

Transformer Mimarisi

2017 yılında Google'ın "Attention Is All You Need" başlıklı çığır açan makalesiyle tanıtılan Transformer mimarisi, doğal dil işleme (NLP) ve yapay zeka alanında bir devrim yaratmıştır. Özyinelemeli sinir ağlarının (RNN) ve Evrişimli Sinir Ağlarının (CNN) uzun süreli bağımlılıkları modellemedeki sınırlamalarını aşmak üzere tasarlanan bu mimari, yalnızca "Özdikkat (Self-Attention)" mekanizmasına dayanmaktadır. Bu araştırma yazısı, Transformer mimarisinin temel bileşenlerini, çalışma prensibini, getirdiği avantajları ve modern dil modelleri üzerindeki dönüştürücü etkisini detaylı bir şekilde incelemektedir.

1. Giriş: Önceki Mimari ve Sınırlamalar

Transformer'dan önce, dizi modelleri (sequence modeling) ve özellikle makine çevirisi, özyinelemeli sinir ağları (RNN'ler) ve onların gelişmiş bir varyantı olan Uzun Kısa-Vadeli Bellek (LSTM) ağları ile domine edilmişti. RNN'ler, girdiyi sıralı olarak işler, her bir adımda bir gizli durum (hidden state) üretir ve bir sonraki adıma aktarır. Bu yaklaşımın iki temel dezavantajı vardır:

- 1. Sıralı İşlem (Sequential Processing):** Bir diziyi işlemek için adım adım ilerleme zorunluluğu, modern GPU'lar gibi paralel işlemcilerin gücünden tam olarak yararlanılmasını engeller. Bu, eğitim sürelerini önemli ölçüde yavaşlatır.
- 2. Uzun Menzilli Bağımlılık Sorunu (Long-Range Dependency Problem):** Uzun dizilerde, erken bir adımdaki bilgi, birçok adım boyunca taşınmak zorunda kalır ve zamanla ya bozulur ya da kaybolur. Bu, bir paragrafın başındaki bir kelimenin sonundaki bir kelimeyle olan ilişkisinin kurulamamasına neden olabilir.

Bu sorunları hafifletmek için "Dikkat (Attention)" mekanizmaları geliştirilmiş olsa da, bu mekanizmalar genellikle RNN çekirdeğine bir eklenti olarak kalmıştır. Transformer ise bu paradigmayı tamamen tersine çevirerek, RNN'yi tamamen ortadan kaldırmış ve dikkati modelin merkezine yerleştirmiştir.

2. Transformer Mimarisi: Temel Bileşenler

Transformer, bir Kodlayıcı (Encoder) ve bir Kodçözücü (Decoder) olmak üzere iki ana kısımdan oluşan bir mimaridir. Her iki taraf da birbirinin aynı olan çok sayıda

katmandan (Nx) yığılmıştır.

2.1. Girdi ve Konumsal Kodlama (Positional Encoding)

RNN'lerin aksine, Transformer girdiyi paralel olarak (tüm kelimeleri aynı anda) işlediği için, kelimelerin dizideki sırasına dair bilgiyi kaybeder. Bu sorunu çözmek için, **Konumsal Kodlama**, her bir kelimenin gömme vektörüne (embedding vector) eklenir. Bu kodlama, sinüs ve kosinüs fonksiyonları kullanılarak her bir konum için benzersiz bir vektör üretir. Böylece model, aynı kelimenin farklı pozisyonlarda farklı anlamlara gelebileceğini öğrenebilir.

2.2. Özdiğer (Self-Attention) Mekanizması

Mimarının kalbinde yer alan bu mekanizma, bir dizideki her bir kelimenin, aynı dizideki diğer tüm kelimelerle olan ilişkisini hesaplamasını sağlar.

- **Sorgu, Anahtar, Değer (Query, Key, Value):** Her bir kelime için üç farklı vektör öğrenilir: Sorgu (Q), Anahtar (K) ve Değer (V). Bu vektörler, girdi vektörlerinin öğrenilmiş ağırlık matrisleri (W^Q , W^K , W^V) ile çarpılmasıyla elde edilir.
- **Dikkat Skoru Hesaplama:** Bir kelimenin diğer tüm kelimelerle olan ilişki katsayısı (dikkat skoru), onun Sorgu vektörü ile diğer kelimelerin Anahtar vektörlerinin nokta çarpımı ile hesaplanır. Bu skorlar, bir softmax fonksiyonundan geçirilerek normalize edilir ve ağırlıklandırılmış bir "dikkat" dağılımı oluşturulur.
- **Çıktı:** Son çıktı, bu ağırlıkların her bir kelimenin Değer (V) vektörleriyle çarpılması ve toplanmasıyla elde edilir. Sonuç olarak, her bir kelimenin yeni temsili (çıkı vektörü), tüm diziden ilgili bilgileri içeren bir bağlamsal temsildir.

