YAPAY ZEKÂYI KANDIRMAK

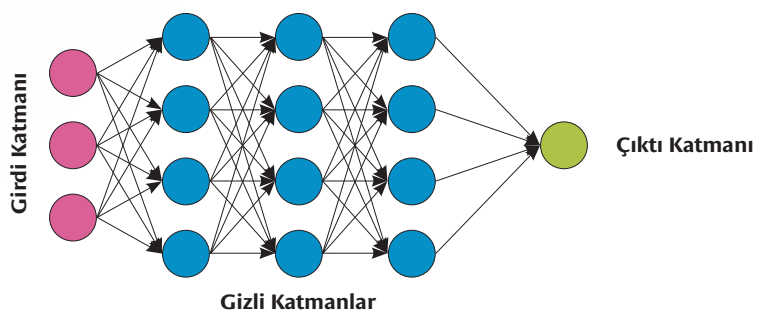
Dr. Mahir E. Ocak [*TÜBİTAK Bilim ve Teknik Dergisi*

Derin öğrenme ile eğitilmiş yapay zekâ uygulamaları, günlük hayatımızın bir parçası oldular. Örüntü tanıma konusunda çok başarılılar. Görüntü ve konuşma da dâhil olmak üzere her türlü veriyi sınıflandırabiliyorlar. Otomatik telefon sistemlerinden, otonom araçları idare etmeye ve hastalara teşhis koymaya kadar pek çok alanda kullanılıyorlar. Ancak yakın zamanlarda yapılan araştırmalar, derin öğrenme ile eğitilmiş yapay zekâ uygulamalarını kandırmak için girdilerde ufak tefek değişiklikler yapmanın bile yeterli olduğunu gösterdi.

Yapay zekâ uygulamaları olmayan şeyleri görebiliyor ve tamamen doğal fotoğraflarda bile beklenmedik hatalar yapabiliyorlar.

Derin Sinir Ağları

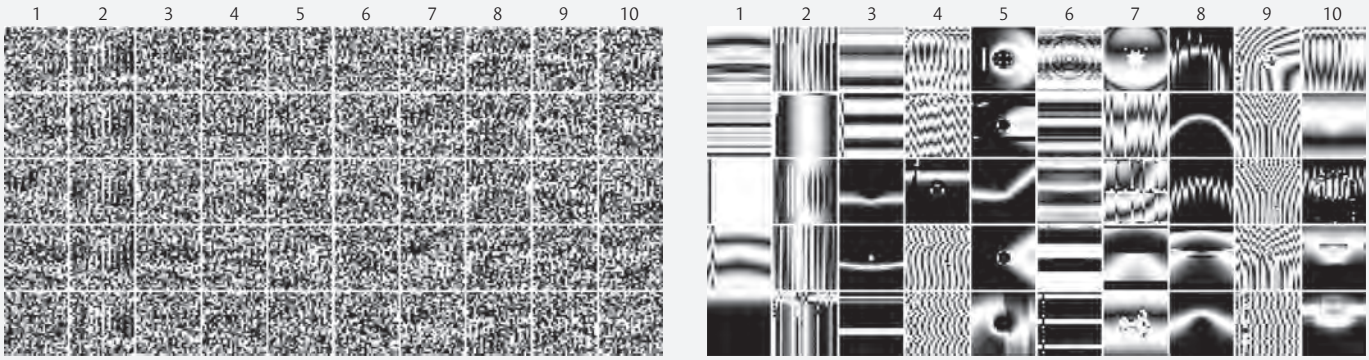
Derin sinir ağları (DNN), katmanlar hâlinde organize olmuş dijital nöronlardan oluşur. Her bir katmandaki nöronlar kendisinden bir önceki ve bir sonraki katmandaki nöronlarla karmaşık biçimlerde bağlantılıdır. Bir katmandaki nöronlar bir önceki katmandaki nöronlardan gelen bilgiyi işler, bulduğu sonucu bir sonraki katmandaki belirli nöronlara iletir.



