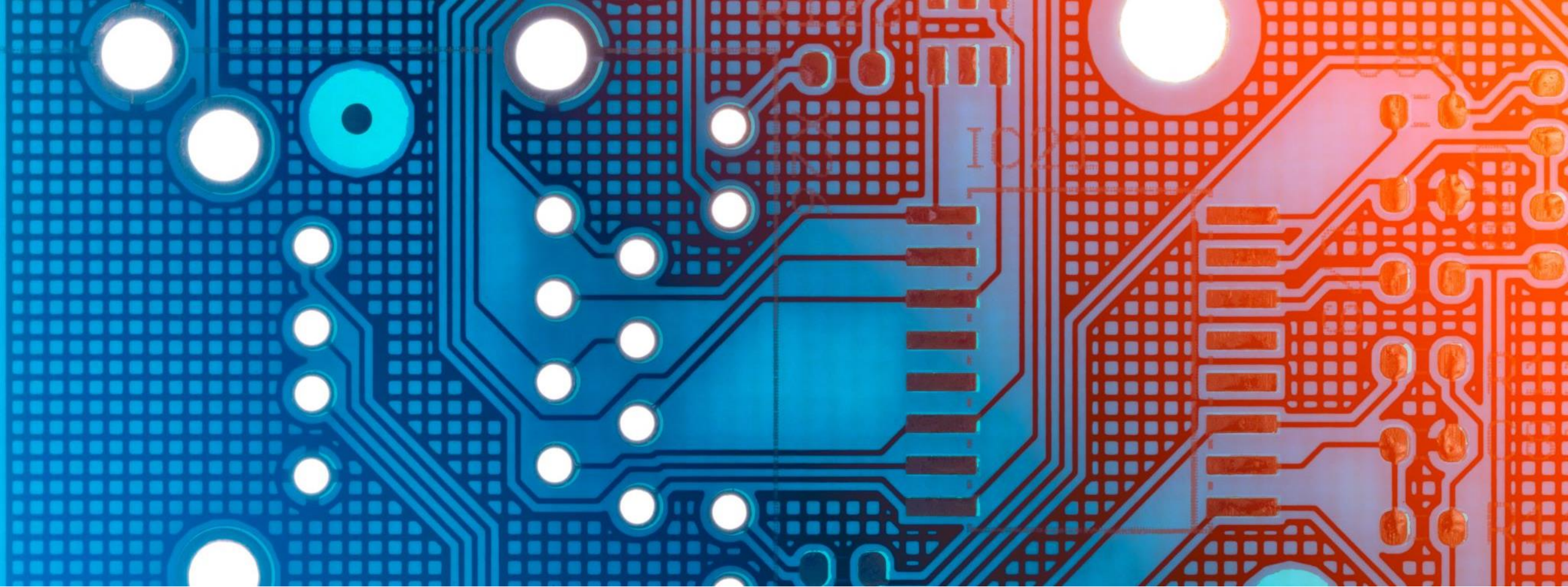




Ders 6

Transformers Mimarisi



Ana Hat

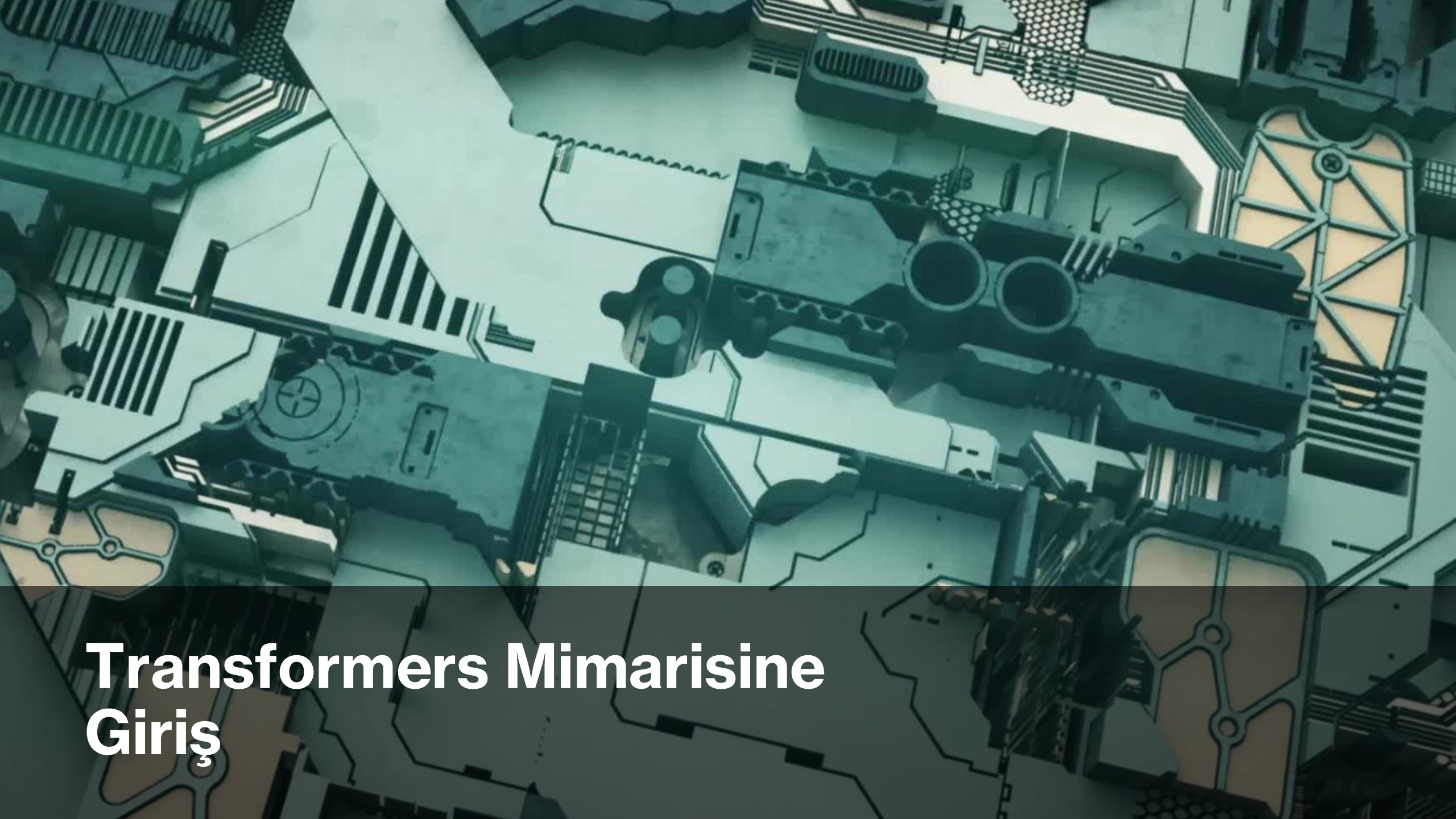
Transformers Mimarisine Giriş

Attention (Dikkat) Mekanizması

Positional Encoding (Konum Kodlaması)

Encoder-Decoder Yapısı

Transformers'in Uygulama Alanları

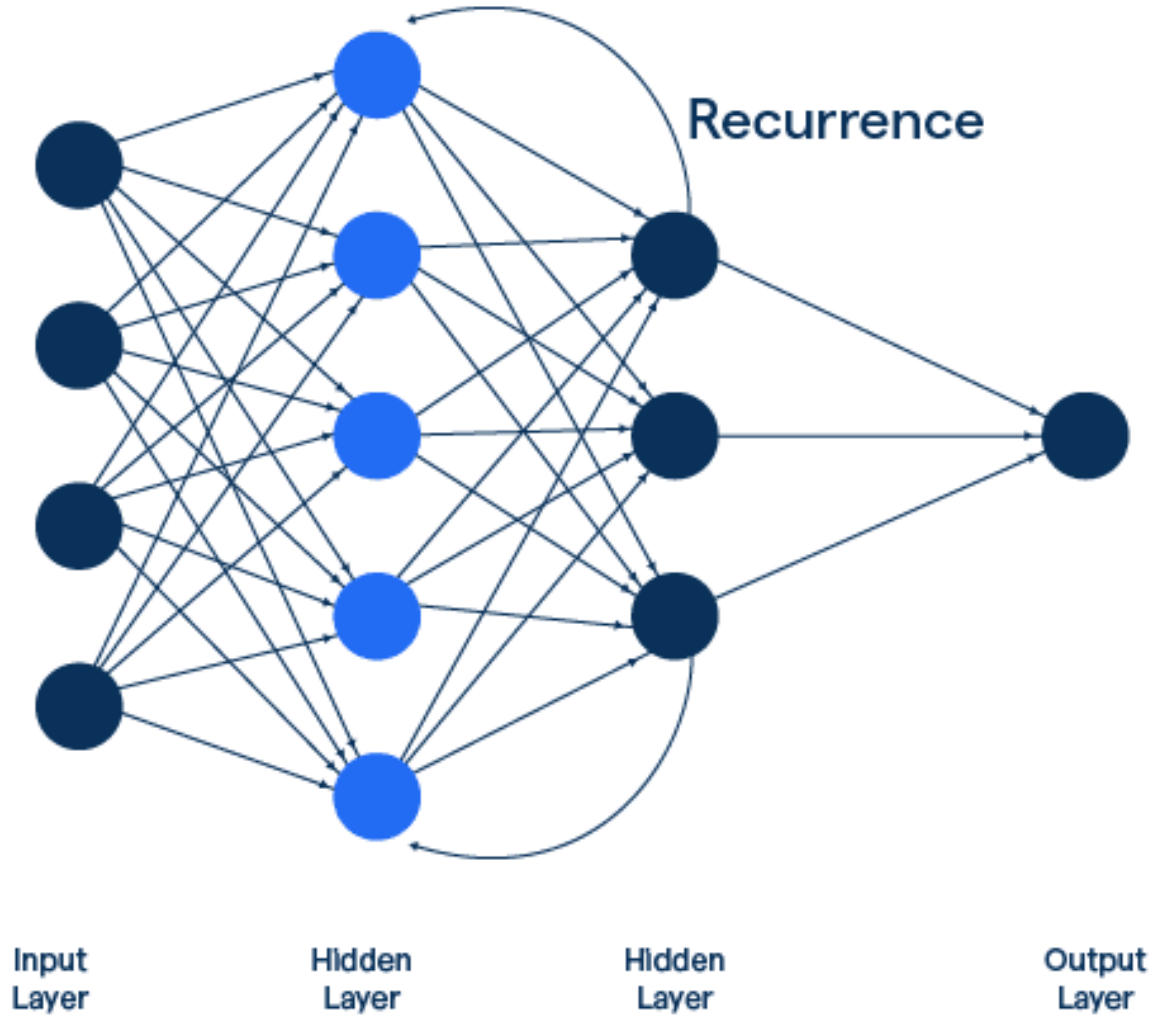


Transformers Mimarisine Giriş

Geleneksel NLP Modellerinin Sınırları

Geleneksel doğal dil işleme (NLP) modelleri, özellikle döngüsel sinir ağları (RNN) ve türevleri (LSTM, GRU), sıralı verileri işlemek için tasarlanmıştır. Bu modeller, metindeki kelimeleri veya sembolleri birer birer işler ve her adımda önceki adımların bilgisini kullanır.

