Model Küçültme Literatür Taraması

Giriş

Yapay zeka (YZ) ve makine öğrenimi (MÖ) modelleri, son on yılda teknolojik gelişmelerin ve veri bilimi uygulamalarının ön saflarına yerleşmiştir. Bu modeller, tıbbi teşhislerden finansal tahminlere, görüntü tanımadan doğal dil işlemeye kadar geniş bir uygulama yelpazesine sahiptir. Bu modellerin oluşturduğu derin öğrenme ağlarının karmaşıklığı, modellerin kullanılacağı cihaz üzerinde önemli bir yük oluşturabilir. Model küçültme, bu zorluğu ele almanın ve MÖ modellerini daha geniş bir kullanıcı kitlesine ve daha geniş bir cihaz skalasına erişilebilir kılmak için kullanılabilen stratejik bir yaklaşımdır.

Model küçültme, genellikle büyük ve hesaplama açısından yoğun modelleri, daha az hesaplama gücü gerektiren, daha hafif ve daha küçük modellere dönüştürme sürecini ifade eder. Bu yöntem, işlem kapasitesi güçlü olmayan cihazlarda modelleri konuşlandırma ihtiyacını gidermeye yöneliktir. Model küçültme teknikleri, modelin boyutunu azaltırken, olabildiğince çok bilgiyi ve öğrenme kapasitesini koruma amacını güder.

Birkaç popüler model küçültme tekniği vardır: Bilgi Damıtma, Parametre Paylaşımı, Kuantizasyon, Matris Faktörizasyonu vb. Her teknik, modelin hesaplama ve bellek kullanımını azaltma yaklaşımını farklı bir açıdan ele alır. Örneğin, Bilgi Damıtma, bir "öğretmen" modelin bilgisini daha küçük bir "öğrenci" modeline aktarırken, Kuantizasyon, model ağırlıklarının hassasiyetini azaltarak model boyutunu küçültür.

Model küçültme, yapay zeka ve makine öğrenimi modellerinin gelecekteki uygulanabilirliği ve erişilebilirliği açısından kritik bir öneme sahiptir. Bu alandaki araştırmalar, modelleri daha verimli, erişilebilir ve sürdürülebilir kılmayı amaçlamaktadır, böylece yapay zeka, daha geniş bir kullanıcı kitlesi ve çeşitli platformlar tarafından kullanılabilir hale gelir.

Bilgi Damıtma Tekniği

Bilgi Damıtma (KD), daha büyük bir modelden (öğretmen) daha küçük bir modele (öğrenci) bilgi aktarmayı esas alan bir model sıkıştırma tekniğidir. Hinton ve arkadaşlarının öncü çalışması, "karanlık bilgiyi" öğretmen modelden öğrenci modele yumuşatılmış çıktı olasılıkları aracılığıyla aktarmanın önemini vurgulayarak KD kavramını tanıtmıştır (Hinton ve ark., 2015). Bu teknik, büyük modellerin hesaplama açısından uygulanabilir olmadığı senaryolarda, örneğin mobil cihazlarda, kenar cihazlarda veya diğer kaynak sınırlı ortamlarda, özellikle kritiktir.

Yaygın uygulaması ve başarısına rağmen, KD'nin karşılaştığı zorluklar yok değildir. Ana zorluklardan biri, model boyutu ile tahmin performansı arasındaki dengeyi yönetmektir. Modelin boyutunu azaltmak bazen tahmin doğruluğunda bir düşüşe yol açabilir, bu da damıtılmış modelin gerçek dünya uygulamalarında hala etkili olmasını sağlamak için dikkatlice yönetilmelidir. Ayrıca öğretmen modelin esas tahmin kabiliyetlerini koruyacak şekilde öğretmen modelden öğrenci modele bilgiyi aktarma süreci de basit olmayan bir konudur ve aktif bir araştırma alanıdır.

Parametre Paylaşımı Tekniği

Parametre paylaşımı, özellikle sinir ağlarında, modelin performansını önemli ölçüde bozmadan hesaplama ve bellek gereksinimlerini azaltmak için kullanılan bir tekniktir. Parametre paylaşımının temel fikri, model içindeki birden fazla birim veya katman için ortak bir parametre (veya ağırlık) seti kullanmaktır, böylece saklanması ve hesaplanması gereken toplam parametre sayısı azaltılır. Parametre paylaşımının klasik bir örneği, Evrişimli Sinir Ağlarında (CNN) gözlemlenir, burada aynı filtre tüm giriş alanına uygulanır, bu da her filtrenin girişte herhangi bir yerde algılanabilen bir özellik öğrenmesini sağlar, böylece modelin karmaşıklığını azaltır ve genelleme yeteneğini artırır.