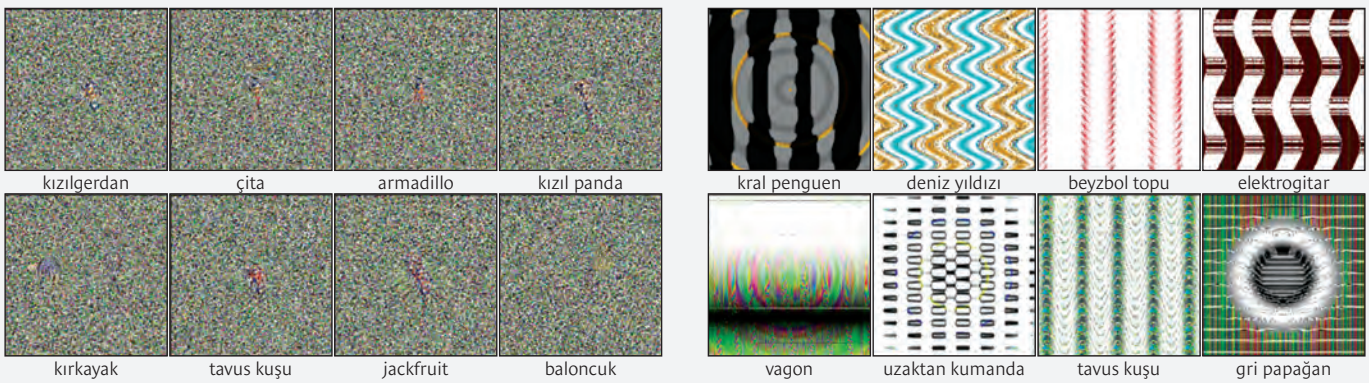
Derin sinir ağlarının bir örneği

Her bir nöronun bilgiyi işleme biçimi öğrenme süreci sırasında şekillenir. Örneğin bir derin sinir ağına çeşitli hayvanları sınıflandırmayı öğrettiğimizi düşünelim. Başlangıçta her bir nörona rastgele komutlar verilir ve çeşitli hayvan resimleri gösterilmeye başlanır. Örneğin gösterilen ilk resmin bir aslan fotoğrafı olduğunu düşünelim. Yapay zekâ uygulamasının, her bir nörona verilen rastgele komutlarla bu resimdeki hayvanın aslan olduğunu tespit etme ihtimali pratikte sıfırdır. Dolayısıyla derin sinir ağının yaptığı tahmin büyük olasılıkla yanlış olacaktır. Uygulamaya yaptığı tahminin yanlış olduğu söylenir ve doğru tahmin yapacak biçimde nöronlara verilen komutlar optimize edilir. Uygulama bir kez daha aynı fotoğrafla karşılaştığında fotoğraftaki hayvanın aslan olduğunu kesinlikle doğru tahmin edecektir. Daha sonra uygulamaya içinde başka bir hayvan, örneğin bir ayı, olan yeni bir fotoğraf gösterilir ve yine resimdeki hayvanı sınıflandırması istenir. Uygulamanın, her bir nörona verilen, belirli bir fotoğraftaki hayvanı aslan olarak sınıflandırması için optimize edilmiş komutlarla bu hayvanı da doğru olarak sınıflandırması beklenemez. Dolayısıyla uygulama büyük olasılıkla yine yanlış tahmin yapacaktır. Uygulamaya yine yanlış tahmin yaptığı söylenir ve nöronlara verilen komutlar bu fotoğraftaki ayıyı da doğru sınıflandıracak biçimde yeniden optimize edilir. Bu süreç, çok çeşitli hayvanların çok sayıda fotoğrafıyla binlerce, milyonlarca kez tekrar edilir. Eğitim süreci tamamlandığında, kendisine gösterilen fotoğraflardaki hayvanların hangi tür olduğunu çeşitli olasılıklarla tahmin eden bir yapay zekâ uygulaması ortaya çıkar. Örneğin böyle bir uygulamaya içinde daha önce hiç karşılaşmadığı bir hayvan bulunan bir fotoğraf gösterdiğinizde, size şöyle bir cevap verebilir: Resimdeki hayvan %95 olasılıkla gergedandır.

Yapay zekâ tarafından rakam olarak algılanan çeşitli görüntüler.