Recurrent Neural Network



Geleneksel NLP Modellerinin Sınırları

- **Uzun Vadeli Bağımlılıkları Öğrenmede Zorluk:** RNN tabanlı modeller, uzun metinlerde veya uzun vadeli bağımlılıklara sahip cümlelerde performans kaybı yaşar.
- **Paralel İşleme Eksikliği:** RNN'ler verileri sıralı olarak işlediği için, hesaplama işlemleri paralelize edilemez. Bu durum, özellikle büyük veri kümeleri üzerinde eğitim yaparken zaman ve kaynak verimliliğini düşürür.
- **Kontekst Bilgisinin Kısıtlılığı:** Geleneksel modeller, genellikle yakın çevredeki kelimelerle sınırlı bir kontekst anlayışına sahiptir. Uzaktaki kelimeler arasındaki ilişkileri yakalamakta zorlanırlar



Transformers'ın Ortaya Çıkışı ve Önemi

2017 yılında Vaswani ve arkadaşları tarafından yayınlanan "Attention is All You Need" makalesiyle Transformer mimarisi tanıtıldı. Bu mimari, NLP alanında devrim oluşturarak birçok görevi yerine getiren modellerin temelini oluşturdu.

- **Attention Mekanizmasının Merkezi Rolü:** Transformers, geleneksel RNN veya CNN yapıları yerine tamamen attention mekanizmasına dayanır. Bu sayede, her kelimenin dizideki diğer tüm kelimelerle olan ilişkisini doğrudan modelleyebilir.
- **Paralel İşlemeye Uygunluk:** Transformers, sıralı işlem gerektirmeyen yapısı sayesinde GPU'lar üzerinde paralel işlemeyi mümkün kılar. Bu, eğitim ve çıkarım süreçlerini önemli ölçüde hızlandırır.
- **Uzun Vadeli Bağımlılıkları Etkili Şekilde Öğrenme:** Self-attention mekanizması sayesinde model, uzak kelimeler arasındaki ilişkileri daha iyi yakalar ve uzun metinlerde daha başarılı sonuçlar üretir.
- **NLP Uygulamalarında Üstün Performans:** Transformer tabanlı modeller, makine çevirisinden metin özetlemeye, soru-cevap sistemlerinden duygu analizine kadar birçok alanda çok iyi sonuçlar elde etmiştir.
- **Büyük Dil Modellerinin Temeli:** GPT, BERT, T5 gibi önde gelen büyük dil modelleri, Transformer mimarisini temel alır ve NLP alanındaki ilerlemeleri hızlandırır.



Transformers'ın Temel Bileşenleri

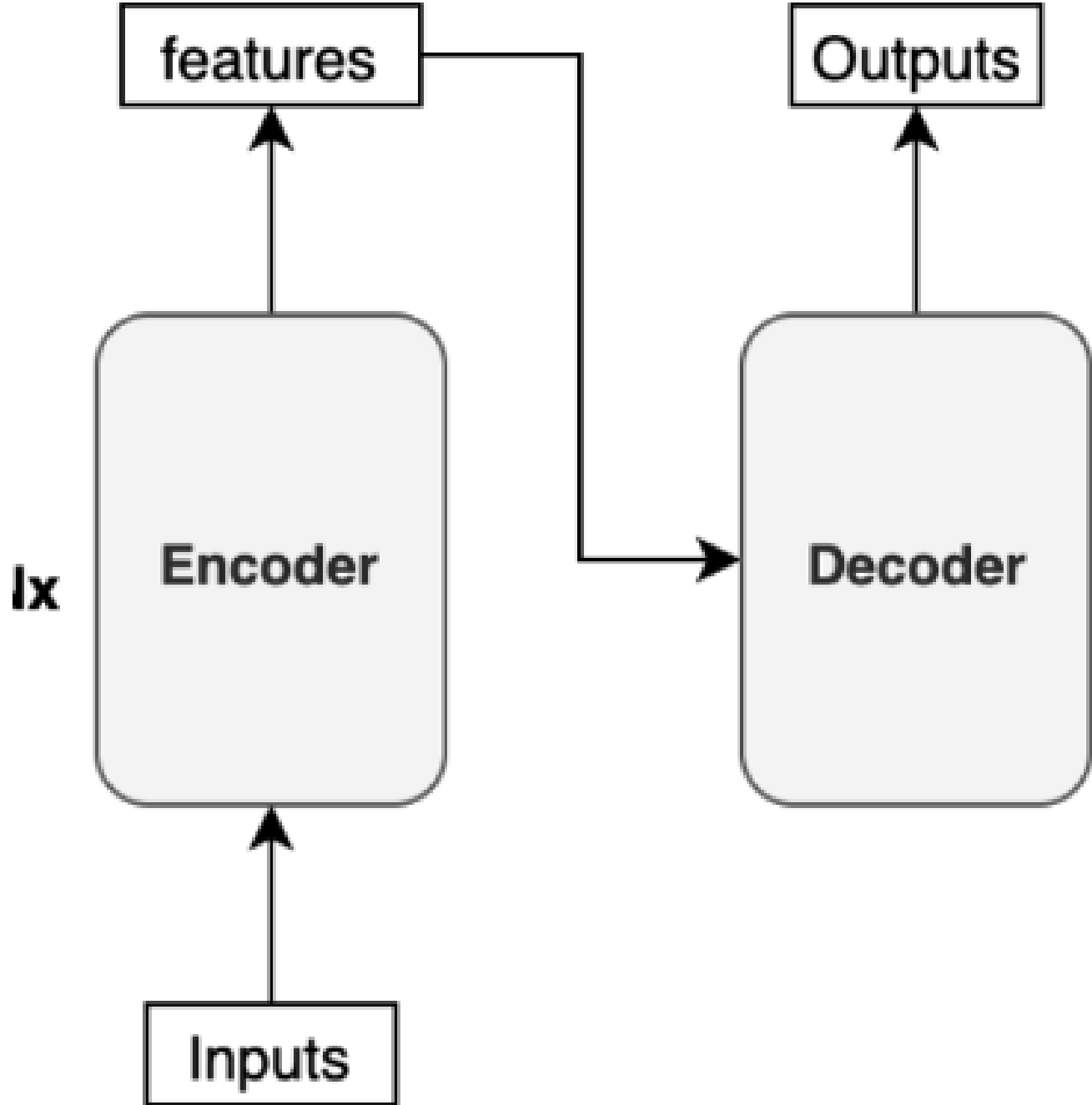
- Transformer mimarisi, karmaşık görünen ancak mantıksal olarak anlaşılması mümkün olan birkaç temel bileşenden oluşur:

Encoder ve Decoder Yapıları:

- Encoder:** Girdi dizisini alır ve her bir ögenin (örneğin kelimenin) temsilini zenginleştirir. Birden fazla encoder katmanından oluşur ve her katman, önceki katmanın çıktısını alarak daha yüksek seviyeli özellikler öğrenir.
- Decoder:** Encoder'dan gelen bilgiyi kullanarak hedef çıktıyı üretir. Aynı şekilde birden fazla decoder katmanından oluşur.

Self-Attention Mekanizması:

- Tanım:** Modelin, bir dizideki bir kelimenin diğer tüm kelimelerle olan ilişkisini öğrenmesine olanak tanır.
- İşlevi:** Her kelime için, diğer kelimelerin o kelimeyle ne kadar ilgili olduğunu belirler ve bu bilgiyi kullanarak kelimenin temsilini günceller.
- Avantajı:** Uzun vadeli bağımlılıkları ve farklı kontekstleri etkili bir şekilde yakalar.



Transformers'ın Temel Bileşenleri

Multi-Head Attention:

