Soru-1

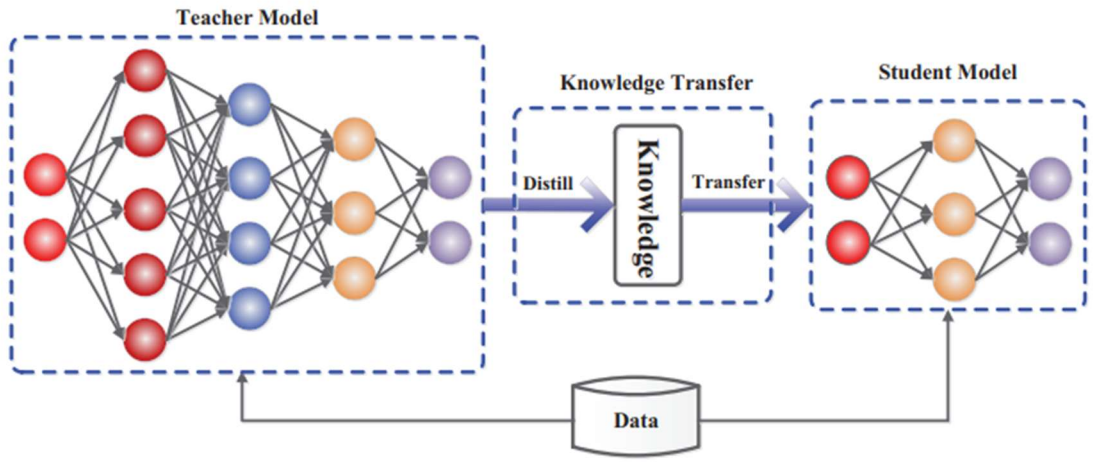
Bu çalışma kapsamında sizden Model küçültme (Knowledge Distillation, parameter sharing, quantisaiton, matrix factorization vb..) teknikleri hakkında 4 sayfalık bir literatür taraması hazırlamanızı istiyoruz.

Günümüzde gelişen dünya ve teknoloji ile birçok projeye imza atılabilmektedir. Bu tekonolojilerin başında yapay zeka ve makine öğrenmesi yöntemleri geliyor. Bu teknolojiler sağlık, finans gibi birçok sektörde kullanılmaktadır. Derin öğrenme modellerinin başarısının arkasında yazılımsal gelişmelerle birlikte yüksek performanslı ekran kartları gibi gelişmiş donanım parçalarının varlığı da çok önemli bir yere sahiptir. Yüksek performans gerektiren bu modelleri daha kısıtlı imkanlar da çalıştırıp aynı başarı oranını elde etmek adına "Model Küçültme" teknikleri kullanılmaktadır. Bu çalışmada da çeşitli model küçültme teknikleri ele alınmıştır.

Knowledge Distillation

Bilgi Distilasyonundaki temel prensip daha önce de kısaca bahsettiğimiz üzere, daha geniş ve kompleks bir model olarak seçilen öğretmen modelin bilgisini, daha hafif ve küçük bir model olarak seçilen öğrenci modele aktarmasıdır. Bu bilgi aktarma işlemi için ilk olarak, önceden eğitilmiş (pre-trained) bir öğretmen modele ihtiyaç duyulmaktadır. Öğretmen modelin, çeşitli işlemlerden geçirilerek öğrenci modele özelliklerini aktarması sağlanır.

Nasıl Çalışır?



Resim. 1 Bilginin damıtılması için genel öğretmen-öğrenci diagramı

Öncelikle, bir sinir ağı modeli genellikle Softmax Fonksiyonu kullanarak çıktılarını olasılık dağılımına çevirir. Softmax, modelin belirli bir sınıfa olan güvenini gösteren bir dağılım üretir. Standart model eğitiminde, gerçek etiketler (ground truth) Hard Predictions (Keskin Tahminler) ile karşılaştırılır. Yani model, doğru olan sınıfa 1, diğerlerine 0 değeri atar. Ancak, Knowledge Distillation sürecinde öğretmen modelin çıktıları doğrudan sert tahminler olarak alınmaz. Bunun yerine, öğretmen modelin ürettiği Soft Labels (Yumuşak Etiketler) kullanılır.

Soft Labels, klasik hard labels yerine öğretmen modelin Softmax sıcaklığı artırılmış çıktılarıdır. Softmax fonksiyonunun sıcaklığı T ile ayarlanarak, çıktıların daha yumuşak ve bilgilendirici hale gelmesi sağlanır. Yüksek sıcaklık değeri, modelin belirsiz olduğu sınıflara da düşük seviyede olasılık atamasını sağlar, böylece öğrenci model hangi sınıfların birbirine daha yakın olduğunu öğrenebilir.