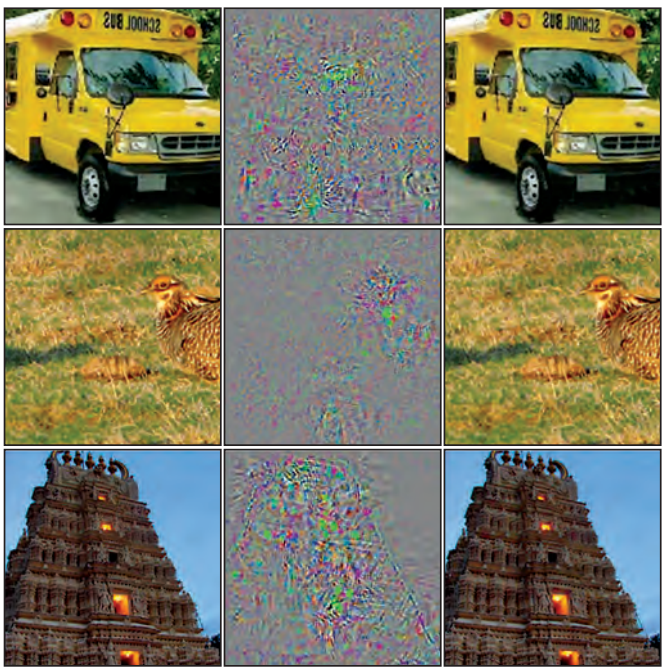


Yapay zekâ tamamen parazitten oluşan resimlerin (solda) ya da soyut resimlerin (sağda) içinde olmayan şeyler görebiliyor.



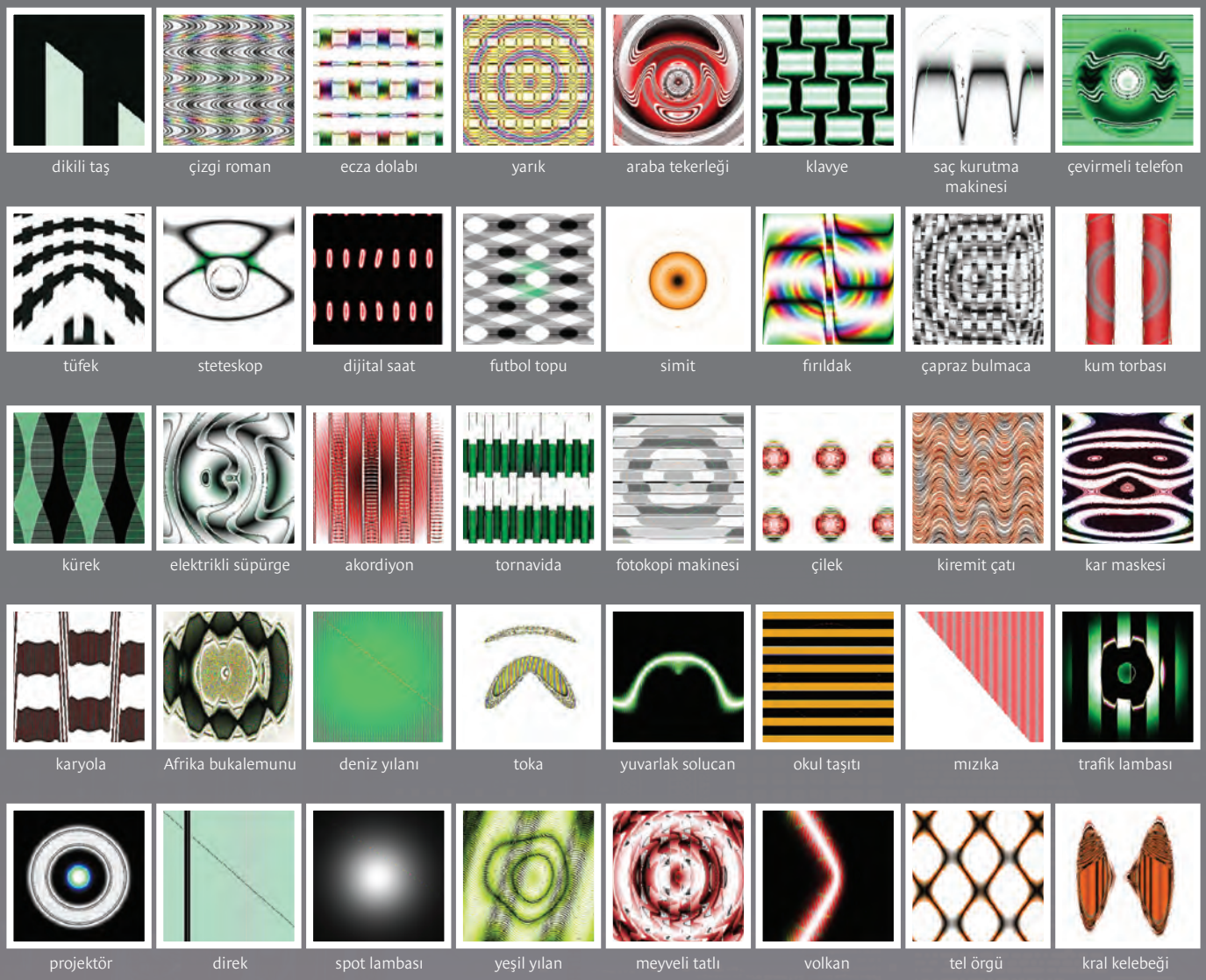
Orijinal resimler (solda),
resimlere eklenen parazit (ortada),
resimlerin parazit eklendikten sonraki hâlleri (sağda).