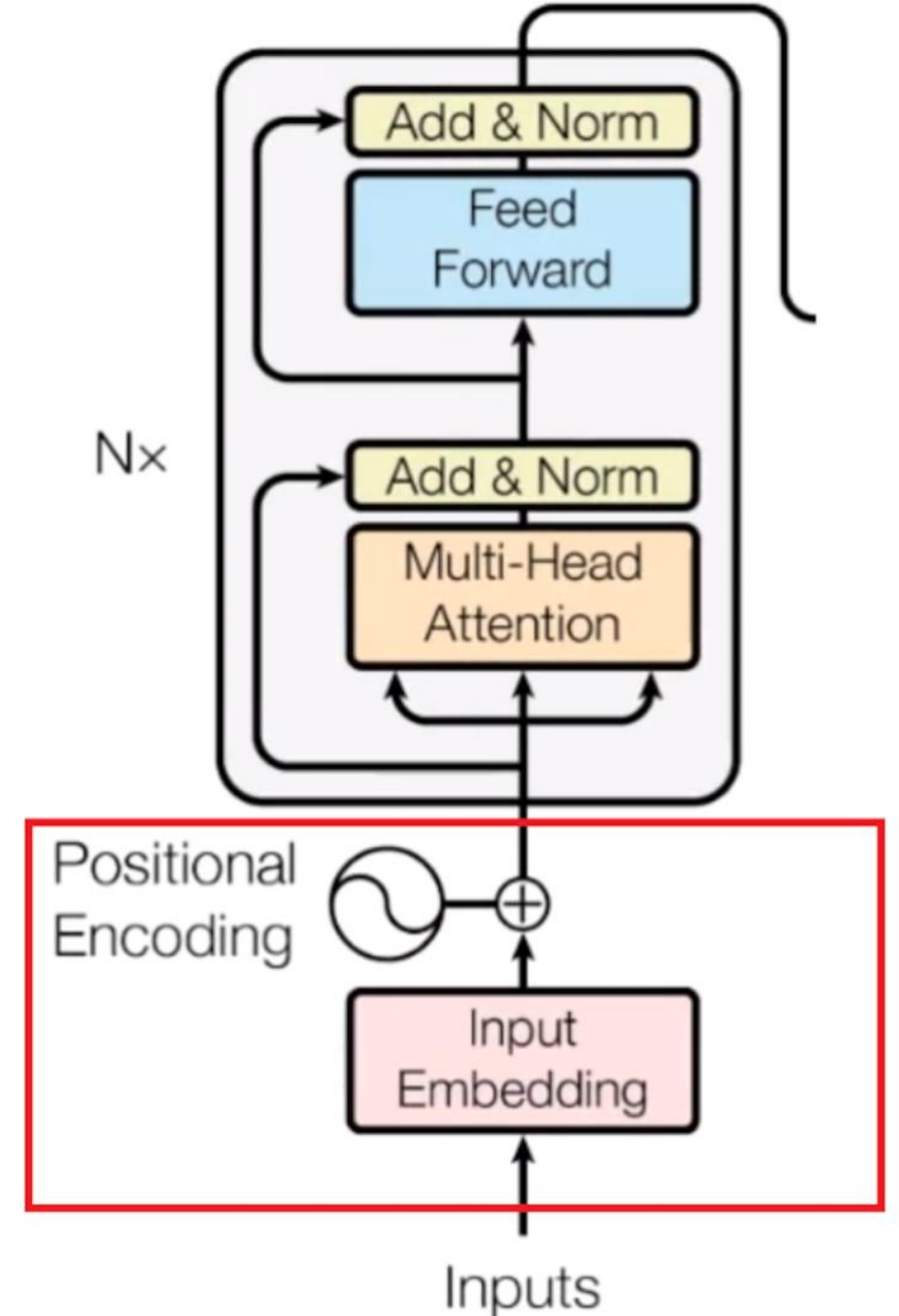
- **Tanım:** Attention hesaplamalarının birden fazla başlık (head) üzerinden paralel olarak yapılmasıdır.
- **İşlevi:** Farklı başlıklar, farklı ilişki türlerini veya özellikleri öğrenebilir. Bu, modelin daha zengin ve çeşitli temsil becerisi kazanmasını sağlar.

Positional Encoding (Konum Kodlaması):

- **Sorun:** Transformers, sıralı işlem yapmadığı için kelimelerin pozisyon bilgisini doğrudan bilemez.
- **Çözüm:** Positional encoding ile her kelimenin konum bilgisi, kelime temsillerine eklenir. Genellikle sinüs ve kosinüs fonksiyonları kullanılarak gerçekleştirilir.
- **Sonuç:** Model, kelimelerin dizideki sırasını ve bunun anlam üzerindeki etkisini öğrenebilir.

Feed-Forward Sinir Ağları:

- **Tanım:** Her encoder ve decoder katmanında bulunan, lineer dönüşümler ve aktivasyon fonksiyonlarından oluşan iki katmanlı tam bağlantılı ağlardır.
- **İşlevi:** Non-lineerlik katarak modelin daha karmaşık ilişkileri öğrenmesini sağlar.



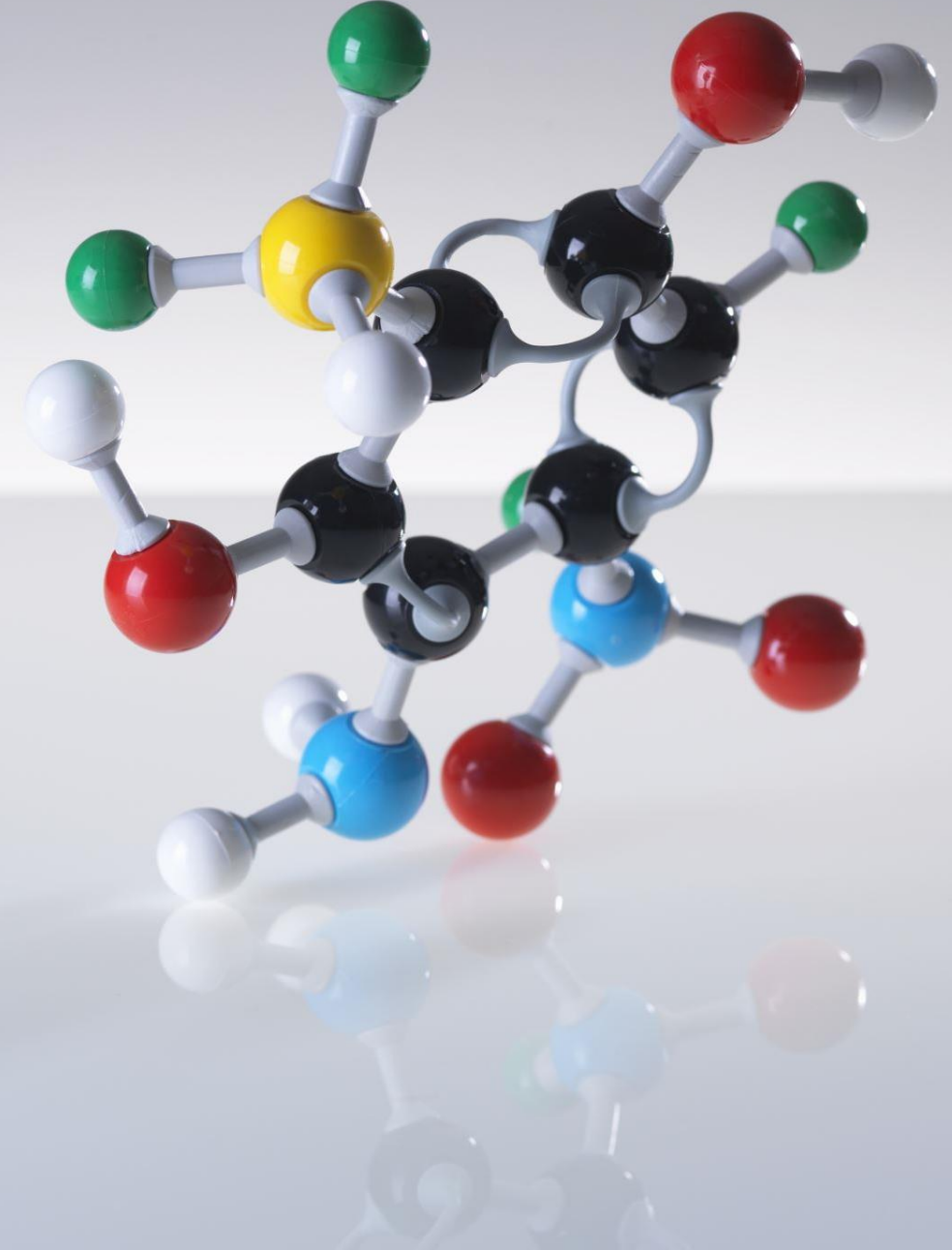
Transformers'ın Temel Bileşenleri

Residual Bağlantılar ve Layer Normalization:

- **Residual Bağlantılar:** Giriş tensorunu doğrudan çıkışa ekleyerek gradyan akışını iyileştirir ve derin ağların eğitimini kolaylaştırır.
- **Layer Normalization:** Her katmandaki aktivasyonları normalleştirerek eğitim sürecini stabilize eder ve daha hızlı yakınsama sağlar.

Masking:

- **İşlevi:** Decoder tarafında, modelin gelecek adımları görmesini engellemek için kullanılır. Bu, modelin sıradaki kelimeyi tahmin ederken yalnızca önceki kelimelere bakmasını sağlar.





Attention Mekanizması

Attention Nedir ve Neden Gereklidir?

Attention mekanizması, derin öğrenme modellerinin giriş verisinin belirli bölümlerine odaklanmasını sağlayan bir tekniktir. Özellikle doğal dil işleme (NLP) ve bilgisayarlı görü alanlarında kullanılır. Bu mekanizma, modelin tüm giriş yerine, görevi yerine getirmek için en önemli olan kısımlara yoğunlaşmasına olanak tanır.



Attention Nedir ve Neden Gereklidir?

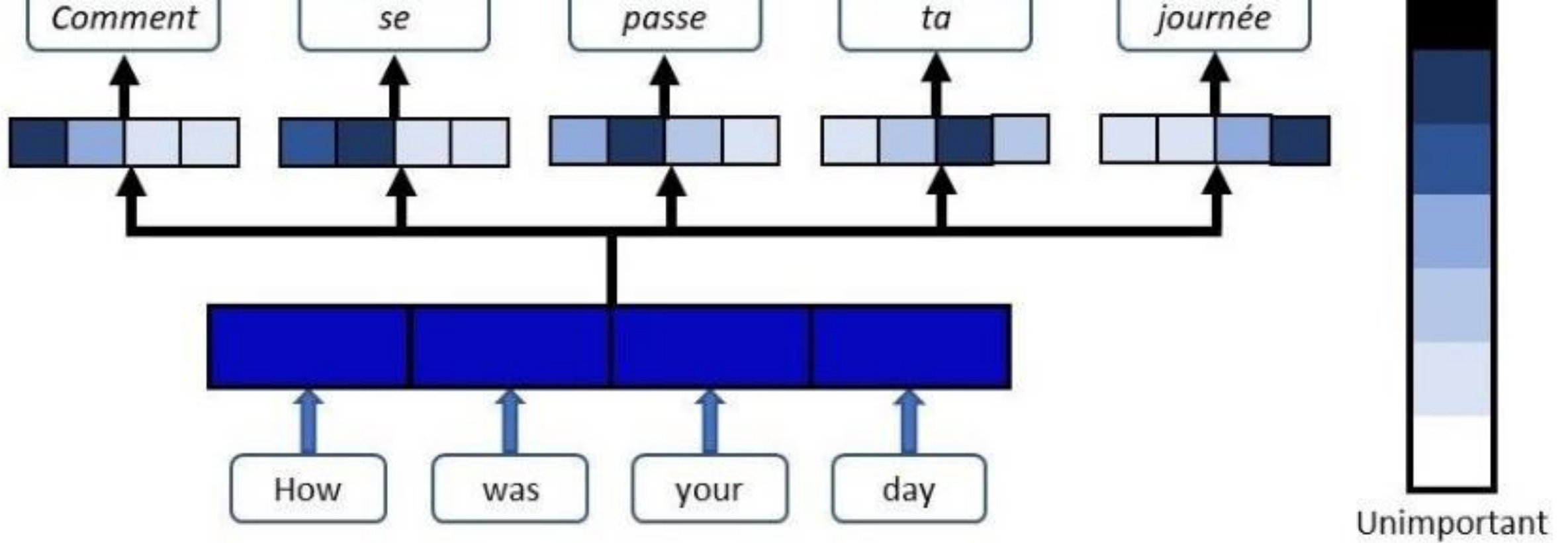
Uzun Dizilerle Çalışma: Geleneksel modeller, uzun metinler veya dizilerle çalışırken bilgi kaybı yaşayabilir. Attention mekanizması, modelin önemli bilgileri seçmesine ve uzun vadeli bağımlılıkları daha iyi yakalamasına yardımcı olur.