Parametre paylaşımını farklı bağlamlar ve uygulamalar üzerinde de etkili bir şekilde uygulamak için çeşitli metodolojiler ve yaklaşımlar geliştirilmiştir. Örneğin, "Parametre Paylaşımı ile Etkili Sinirsel Mimari Araması" üzerine yapılan bir çalışma hem hızlı hem de ucuz olan otomatik model tasarımı için bir yöntem sunar, tüm yöntemler arasında post-egitim işleme olmaksızın yeni bir standart belirler ve çok daha az GPU-saat kullanarak güçlü deneysel performanslar sunar (Pham ve ark., 2018). Başka bir dikkate değer yaklaşım olan "Çok-Görevli Öğrenmede Görev Tabanlı Parametre Paylaşımı", bir temel modeli, görev özel bir katman alt kümesini adaptif olarak değiştirerek yeni bir göreve uyarlar, uygulaması basit olmasına rağmen en iyi performansı elde eder. (Wallingford ve ark., 2022) Bu metodolojiler, parametre paylaşımının çeşitli model mimarileri ve uygulamalarındaki çok yönlülüğünü ve etkinliğini vurgular.

Parametre paylaşımı, geniş bir alan ve görev yelpazesi üzerinde uygulama bulmuştur, bunlar arasında görüntü sınıflandırma, nesne algılama ve çok görevli öğrenme bulunmaktadır. Örneğin, "Seçici Parametre Paylaşımı Tekniği ile Çok Ajanlı Güçlendirilmiş Öğrenme Ölçeklemesi" adlı çalışma, parametre paylaşımından fayda sağlayabilecek ajanları otomatik olarak belirlemek için bir yöntem önerir, parametre paylaşımının artan örnek verimliliğini, eğitim süresini azaltmak ve sonuçları artırmak için birden fazla bağımsız ağın gösterim kapasitesi ile birleştirir (Christianos ve ark., 2021).

Ancak avantajlarına rağmen, parametre paylaşımı kendi zorluklarıyla birlikte gelir. Anahtar zorluklardan biri, hangi parametrelerin paylaşılması gerektiğini ve bu paylaşımın, modelin görülmemiş verilere iyi bir şekilde genelleme yeteneğini korumasını sağlamak üzere nasıl yapılandırılması gerektiğini belirlemektir. Parametre azaltma ile model performansını koruma arasında bir denge sağlamak hayati öneme sahiptir ve çeşitli görevler ve alanlar üzerinde parametreleri adaptif ve etkili bir şekilde paylaşma yöntemlerini geliştirmek aktif bir araştırma alanıdır.

Kuantizasyon Tekniği

Derin öğrenme bağlamında kuantizasyon, bir sinir ağındaki ağırlıkların ve aktivasyonların olası değerlerini, genellikle model çıkarımının bellek ayak izini ve hesaplama maliyetini önemli ölçüde azaltan ayrık bir kümeye sınırlama sürecini ifade eder. Han ve arkadaşlarının öncü çalışması "Derin Sıkıştırma: Budama, Eğitimli Niceleme ve Huffman Kodlama ile Derin Sinir Ağı Sıkıştırma", derin sıkıştırma için üç aşamalı bir boru hattı tanıtmış; budama, eğitimli kuantizasyon ve Huffman kodlama, sinir ağlarının depolama gereksinimlerini doğruluklarını etkilemeden 35 ile 49 kat arasında azaltma başarısını göstermiştir (Han ve ark., 2016).

Kuantizasyon, model boyutu ve hesaplama verimliliği açısından belirgin avantajlar sunarken, özellikle model doğruluğunu kuantizasyon sonrası koruma ile ilgili zorluklar getirir.

Ağırlıkların ve aktivasyonların sayısal hassasiyetindeki azalma, genellikle dikkate alınması gereken model performansında bir bozulmaya yol açabilir. Ayrıca, kuantizasyon için optimal bit genişliğini belirleme ve eğitim veya doğrulama verilerine erişim olmaksızın kuantizasyon sürecini yönetme, alandaki araştırmacılar ve uygulayıcılar için basit olmayan zorluklardır.