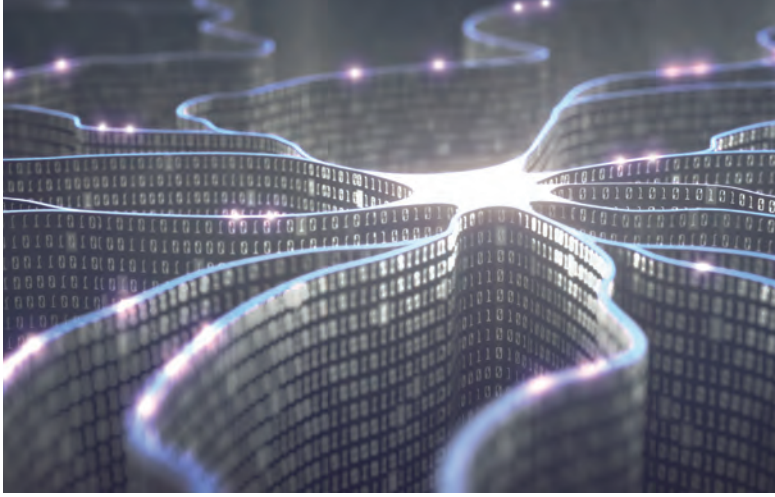
İnsan gözü orijinal resimler ve parazitli resimler arasındaki farkı algılayamıyor. Ancak orijinal resimler yapay zekâ tarafından doğru sınıflandırıldığı hâlde parazitli resimler yanlış sınıflandırılıyor.



Derin öğrenme, yapay zekâya bir şeyler öğretmenin tek yolu değil. Ancak başka yöntemlere göre çok daha hızlı bir biçimde öğrenmeye imkân vermesi sebebiyle çok yararlı bulunuyor ve sıklıkla tercih ediliyor. Derin öğrenmenin uygulama alanları arasında konuşma algılama sistemlerinden, görüntü tanıma sistemlerine, yeni ilaç keşfinden kanserli hücreleri tespit etmeye kadar pek çok şey sayılabilir.

Yapay zekâ tarafından yanlış sınıflandırılan çeşitli görüntüler.