Bağlamsal Anlamlandırma: Kelimelerin veya sembollerin bağlam içinde doğru anlamlarını belirlemeye yardımcı olur. Örneğin, "bank" kelimesinin "finans kurumu" mu yoksa "nehir kıyısı" mı olduğunu bağlama göre ayırt edebilir.

Paralel İşleme: Attention mekanizması, hesaplamaların paralel yapılmasına olanak tanır, bu da eğitim ve çıkarım süreçlerini hızlandırır.

Performans Artışı: Makine çevirisi, metin özetleme ve soru-cevap gibi görevlerde modelin doğruluğunu ve etkinliğini artırır.





Self-Attention Nedir?

Self-attention, bir dizideki her ögenin (örneğin, bir cümledeki kelimenin) aynı dizideki diğer tüm öğelerle olan ilişkisini öğrenmesini sağlayan bir attention türüdür. Bu mekanizma, modelin kelimeler arasındaki bağımlılıkları ve ilişkileri anlamasına yardımcı olur.

Self-Attention Nasıl Çalışır?

Bağlam Yakalama: Her kelime için, diğer kelimelerin o kelimeyle ne kadar ilişkili olduğunu hesaplar.

Özelliklerin Ağırlıklandırılması: İlgili kelimelere daha fazla, ilgisiz olanlara daha az ağırlık verilir.

Esnek Bağımlılık Modelleme: Hem yakın hem de uzak kelimeler arasındaki ilişkiler eşit derecede dikkate alınabilir.

Avantajları

Paralel İşleme İmkânı: RNN'lerin aksine, self-attention mekanizması paralel hesaplamalara izin verir.

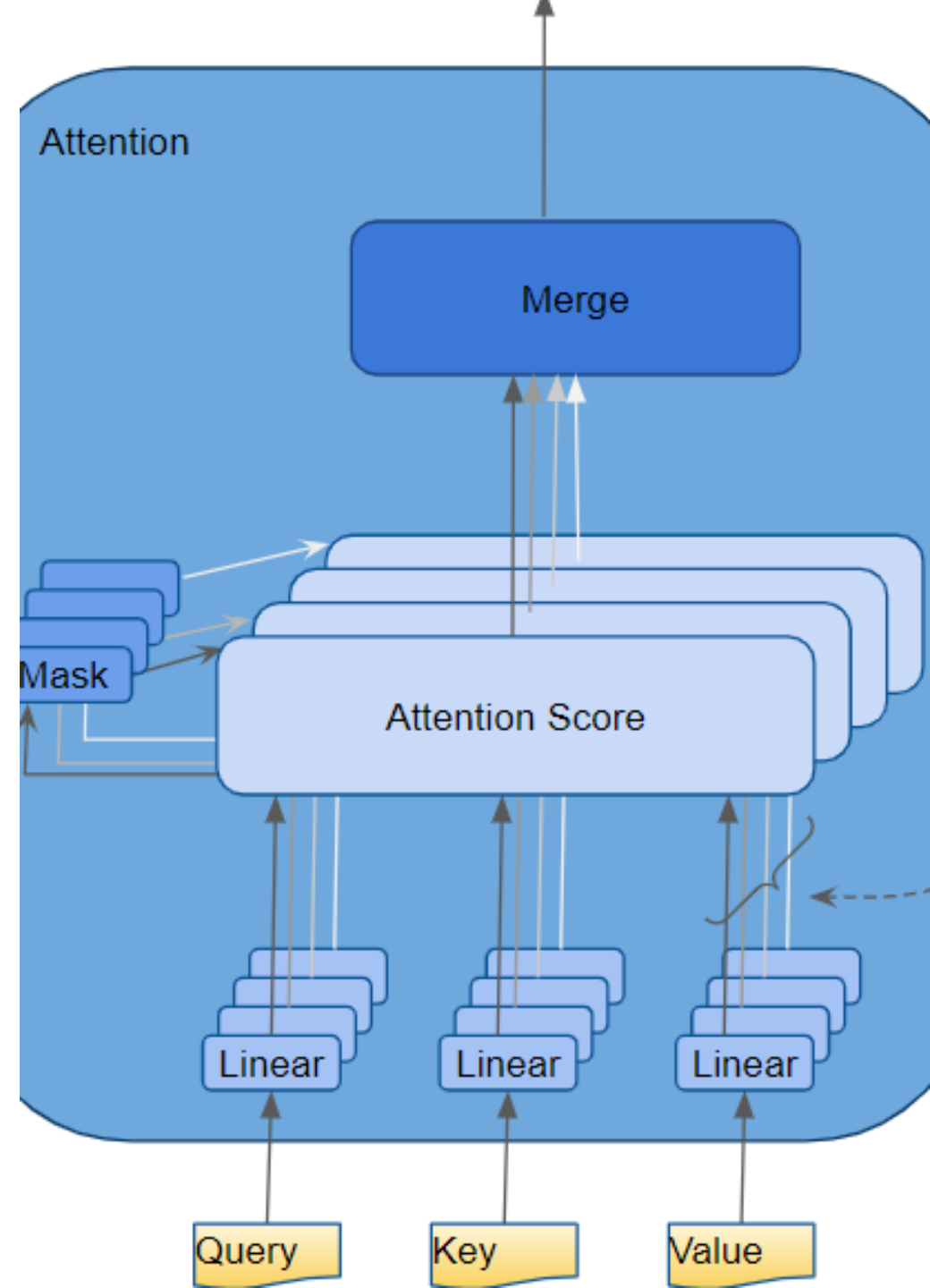
Uzun Vadeli Bağımlılıklar: Uzak kelimeler arasındaki ilişkileri etkin bir şekilde yakalar.

Bağlamsal Temsil: Kelimelerin anlamlarını, cümle içindeki tüm diğer kelimelerle olan ilişkileri üzerinden belirler.



Multi-Head Attention Nedir?

Multi-head attention, attention mekanizmasının birden fazla "başlık" (head) üzerinden paralel olarak uygulanmasıdır. Her bir başlık, farklı bir alt uzayda attention hesaplaması yapar ve böylece modelin farklı türde ilişkileri öğrenmesine olanak tanır.



Multi-Head Attention Nasıl Çalışır?

Başlıkların Oluşturulması: Girdi verisi, lineer dönüşümlerle farklı parçalara (başlıklara) ayrılır.

Paralel Attention Hesaplamaları: Her başlık üzerinde ayrı ayrı attention hesaplamaları yapılır.

Başlıkların Birleştirilmesi: Başlıklardan elde edilen çıktılar birleştirilir ve yeniden bir lineer dönüşümden geçirilir.

Avantajları

Çeşitli Özelliklerin Öğrenilmesi: Farklı başlıklar, kelimeler arasındaki farklı ilişkileri yakalayabilir.

Modelin Kapasitesini Artırma: Modelin daha zengin ve detaylı temsil öğrenmesini sağlar.

Paralel İşleme: Hesaplamalar paralel olarak yapıldığı için verimlilik artar.



Attention mekanizması örnekleri

Örnek 1: Makine Çevirisi

İngilizce "The cat sat on the mat" cümlesini Türkçeye çevirelim: "Kedi paspasın üzerine oturdu." Çeviri sırasında, "cat" kelimesi "kedi" kelimesine karşılık gelir. Attention mekanizması, çeviri işlemi sırasında hedef dilde üretilen her kelime için kaynak dildeki hangi kelimelere daha fazla odaklanılması gerektiğini belirler.

- **"Kedi" kelimesini üretirken**, model en çok "cat" kelimesine dikkat eder.
- **"Üzerine" kelimesini üretirken**, model "on" kelimesine odaklanır.

Bu sayede, doğru ve akıcı bir çeviri elde edilir.



Attention mekanizması örnekleri

Örnek 2: Anlam Belirsizliğinin Giderilmesi

Cümle: "Kalem ucu kırıldı, yazamıyorum."

"Burada "ucu" kelimesinin "kalem" ile ilişkili olduğunu anlamak önemlidir. Attention mekanizması, "ucu" kelimesi ile "kalem" kelimesi arasındaki bağlantıyı kurarak anlamın doğru yorumlanmasına yardımcı olur.



Attention mekanizması örnekleri

Örnek 3: Uzun Vadeli Bağlantıların Yakalanması

Cümle: "Ahmet, dün gördüğü filmi çok beğendi ve arkadaşlarına tavsiye etti."

"Arkadaşlarına tavsiye etti" ifadesindeki "etti" fiilinin öznesi "Ahmet"tir. Arada başka kelimeler olmasına rağmen, attention mekanizması bu bağlantıyı kurabilir ve cümlenin doğru anlaşılmasını sağlar



Attention mekanizması örnekleri

Örnek 4: Self-Attention ile Kelime İlişkilerinin Belirlenmesi

Cümle: "Onların evine gittik ve onlarla birlikte yemek yedik."