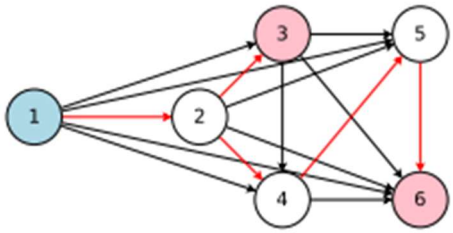
Öğrenci model, hem soft labels'ı hem de klasik ground truth etiketlerini öğrenir. Bu süreçte Distillation Loss (Damıtma Kaybı) hesaplanır. Distillation Loss, öğretmen modelin ürettiği soft labels ile öğrenci modelin tahminleri arasındaki farkı ölçer. Aynı zamanda öğrenci modelin klasik Student Loss (Öğrenci Kaybı) olarak bilinen çapraz entropi kaybı da gerçek etiketlerle kıyaslanarak hesaplanır. Sonuç olarak, öğrenci model bu iki kaybı minimize ederek hem öğretmen modelin bilgilerini alır hem de klasik eğitim sürecini devam ettirir.

Faydaları ve Uygulamaları

Bilgi Distilasyonun birçok faydası ve uyugulama alanı vardır. Cep telefonları veya gömülü sistemler gibi sınırlı hesaplama kaynaklarına sahip uç cihazlarda dağıtıma uygun daha küçük, daha verimli modellerin oluşturulmasına olanak tanır. Daha küçük modeller doğal olarak daha hızlı çıkarım süreleri sağlar; bu da otonom sürüş, robotik süreç otomasyonu (RPA) ve güvenlik sistemleri gibi gerçek zamanlı uygulamalar için gerekli olanakları sağlar. YOLO veya benzer nesne algılama modelleri, uç cihazlardaki gerçek zamanlı uygulamalarda kullanılmak üzere damıtılabilir. Örneğin, akıllı şehirlerde, trafik kavşaklarında doğrudan uç bilişim cihazlarında çalışan verimli trafik izleme ve yönetimi için damıtılmış modeller kullanılabilir.

Parameter Sharing

Bu teknik, modelde kullanılan parametrelerin (ağırlıkların) bir kısmını tekrar kullanarak model boyutunu azaltır. Aynı ağırlıkları farklı katmanlarda veya nöronlar arasında paylaşarak modelin toplam parametre sayısını azaltır. Transformer tabanlı modellerde, katmanlar arasında parametre paylaşımı yapılabilir.



Resim 2. Grafik arama alanının tamamını temsil ederken, kırmızı oklar arama uzayında kararlaştırılan bir modeli tanımlar bir kontrolör tarafından. Burada düğüm 1 modelin girişidir, oysa 3. ve 6. düğümler modelin çıktılarıdır

Faydaları ve Uygulamaları

"Parameter Sharing" tekniğinin birçok fayda sağlar ve uygulama alanı vardır. Avantajları arasında bellek kullanımını ciddi ölçüde azaltır ve hesaplama maliyetlerini düşürerek modelin hızını artırır. Kulllanım alanları arasında transformes modeller, RNN, CNN gibi sinir ağlarında sık kullanılan yöntemdir.

Quantisaiton

Quantisaiton, yapay zeka modellerinin dağıtımında hesaplama ve bellek maliyetlerini azaltmak için makine öğreniminde kullanılan önemli bir optimizasyon tekniğidir. Bir sinir ağının ağırlıklarını ve aktivasyonlarını yüksek hassasiyetli kayan noktalı sayılardan (32 bit kayan noktalı sayılar gibi) 8 bit tam sayılar gibi daha düşük hassasiyetli formatlara dönüştürerek çalışır. Bu işlem model boyutunu önemli ölçüde azaltır ve çıkarım hızını artırır, bu da onu kaynak kısıtlı cihazlarda dağıtım için ideal hale getirir.

Nasıl Çalışır?

Quantisaiton, çeşitli şekiilerde çalışabilmektedir. Post-training Quantisation, eğitilmiş modelin ağırlıklaeı düşük bit seviyesine indirilir. Quantization-Aware Training (QAT), model baştan itibaren kuantize edilmiş ağırlıklarla eğitilir. Dynamic Quantization, model çalıştırma sırasında belirli kısımlar kuantize edilir. Bu şekilde modelin hafıza ihtiyacını ciddi şekilde düşürür. Çalışma hızını arttırarak donanımsal optimizasyan sağlar.