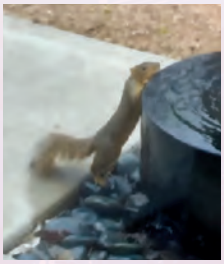




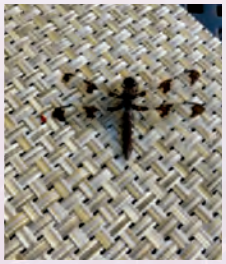
Hatalar

Derin öğrenmeyle eğitilen yapay zekâ uygulamaları her ne kadar pek çok alanda başarılı olsa ve hatta bazı işleri insanlardan daha iyi becerse bile bu durum mükemmel oldukları anlamına gelmiyor. Son yıllarda yapılan çalışmalar, derin öğrenme ile eğitilmiş yapay zekâ uygulamalarının çok basit hatalar yapabileceğini ve kolaylıkla kandırılabileceğini gösterdi.

Yapay zekânın nasıl kandırılabileceğiyle ilgili ilk makale, Google araştırmacılarından Christian Szegedy ve arkadaşları tarafından 2013 yılında yayımlanmıştı. Araştırmacıların yaptığı çalışmalar, herhangi bir görüntü üzerinde ufak tefek oynamalar yaparak DNN'leri yanlışa sürüklemenin mümkün olduğunu gösteriyordu. Örneğin bir hayvan resmini alıp bazı piksellerde ufak, sistematik değişiklikler yaparak yapay zekâyı gördüğünün aslında bir araba fotoğrafı olduğuna ikna etmek mümkündü. Üstelik bu durum belirli bir yapay zekâ uygulamasının eğitimindeki eksikliklerden de kaynaklanmıyordu. Üzerinde oynamalar yapılmış resim, tamamen farklı verilerle eğitilmiş herhangi başka bir yapay zekâ uygulaması tarafından da yanlış sınıflandırılıyordu.



sincap deniz aslanı (%99)



yusufçuk böceği rögar kapağı (%99)

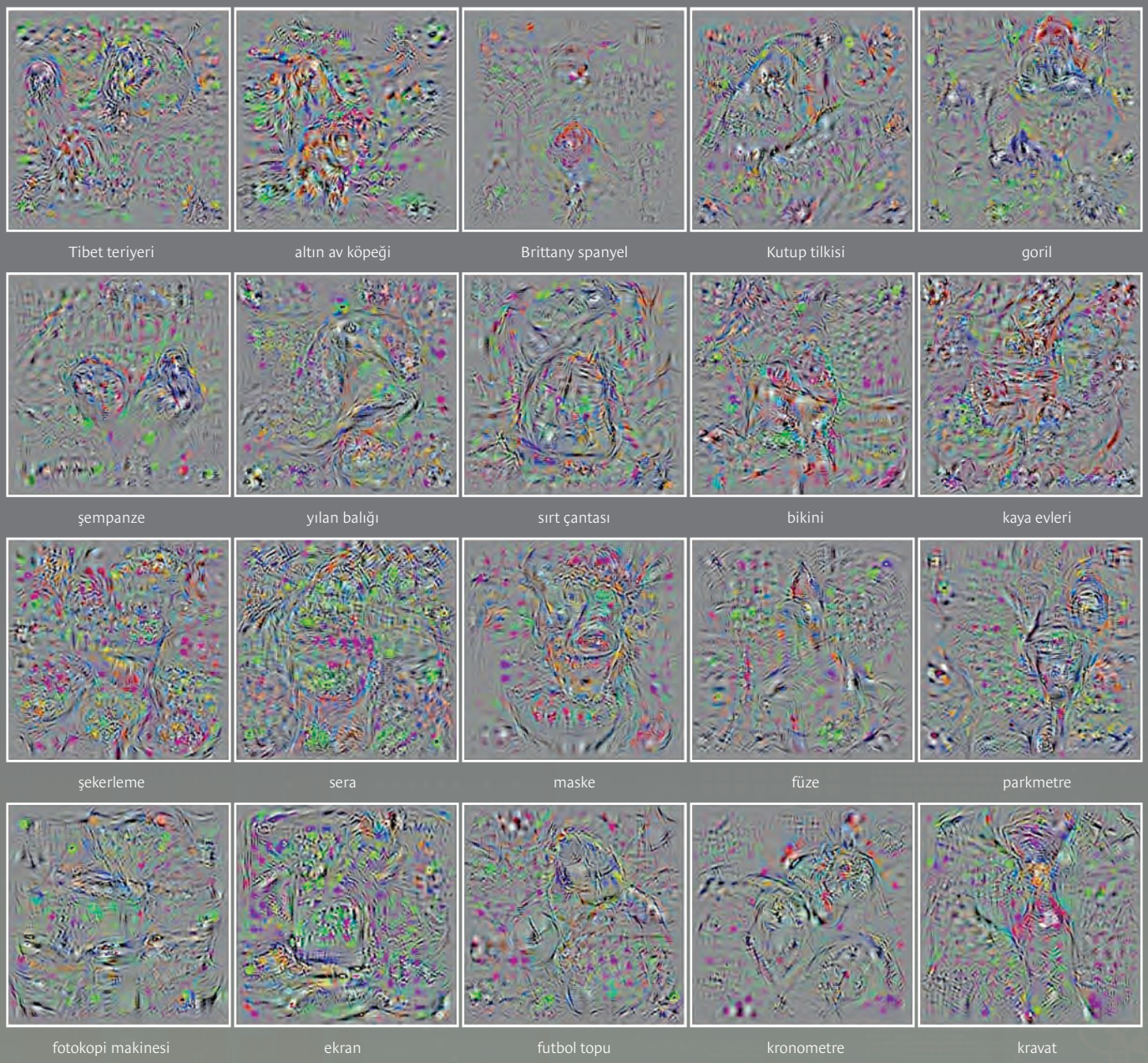
Bir yıl sonra Anh Nguyen, Jason Yosinski ve Jeff Clune, yapay zekâya olmayan şeyleri gördürmenin de mümkün olduğunu gösterdi. Örneğin derin öğrenme ile eğitilmiş yapay zekâ uygulamaları, soyut dalgalı çizgilerin içinde elektronik gitarlar ya da tamamen parazitli (her bir pikseli rastgele renklere boyanmış) resim dosyalarının içinde tavus kuşları görebiliyordu.