"Onların" ve "onlarla" kelimeleri arasındaki ilişkiyi self-attention mekanizması sayesinde model öğrenebilir.

Bu, modelin zamirlerin hangi isimlere atıfta bulunduğunu anlamasına yardımcı olur.





Positional Encoding (Konum Kodlaması)

Sekans Bilgisinin Modelde Temsili

Transformers, geleneksel RNN veya LSTM modellerinin aksine, sıralı veriyi işlemek için tekrarlayan yapılar kullanmaz. Bu durumda, modelin kelimelerin dizideki sırasını nasıl bileceği önemli bir soru haline gelir. Doğal dilde kelimelerin sırası anlamı önemli ölçüde etkiler. Örneğin, "Kedi köpeği kovaladı." ile "Köpek kediye kovaladı." cümleleri aynı kelimeleri içerir ancak farklı anlamlara sahiptir.

$$P(k, 2i) = \sin\left(\frac{k}{n^{2i/d}}\right)$$

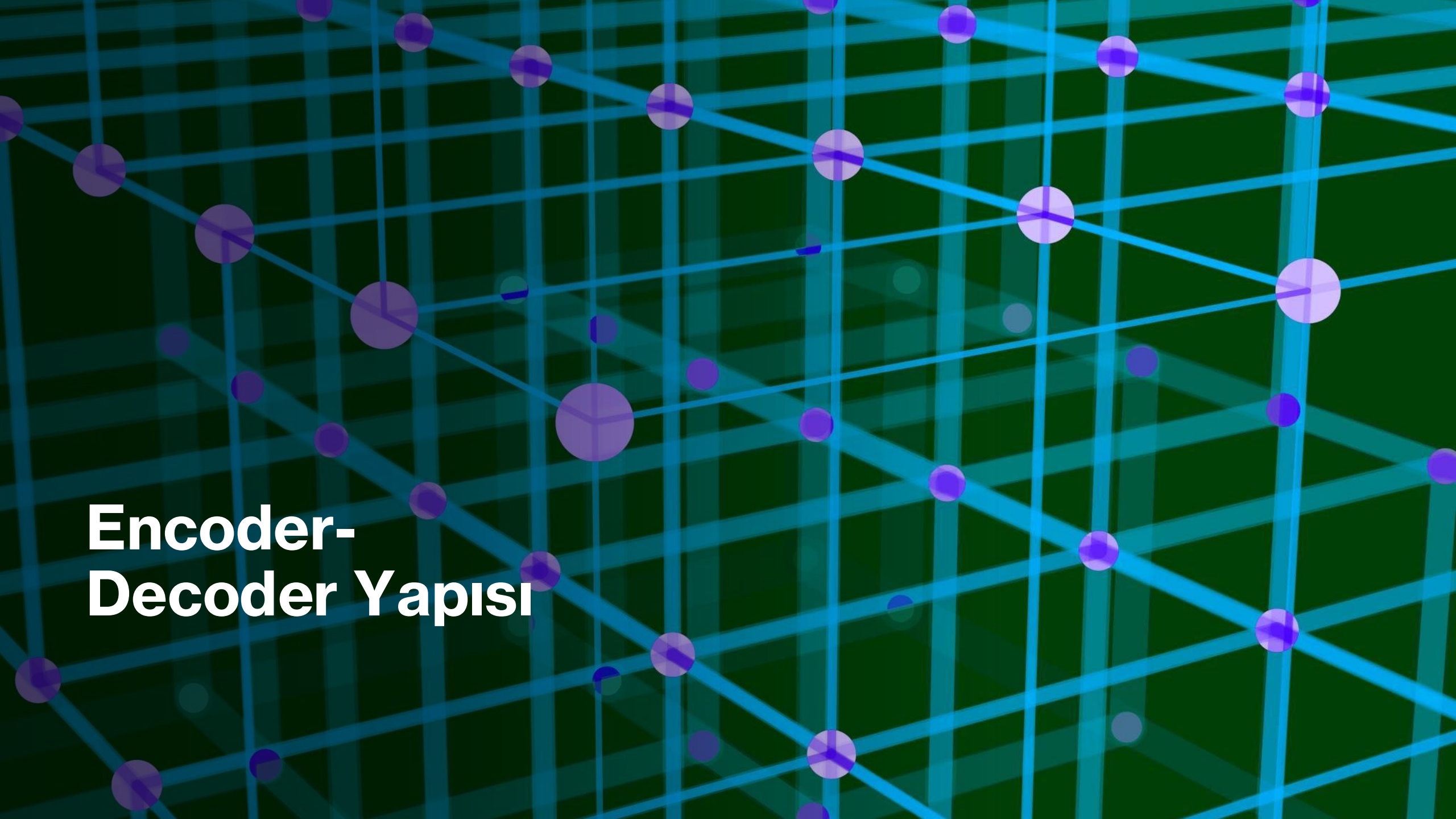
$$P(k, 2i + 1) = \cos\left(\frac{k}{n^{2i/d}}\right)$$

Sekans Bilgisinin Modelde Temsili

Transformers'ın bu sıralı bilgiyi modellemesi için **Positional Encoding (Konum Kodlaması)** kullanılır. Bu kodlama yöntemi, modelin kelimelerin veya token'ların dizideki konumlarını öğrenmesine olanak tanır. Bu sayede, model sıralı bağımlılıkları ve kelimelerin bağlam içindeki rollerini daha iyi anlar.

Sequence	Index of token, k	Positional Encoding Matrix with $d=4$, $n=100$			
		$i=0$	$i=0$	$i=1$	$i=1$
I	0	$P_{00}=\sin(0)$ = 0	$P_{01}=\cos(0)$ = 1	$P_{02}=\sin(0)$ = 0	$P_{03}=\cos(0)$ = 1
am	1	$P_{10}=\sin(1/1)$ = 0.84	$P_{11}=\cos(1/1)$ = 0.54	$P_{12}=\sin(1/10)$ = 0.10	$P_{13}=\cos(1/10)$ = 1.0
a	2	$P_{20}=\sin(2/1)$ = 0.91	$P_{21}=\cos(2/1)$ = -0.42	$P_{22}=\sin(2/10)$ = 0.20	$P_{23}=\cos(2/10)$ = 0.98
Robot	3	$P_{30}=\sin(3/1)$ = 0.14	$P_{31}=\cos(3/1)$ = -0.99	$P_{32}=\sin(3/10)$ = 0.30	$P_{33}=\cos(3/10)$ = 0.96

Positional Encoding Matrix for the sequence 'I am a robot'

The background features a complex, abstract pattern of intersecting lines in various shades of green and blue, creating a grid-like effect. Scattered throughout this grid are numerous circles of different sizes and shades of purple and blue. Some circles are solid, while others appear to be outlines or have a lighter center. The overall composition is dynamic and tech-oriented.

Encoder- Decoder Yapısı

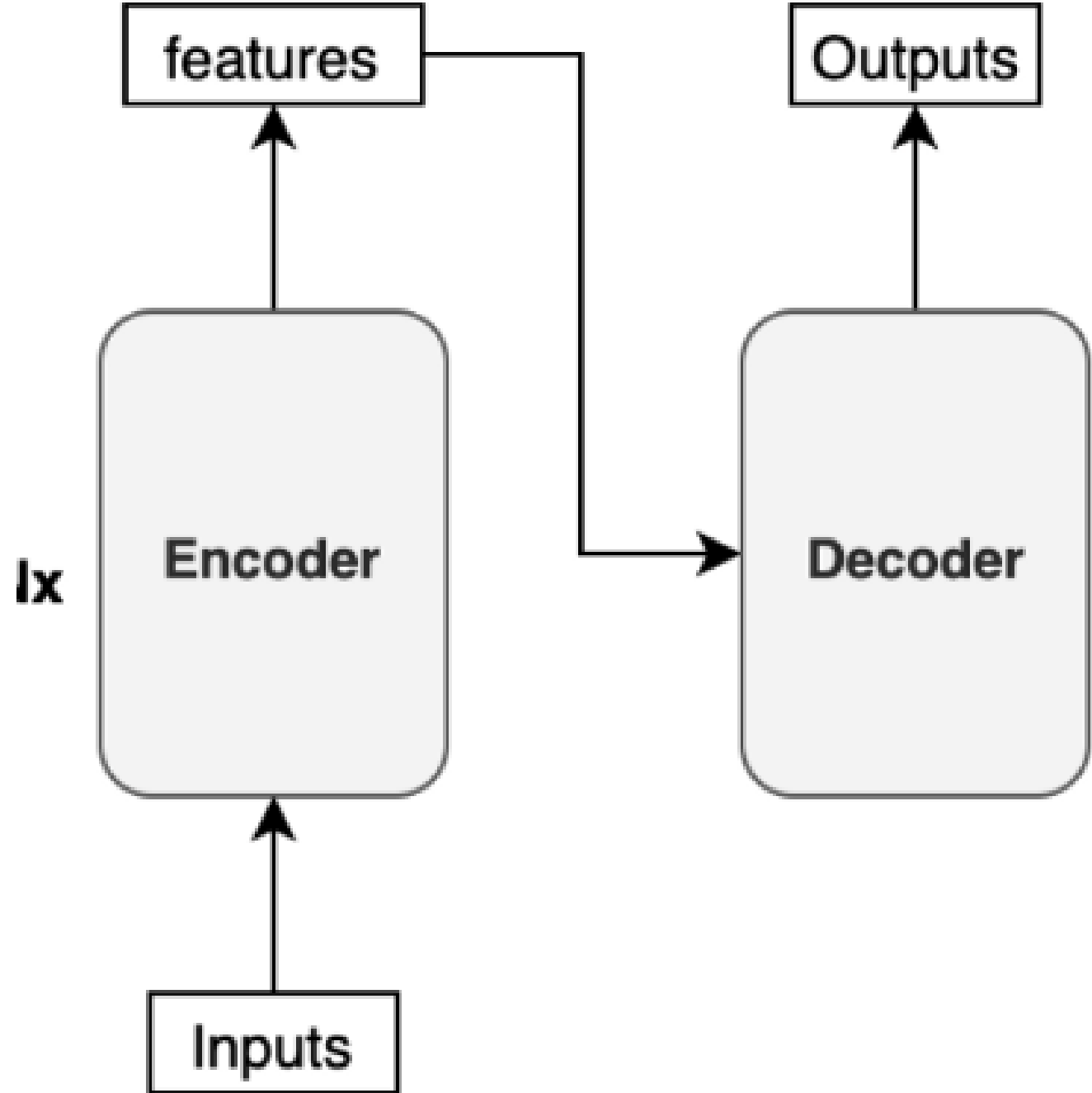
Encoder ve Decoder Modüllerinin İşlevleri

Encoder Modülü:

- **Girdi İşleme:** Encoder, giriş dizisini (örneğin bir cümleyi) alır ve bu diziyi ardışık işlemlerle daha yüksek düzeyli temsilcilere dönüştürür.
- **Temel Görevi:** Giriş verisinin anlamını ve bağlamsal ilişkilerini yakalamaktır.
- **Katmanlar:** Birden fazla encoder katmanından oluşur ve her katman self-attention ve feed-forward sinir ağlarını içerir.
- **Self-Attention Kullanımı:** Encoder'daki self-attention mekanizması, giriş dizisindeki her öğenin diğer öğelerle olan ilişkisini öğrenmesini sağlar.