Uygulamalar

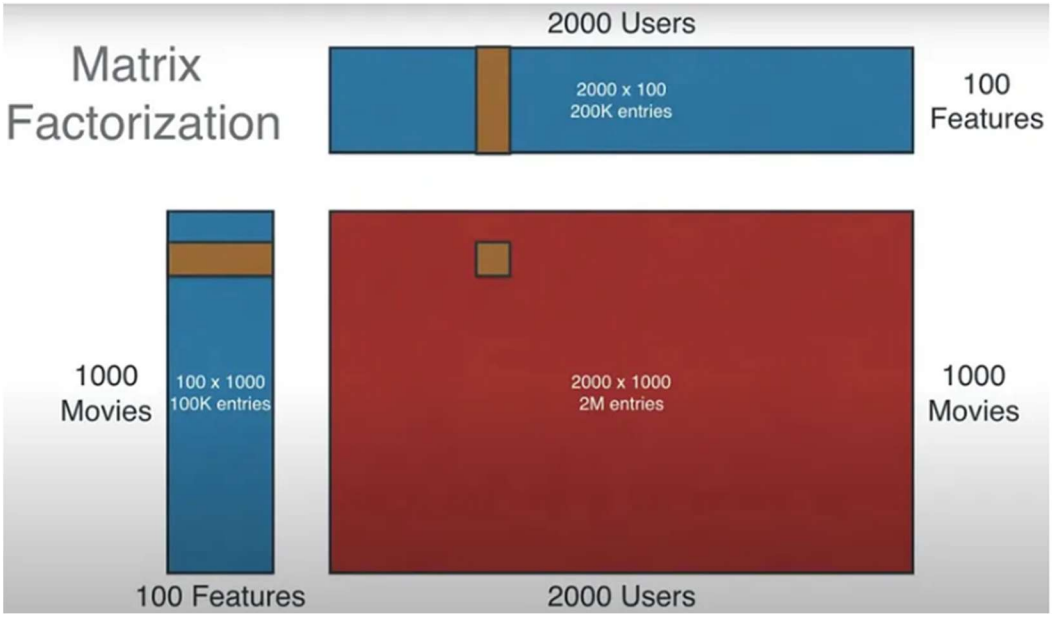
Quantisaiton, çeşitli birçok yerde kullanılmaktadır. Mobil cihazlar, akıllı telefonlar genellikle görüntü tanıma ve doğal dil işleme gibi cihaz yapay zeka özeliikleri için "Quantisaiton" modelleri kullanılır. Akıllı şehirler veya endüstriyel otomasyon gibi senaryolarda, AI modelleri gerçek zamanlı veri işleme için çok sayıda uç cihaza yerleştirilir. Quantisaiton, genellikle sınırlı işlem gücü ve belleğe sahip olan bu cihazlarda verimli model sunumu sağlamak için hayati önem taşır.

Matrix Factorization

Netflix tarafından yapılan açıklamalara göre platformun ücretli abone sayısı, son üç aylık periyotta 1.75 milyon artarak, Mart 2023 itibari ile 232.5 milyon oldu. İçerik sayısının ise dünya çapında 17.000'den fazla olduğu düşünülüyor. Bunu Netflix yapabilmelerini sağlayan şeylerden biride "Matrix Factorization".

Nasıl Çalışır?

Matris Ayrıştırma, büyük ve yüksek boyutlu veri matrislerini daha küçük bileşenlere ayırarak bellek ve hesaplama yükünü azaltan bir tekniktir. Yapay zeka modellerinde, özellikle sinir ağı ağırlıklarının sıkıştırılması, öneri sistemleri ve büyük ölçekli veri analizlerinde kullanılır. Amaç, büyük bir ağırlık matrisini iki veya daha fazla düşük boyutlu matrisin çarpımı şeklinde yeniden ifade ederek modeli küçültmek ve hızlandırmaktır.



Matris Ayrıştırmasında çeşitli yöntemler kullanılmaktadır. Bunlar , SVD - Tekil Değer Ayrıştırması , NMF - Negatif Olmayan Matris Ayrıştırma , PMF , Logistic PMF , Bayesian PMF gibi çeşitli yöntemler kullanılır.

Faydaları ve Uygulamaları

Matris Ayrıştırmasının birçok avantajı ve uygulama alanı bulunmaktadır. Modelin hesaplama yükünü azaltarak bellek kullanımı düşürür ve daha kısa sürede elde edim sağlar. Buna bağlı olarakda düşük donanımlarda çalıştırmayı mümkün kılar. Birçok kulllanım alanı olan Matris Ayrıştırmasının başlıca öneri sistemleri, görüntü işlemei doğal dil işleme gibi birçok alanda kullanıldığı görülmektedir.

Kaynakça

[1]. *https://medium.com/crypttech-research/daha-hafif-daha-h%C4%B1zl%C4%B1-derin-%C3%B6%C4%9Frenme-modellerinde-bilgi-distilasyonu-bed63b76cfcf*

[2]. *https://www.researchgate.net/figure/An-intuitive-example-of-hard-and-soft-targets-for-knowledge-distillation-in-Liu-et-al_fig2_342094012*

[3]. *https://www.ultralytics.com/tr/glossary/knowledge-distillation*

[4]. *https://doi.org/10.48550/arXiv.1802.03268*

[5]. *https://doi.org/10.48550/arXiv.1712.05877*

[6]. *https://www.ultralytics.com/tr/glossary/model-quantization*

[7]. Salakhutdinov, Ruslan & Mnih, Andriy. (2008). Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. Proceedings of the 25th International Conference on Machine Learning. 25. 880-887. 10.1145/1390156.1390267.

[8]. https://mustafaserdarkonca.medium.com/matrix-factorization-part-1-%C3%B6neri-sistemleri-ccd77d3b18b4