Uygulamalar alanında, kuantizasyon, hesaplama kaynakları ve bellek sınırlı olan cihazlarda, örneğin İnternet Nesneleri (IoT) cihazları, cep telefonları ve kenar cihazları gibi, derin öğrenme modellerinin konuşlandırılmasını sağlamada kilit bir rol oynamıştır. Örneğin, "BRECQ: Post-Training Quantization'ın Limitini Blok Yeniden Yapılandırma ile Zorlama" çalışması, ilk kez INT2 bit genişliğine kadar verimli ve esnek kuantizasyonu mümkün kılan yeni bir Post-Training Quantization (PTQ) çerçevesi tanıtmış ve PTQ'nun, Quantization-Aware Training (QAT) ile karşılaştırılabilir 4-bit ResNet ve MobileNetV2'yi elde edebileceğini ve kuantize modellerin üretiminde 240 kat daha hızlı olabileceğini kanıtlamıştır (Li ve ark., 2021).

Martis Faktarizasyonu Tekniği

Matris Faktörizasyonu (MF), özellikle öneri sistemlerindeki uygulamalarıyla bilinen, makine öğrenimi ve veri madenciliği alanında geniş bir tanınırlığa sahip bir tekniktir. MF'nin temel ilkesi, büyük bir matrisi daha küçük matrislere ayırmaktır, bu da orijinal matrisi daha az hesaplama karmaşıklığı ile yaklaştırmak için kullanılabilir. MF'nin klasik bir uygulaması olan Netflix Ödül yarışması, matris faktörizasyonu modellerinin, ek bilgilerin dahil edilmesine izin vererek, ürün önerileri üretme konusunda klasik en yakın komşu tekniklerinden üstün olduğunu göstermiştir (Koren ve ark., 2009).

Matris Faktörizasyon teknikleri genellikle orijinal matrisin düşük sıralı bir yaklaşımını arar, bu da Tekil Değer Ayrıştırması (SVD) gibi çeşitli yöntemlerle elde edilebilir. SVD, bir matrisi üç başka matrise ayıran popüler bir matris faktörizasyon yöntemidir ve düşük sıralı yaklaşımlar oluşturma temelini sağlar. Öneri sistemleri bağlamında, SVD ve varyantları, etkili ve doğru modeller oluşturma konusunda belirleyici olmuştur. Örneğin, Koren ve arkadaşlarının çalışması, öneri sistemlerinde işbirlikçi filtreleme için matris faktörizasyonunu kullanmış ve Netflix Ödül yarışmasında etkinliğini göstermiştir (Koren ve ark., 2009). Ayrıca, matris girişlerini negatif olmayan olarak sınırlayan Negatif Olmayan Matris Faktörizasyonu (NMF) gibi varyantlar da çeşitli alanlarda, yüz parçalarını ve metinlerin semantik özelliklerini öğrenme gibi, araştırılmış ve uygulanmıştır (Lee ve ark., 1999) (Lee ve ark., 2000).

Matris Faktörizasyonu, özellikle öneri sistemlerinin geliştirilmesinde özellikle etkili olmuştur. Örneğin, Olasılıksal Matris Faktörizasyonu (PMF), gözlem sayısı ile doğrusal olarak ölçeklenir ve Netflix veri kümesi gibi büyük, seyrek ve çok dengesiz veri kümelerinde iyi performans gösterir (Salakhutdinov ve ark., 2007). Ayrıca, MF, semantik özellik öğrenme gibi çeşitli diğer uygulamalarda da kullanılmıştır, burada NMF için algoritmalar, yüz parçalarını ve metinlerin semantik özelliklerini öğrenme yetenekleri için analiz edilmiştir.

Öneri sistemlerinde kullanıcı-öğe etkileşim verilerinin genellikle seyrek ve dengesiz olması gibi, MF ile ilgili olarak veri seyrekliği ile başa çıkma, hesaplama karmaşıklığını yönetme ve gürültülü veriler karşısında sağlamlığı sağlama gibi zorluklar devam etmektedir.

KAYNAKÇA

Christianos, F., Papoudakis, G., Rahman, M. A., & Albrecht, S. V. (2021, July). Scaling multi-agent reinforcement learning with selective parameter sharing. In International Conference on Machine Learning (pp. 1989-1998). PMLR.

Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. arXiv preprint arXiv:1510.00149.

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788-791.

Lee, D., & Seung, H. S. (2000). Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 13.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., ... & Gu, S. (2021). Brecq: Pushing the limit of post-training quantization by block reconstruction. arXiv preprint arXiv:2102.05426.

Mnih, A., & Salakhutdinov, R. R. (2007). Probabilistic matrix factorization. Advances in neural information processing systems, 20.

Pham, H., Guan, M., Zoph, B., Le, Q., & Dean, J. (2018, July). Efficient neural architecture search via parameters sharing. In International conference on machine learning (pp. 4095-4104). PMLR.

Wallingford, M., Li, H., Achille, A., Ravichandran, A., Fowlkes, C., Bhotika, R., & Soatto, S. (2022). Task adaptive parameter sharing for multi-task learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7561-7570).