Decoder Modülü:

- **Çıktı Üretme:** Decoder, encoder'dan gelen bilgilere dayanarak hedef çıktıyı (örneğin çeviri cümlesini) üretir.
- **Temel Görevi:** Giriş verisine uygun ve anlamlı bir çıktıyı adım adım oluşturmaktır.
- **Katmanlar:** Birden fazla decoder katmanından oluşur ve her katman masked self-attention, encoder-decoder attention ve feed-forward sinir ağlarını içerir.



Encoder-Decoder Etkileşimi

Bilgi Akışı:

- Encoder, giriş dizisini işleyerek gizli temsiller üretir.
- Decoder, bu temsilleri alarak hedef diziyi üretir.

Encoder'dan Decoder'a Geçiş:

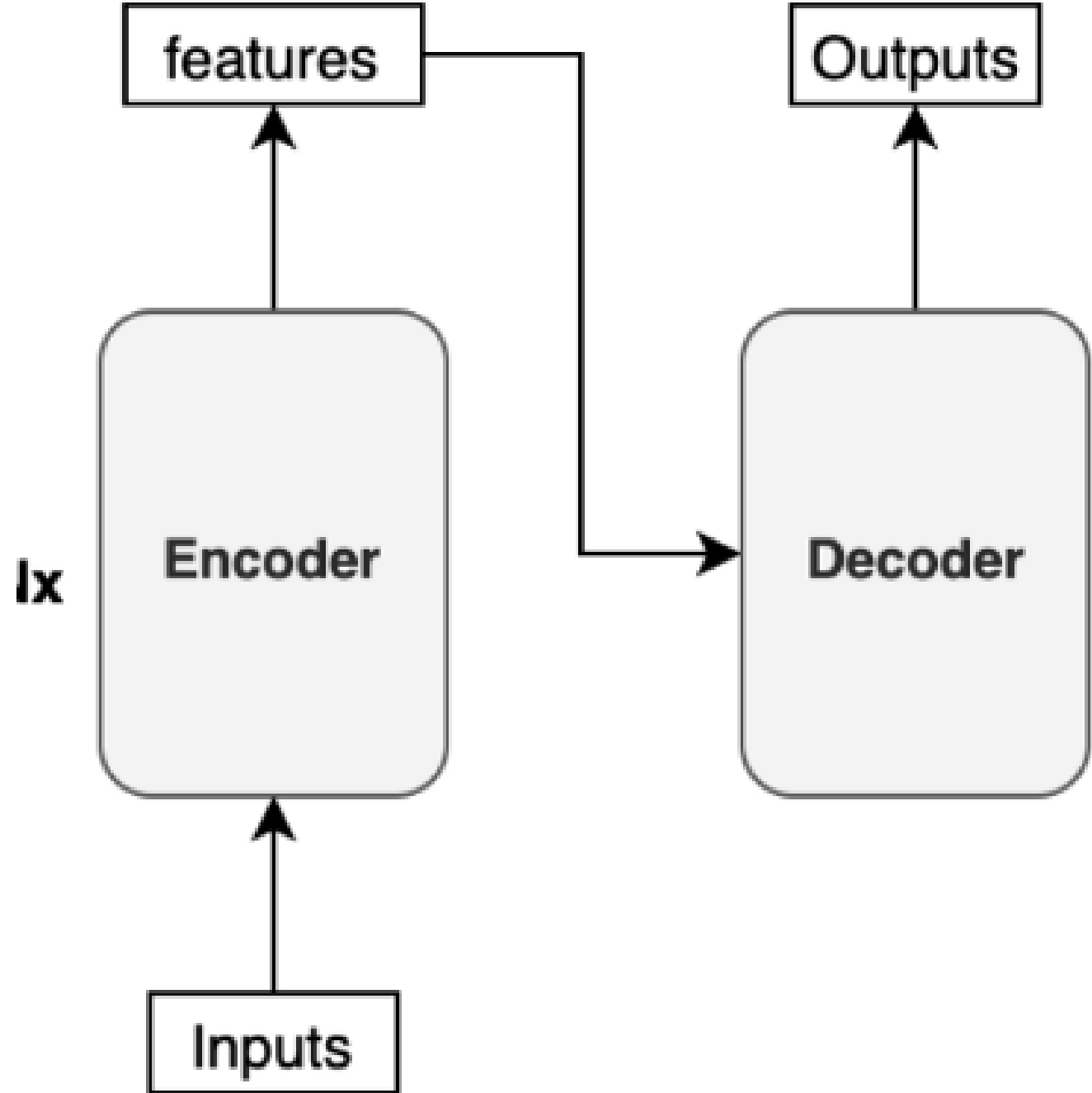
- Encoder'ın her katmanından elde edilen çıktılar, decoder'ın attention mekanizmalarında kullanılır.
- Encoder-Decoder Attention, decoder'ın giriş dizisindeki önemli bilgilere odaklanmasını sağlar.

Etkileşimin Önemi:

- Bu etkileşim sayesinde, decoder hem kendi oluşturduğu çıktıyı hem de orijinal girdiyi dikkate alarak daha doğru ve tutarlı sonuçlar üretir.

İki Modülün Birlikte Çalışması:

- Encoder ve decoder modülleri birlikte çalışarak karmaşık görevleri (örneğin makine çevirisi) etkili bir şekilde gerçekleştirebilir.



Masking ve Geleceđi Görme Problemi

Geleceđi Görme Problemi Nedir?

Modelin, tahmin etmesi gereken bir kelimeyi veya gelecek adımları önceden görerek tahmin yapması istenmez. Bu durum, modeli hile yapmaya yönlendirir ve gerçekçi olmayan sonuçlara yol açar.



Masking ve Geleceği Görme Problemi

Masking (Maskeleme):

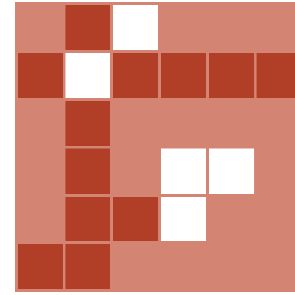
- **Amaç:** Geleceğe ait bilgilerin modele sızmasını engellemek.

Nasıl Uygulanır:

- **Decoder'da Masked Self-Attention:** Masking, decoder'ın attention mekanizmasında uygulanır. Bu, modelin sadece önceki kelimelere bakarak sonraki kelimeyi tahmin etmesini sağlar.
- **Üçgen Şeklinde Maske:** Genellikle alt üçgen bir matris kullanılarak gelecekteki pozisyonlar maskelenir.

Örnek:

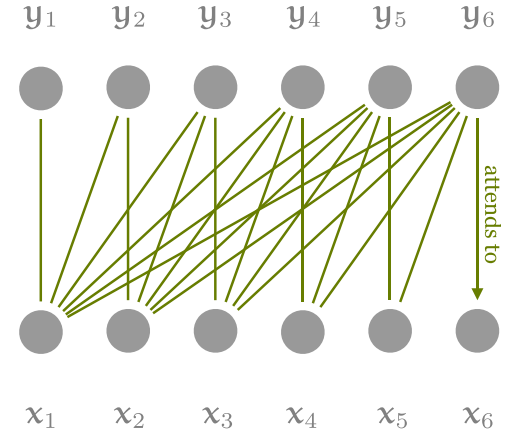
- Bir dil modelinde, "Ben bugün __" cümlesindeki boşluğu doldururken, model "gidiyorum" kelimesini tahmin ederken "spora" kelimesini önceden görmemelidir.



raw attention weights



mask



Maskellemenin Önemi

1

Doğru Öğrenme: Modelin dizinin sıralı doğasını dikkate alarak öğrenmesini sağlar.

2

Otantik Üretim: Doğal ve tutarlı cümleler oluşturmaya yardımcı olur.