Çok Kafalı Dikkat (Multi-Head Attention), modelin aynı anda dizinin farklı bölgelerine odaklanabilmesini sağlar. Farklı öğrenilmiş ağırlık matrisleri kullanılarak hesaplanan birden fazla dikkat katmanının (kafaların) çıktıları birleştirilir. Bu, bir kelimenin sözdizimsel (bir kafa), anlamsal (başka bir kafa) vb. farklı yönlerden ilişkilerini aynı anda modelleyebilme esnekliği sağlar.

2.3. İleri Beslemeli Ağlar (Feed-Forward Networks)

Özdiğer katmanından sonra, her alt katmanda bir de İleri Beslemeli Sinir Ağı bulunur. Bu ağ, her bir konumdaki temsili bağımsız olarak dönüştürür. Genellikle bir doğrusal dönüşüm, bir ReLU aktivasyonu ve bir başka doğrusal dönüşümden oluşur.

2.4. Artık Bağlantılar ve Katman Normalleştirme (Residual Connections & Layer Normalization)

Mimarideki her bir alt katman (öz dikkat veya ileri beslemeli ağ), bir **Artık Bağlantı (Residual Connection)** ile çevrilidir ve ardından **Katman Normalleştirme (Layer Norm)** uygulanır. Bu, derin ağların eğitimi stabilize eder ve gradyan kaybolması (vanishing gradient) problemini hafifletir. Artık bağlantı, girdiyi katmanın çıktısına ekleyerek orijinal bilginin korunmasını sağlar.

3. Kodlayıcı-Kodçözücü Etkileşimi ve Çıktı Üretimi

Kodlayıcı, girdi dizisinin bir temsilini oluşturur. Kodçözücü ise, çıktı dizisini oluştururken (örneğin, çevrilmiş cümle) bu temsili kullanır. Kodçözücüdeki ilk dikkat katmanı **maskelenmiş çok kafalı dikkattir**. Bu maskeleye, eğitim sırasında modelin bir sonraki kelimeyi tahmin ederken yalnızca önceki kelimelere (gelecekteki kelimelere değil) bakmasını sağlar, böylece "hile yapmasını" engeller.

Kodçözücünün ikinci dikkat katmanı, **Kodlayıcı-Kodçözücü Dikkat** katmanıdır. Burada, Kodçözücünden gelen Sorgular (Q), Kodlayıcının çıktısından gelen Anahtarlar (K) ve Değerler (V) ile etkileşime girer. Bu mekanizma, Kodçözücünün, çıktının her bir adımında, girdi dizisinin hangi bölümlerine en çok odaklanması gerektiğini ("dikkat etmesi gerektiğini") öğrenmesini sağlar.

4. Avantajlar ve Etkisi

- **Yüksek Paralellik:** Tüm girdi dizisi aynı anda işlenebildiği için eğitim süreleri büyük ölçüde kısalmıştır.
- **Üstün Performans:** Uzun menzilli bağımlılıkları çok daha etkili bir şekilde yakalayarak, çeşitli NLP görevlerinde yeni state-of-the-art sonuçlar elde edilmiştir.
- **Ölçeklenebilirlik:** Donanım kısıtlamaları dışında, daha büyük veri kümeleri ve daha fazla parametre ile kolayca ölçeklenebilir.

Transformer'ın bu avantajları, onu GPT (Generative Pre-trained Transformer), BERT (Bidirectional Encoder Representations from Transformers) ve günümüzün devasa Büyük Dil Modelleri (LLM'ler) gibi birçok devrimci modelin temel yapı taşı haline getirmiştir. Etkisi yalnızca NLP ile sınırlı kalmamış, bilgisayarlı görü (Vision Transformers - ViT) ve ses sentezi gibi alanlara da yayılmıştır.

5. Sonuç

Transformer mimarisi, yapay zeka tarihindeki en önemli mimari yeniliklerden biridir. Özdikkat mekanizmasına dayalı basit ama güçlü tasarımı, alanda bir paradigma değişimine öncülük etmiş ve modern dil modellerinin ve yapay zeka uygulamalarının bugünkü yeteneklere ulaşmasının önünü açmıştır. Paralel işleme gücü, bağımlılık yakalama kapasitesi ve ölçeklenebilirliği ile önümüzdeki yıllarda da temel bir araştırma ve uygulama alanı olmaya devam edeceği açıktır.