Daha sonraları DDN'lerin yaptığı farklı türlerde başka hatalar da keşfedildi. Örneğin geçen sene Anh Nguyen ve arkadaşları bir görüntüdeki nesneleri döndürmenin yapay zekâyı yanıltmak için yeterli olduğunu gösterdi. Örneğin yapay zekâ uygulamaları farklı açıdan çekilmiş "Dur" işareti levhalarını gülle ya da raket olarak sınıflandırabiliyordu.

Geçtiğimiz sene içinde Dan Hendrycks ve arkadaşları üzerinde hiçbir oynama yapılmamış, tamamen doğal görüntülerin bile iyi eğitilmiş yapay zekâ uygulamaları tarafından yanlış sınıflandırılabileceğini gösterdi. Örneğin yusufçuk böceğini rögar kapağı ya da mantarı çubuk kraker olarak sınıflandırmak gibi...

Yapay zekâ uygulamalarının yaptığı hatalar sadece görüntü tanımayla ilgili değil. Verileri sınıflandırmak için derin sinir ağları kullanan herhangi bir yapay zekâ uygulamasını kandırmak da mümkün. Örneğin Sandy Huang ve arkadaşları, 2017 yılında, görüntülerdeki birkaç pikselde ufak tefek değişiklikler yaparak Atari oyunları oynamak için eğitilmiş DNN'lere oyun kaybettirmeyi başardı.

Yapay zekâ tarafından yanlış sınıflandırılan çeşitli görüntüler.



Üzerine birkaç şerit eklenmiş "Dur" işareti yapay zekâ tarafından "45" (hız sınırı) olarak algılanıyor.