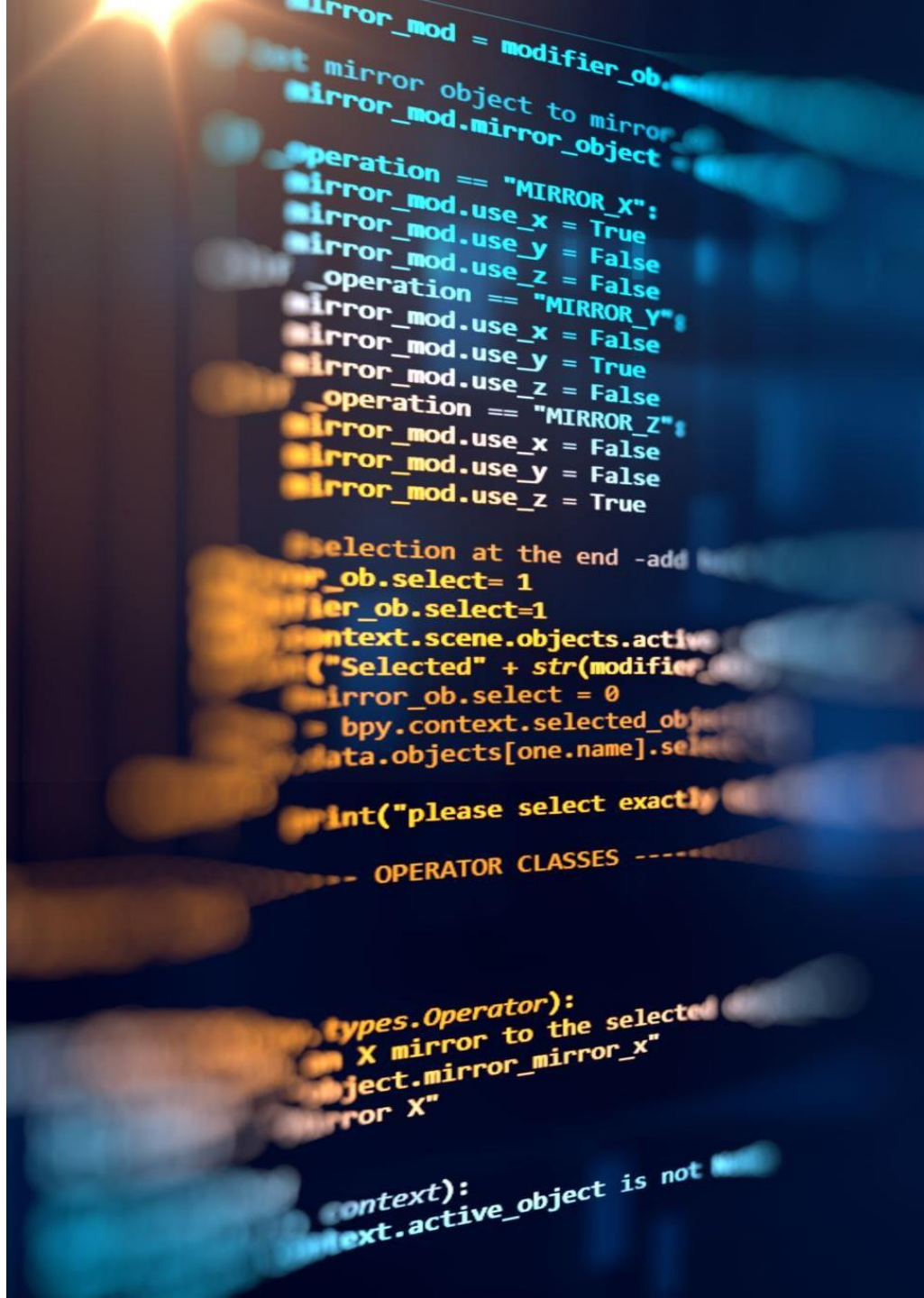
Transformerlarda Encoder-Decoder Uygulamaları

Makine Çevirisi:

- **Uygulama:** Bir dildeki cümleyi başka bir dile çevirme.
- **Nasıl Çalışır:**
 - Encoder, kaynak dildeki cümleyi işler.
 - Decoder, hedef dildeki cümleyi üretirken encoder'dan gelen bilgileri kullanır.

Metin Özeti Oluşturma:

- **Uygulama:** Uzun bir metni daha kısa bir özet haline getirme.
- **Nasıl Çalışır:**
 - Encoder, orijinal metni işler.
 - Decoder, metnin özetini üretir.



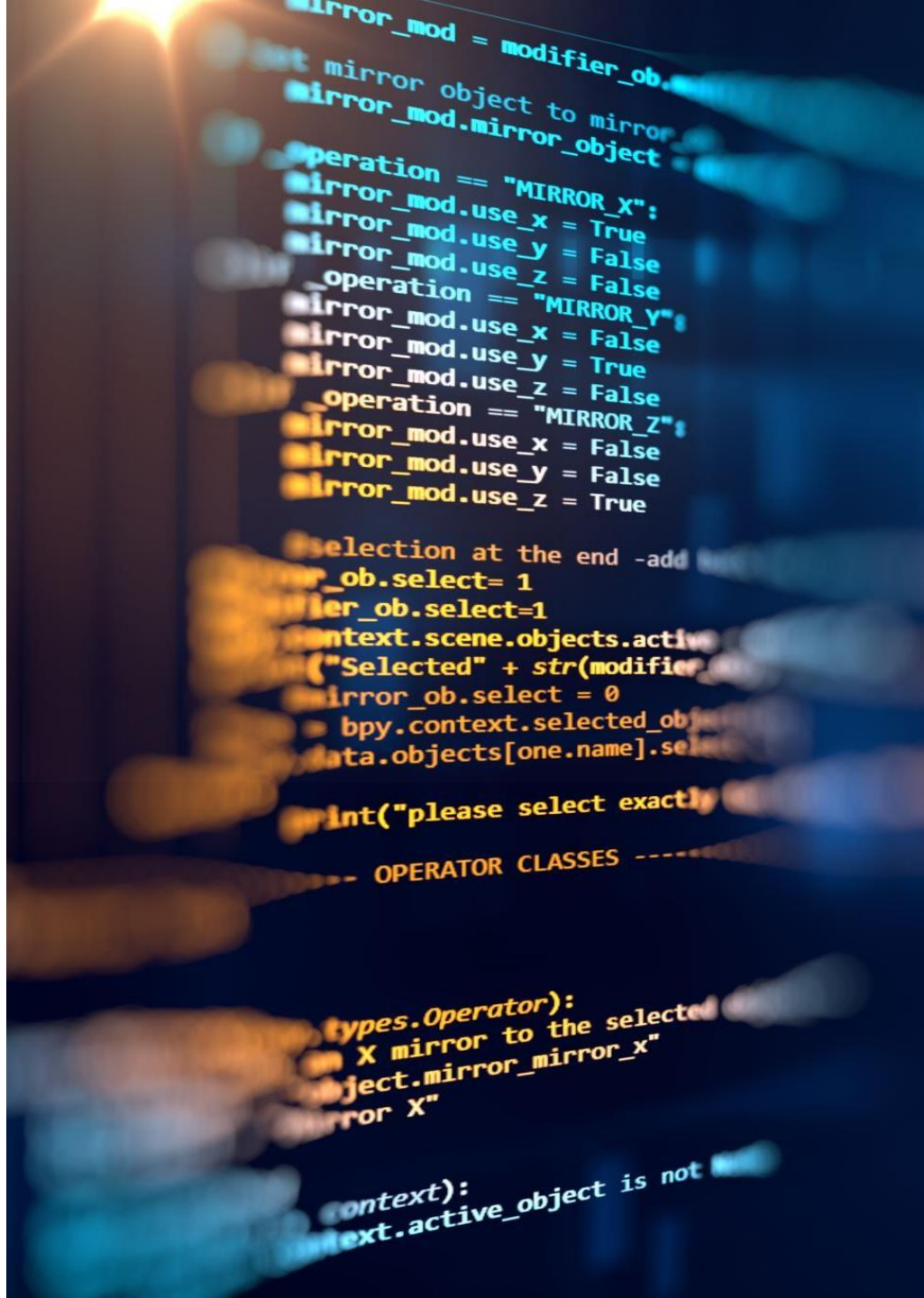
Transformerlarda Encoder-Decoder Uygulamaları

Soru-Cevap Sistemleri:

- **Uygulama:** Bir metne dayalı olarak soruları cevaplama.
- **Nasıl Çalışır:**
 - Encoder, soru ve ilgili metni işler.
 - Decoder, soruya uygun cevabı üretir.

Görüntü Altyazılama (Image Captioning):

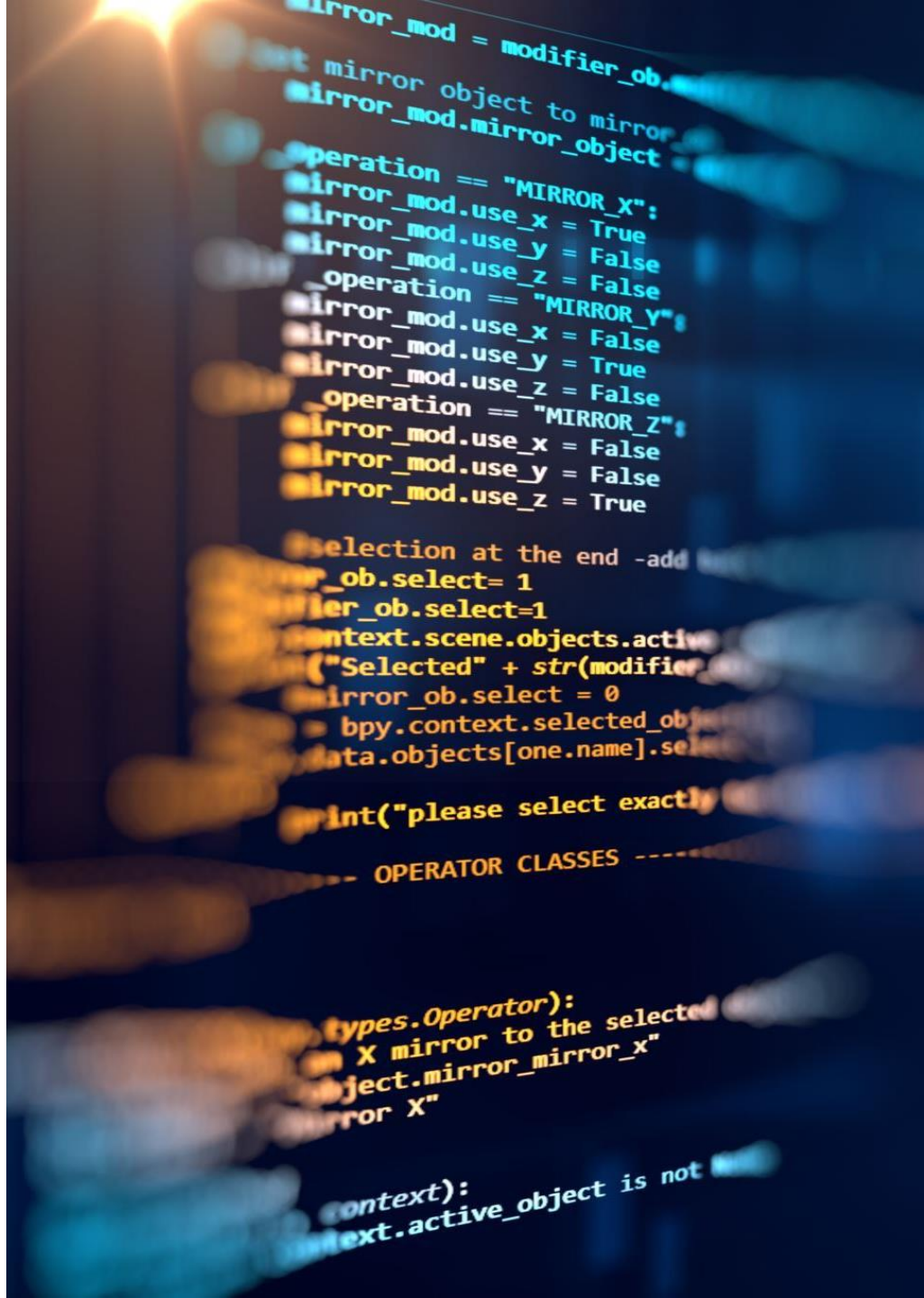
- **Uygulama:** Görüntülere uygun açıklamalar veya altyazılar oluşturma.
- **Nasıl Çalışır:**
 - Encoder, görüntüyü işler ve özelliklerini çıkarır.
 - Decoder, bu özelliklere dayanarak açıklama metni üretir.



