	1536	3.125681	8.768707	15.744841

32 batch size için iki tabloda da geçen sürelerin yaklaşık aynı oranlarda değiştiri görülmekte.

Fakat yüksek batch size değerlerinde epoch sayısı arttıkça geçen süredeki artışın GPU için daha fazla olduğunu görüyoruz.

Bunu örneklerdirmek gerekirse, batch size 512 için epoch 5 ve epoch 10 arasında TPU yaklaşık 1,5 kat artmışken, GPU yaklaşık 2 kat artmıştır.

Aynı şekilde batch size 1536 için epoch 5 ve epoch 10 arasında TPU yaklaşık 1,3 kat artmışken, GPU yaklaşık 2 kat artmıştır.

Bu sonuçlara bakarak batch size değerini ve epoch sayısını arttıracak olursak, TPU hızının GPU hızına yaklaşabileceği ve geçebileceği yorumunu yapabiliriz. Bunda aynı zamanda kurulan modelin karmaşıklığının ve kullanılan veri setinin boyutunun da önemi olacaktır.