Transformerlarda Encoder-Decoder Uygulamaları

Konuşma Tanıma ve Sentezleme:

- **Uygulama:** Ses verisini metne dönüştürme veya metinden ses üretme.
- **Nasıl Çalışır:**
 - Encoder, ses sinyalini işler.
 - Decoder, metin çıktısı veya sentezlenmiş ses üretir.



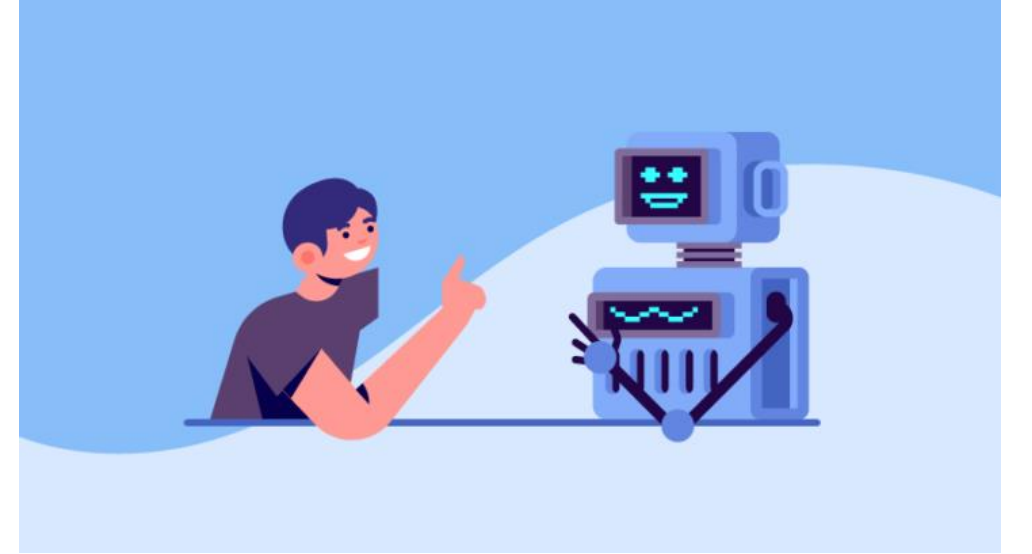
Transformer'ın Uygulama Alanları



Makine Çevirisi

Transformers, makine çevirisi alanında devrim niteliğinde gelişmelere yol açmıştır. Geleneksel RNN tabanlı modellerin sınırlamalarını aşarak, daha etkili ve verimli çeviri modelleri oluşturulmasını sağlamıştır.

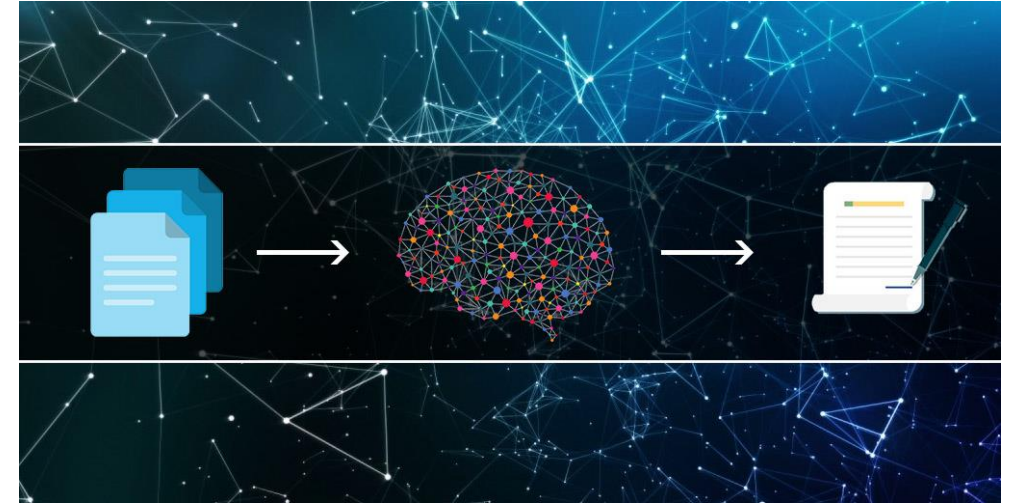
- **Üstün Performans:** Attention mekanizması sayesinde, model kaynak ve hedef dil arasındaki karmaşık ilişkileri daha iyi öğrenir. Bu, daha akıcı ve doğru çevirilerin üretilmesine olanak tanır.
- **Paralel İşleme:** Transformers'ın paralel hesaplama yeteneği, büyük veri setleri üzerinde hızlı eğitim imkanı sunar. Bu, çeviri modellerinin ölçeklenebilirliğini artırır.
- **Pratik Uygulamalar:** Google Translate ve diğer çeviri hizmetleri, Transformer tabanlı modelleri kullanarak çeviri kalitesini büyük ölçüde geliştirmiştir.



Metin Özeti Oluřturma

Transformers, uzun metinlerin önemli bilgilerini özetleyerek daha kısa ve anlaşılır metinler oluřturma konusunda etkilidir.

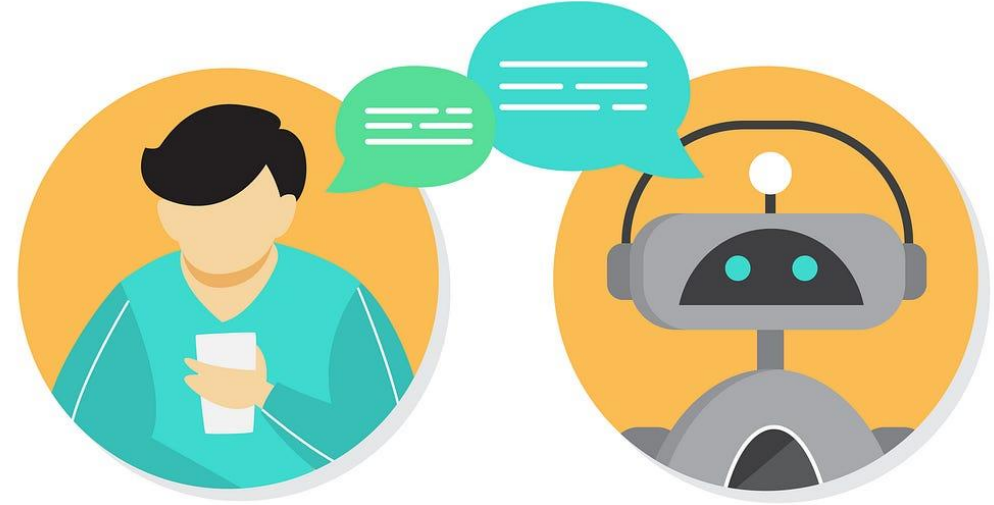
- **Abstraktif Özetleme:** Transformers, metindeki bilgileri olduėu gibi aktarmak yerine, kendi cümlelerini kurarak özetler oluřturabilir. Bu, insan benzeri ve özgün özetlerin üretilmesini sağlar.
- **Dikkat Mekanizması ile Önemli Bilgilerin Seçimi:** Attention mekanizması, metindeki önemli cümle ve kelimelere odaklanarak özette yer alması gereken bilgileri belirler.
- **Uygulama Alanları:** Haber özetleme, akademik makalelerin özetlenmesi, hukuki belgelerin kısaltılması gibi pek çok alanda kullanılır.



Soru-Cevap Sistemleri

Transformers, bir metin veya veri tabanı üzerinde sorulan sorulara doğru ve tutarlı cevaplar verme konusunda başarılı modeller oluşturulmasını sağlar.

- **Metin Anlamlandırma:** Self-attention mekanizması sayesinde model, hem sorunun hem de metnin bağlamını derinlemesine anlayabilir.
- **Doğru Bilgi Erişimi:** Model, soruyla ilgili metin bölümlerine odaklanarak doğru cevapları çıkarır.
- **Etkileşimli Uygulamalar:** Akıllı asistanlar (örneğin Siri, Alexa), müşteri hizmetleri chatbotları ve eğitim amaçlı soru-cevap platformlarında kullanılır.



Dil Modelleme ve Metin Üretimi

Transformers, dil modelleme ve doğal dil üretimi alanlarında da öncü rol oynar.

- **Metin Tamamlama ve Üretme:** Transformers, verilen bir başlangıç metnini anlamlı ve tutarlı bir şekilde devam ettirebilir veya tamamen yeni metinler üretebilir.
- **Yaratıcı Yazma:** Hikaye yazımı, şiir oluşturma ve yaratıcı içerik üretimi gibi görevlerde kullanılır.
- **Önceden Eğitilmiş Büyük Dil Modelleri:** GPT-2, GPT-3 ve benzeri modeller, metin üretimi ve dil anlama görevlerinde kullanım için geniş bir potansiyel sunar.
- **Kişiselleştirilmiş İçerik Oluşturma:** Kullanıcı tercihleri ve stiline göre özelleştirilmiş metinler üretilebilir.