Sorunlar ve Çözümler

Derin sinir ağlarını güçlü yapan şey çok katmanlı olmaları ve bu sayede verilerdeki çeşitli örüntüleri kolaylıkla algılayabilmeleridir. Ancak bu durum derin sinir ağlarının aynı zamanda sınıflandırdıkları nesnelerin en belirgin özellikleri hakkında bir fikir edinmelerini de zorlaştırır. Örneğin bir insan için bir kuşun en belirgin özelliği gövde şekli ve özellikle de kanatlarıdır. Ancak kuşları sınıflandırmak için eğitilmiş bir yapay zekâ uygulaması için arka plandaki renkler de en az kanatlar kadar önemli olabilir. Verilerdeki ufak tefek değişiklikler sebebiyle hata yapmalarının önüne geçmenin bir yolu, yapay zekâ uygulamalarını daha iyi eğitmek olabilir. Uygulamaları hataya sürükleyen örnekler eğitimlere dâhil edilerek daha az hata yapmaları sağlanabilir. Ancak bazı araştırmacılara göre, derin sinir ağlarını belirli türden saldırılar karşısında hata yapmalarını engelleyecek biçimde eğitmek, başka türden saldırılar karşısında daha savunmasız kalmalarına sebep olabilir.

Daha iyi bir çözüm, derin sinir ağlarının kullandığı algoritmaları, çeşitli matematiksel kısıtlar koyarak, verilerdeki ufak değişikliklerden etkilenmeyecekleri şekilde düzenlemek olabilir.

Derin öğrenme ile ilgili en temel sorunlardan biri algıladıkları görüntüleri anlamamaları. Örneğin, bir yapay zekâ uygulaması, üzerine birkaç kısa çizgi eklenmiş bir "Dur" işaretini "45" (hız sınırı) olarak algılayabiliyor. Bu durumun nedeni uygulamanın trafik levhalarındaki harfleri ve rakamları algılayıp okumaması, sadece görüntülerdeki pikseller arasında çeşitli ilişkiler kurması. Bu sorunun bir çözümü derin sinir ağlarını "sembolik yapay zekâ" ile bir araya getirmek olabilir. Özellikle 1950-80 döneminde üzerinde yoğun araştırmalar yapılan sembolik yapay zekâda makinelere dünyanın birbirinden bağımsız nesnelerden oluştuğu ve bu nesnelerin birbirleriyle hangi biçimlerde ilgili olduğu öğretiliyor.

Yapay zekânın eğitim biçimleri ne kadar geliştirilse geliştirilsin, bir uygulamanın kendine öğretilenlerden daha fazlasını yapması beklenemez. Bu yüzden sadece eğitim yöntemlerinin değil verilen bilgilerin de zenginleştirilmesi gerekiyor. Örneğin farklı açılardan fotoğrafı çekilmiş "Dur" işareti levhalarını yapay zekâ uygulamalarının gülleye ya da rakete benzetmesinin sebebi dünyanın üç boyutlu olduğunu bilmemeleri. Sadece iki boyutlu fotoğraflara bakarak bu nesnelerin üç boyutlu yapısı hakkında bir fikir edinemiyorlar. Yaptıkları hatalar da bu durumdan kaynaklanıyor. Bu yüzden, gerçek ya da sanal üç boyutlu ortamda eğitim vermek yapay zekânın dünyayı daha iyi kavramasını sağlayacaktır.

Yapay zekâyı nedensellik (olaylar arasındaki sebep sonuç ilişkileri) hakkında bilgilendirmenin yolu, uygulamalara deneyler ve keşif yapma imkânı vermekten geçiyor. Bugün bilgisayar oyunları oynayan yapay zekâ uygulamaları zaten "sanal ortamda" bu şekilde eğitiliyorlar. Uygulamalara önce bir "amaç" veriliyor, daha sonra uygulama oyunun kurallarıyla uyumlu çeşitli hamleler yaparak deneme yanılma yoluyla bu amaca ulaşmak için neler yapması gerektiğini öğreniyor.

Gerçek dünya sanal ortamdan çok daha zengindir. Bugün Berkeley'deki Kaliforniya Üniversitesinden araştırmacılar bir robotik kol yardımıyla derin öğrenme yöntemiyle yapay zekâyı aletler kullanması için eğitiyorlar. Robot, kendisine verilen alet kutusundaki aletleri eline alıyor, inceliyor ve ne işe yaradıklarını anlamaya çalışıyor. Bu tarz bir eğitim, yapay zekâya sadece iki boyutlu fotoğraflarla öğrenebileceğinden çok daha fazlasını veriyor.

Şu an için robotlara gerçek dünyada eğitim vermekle ilgili en önemli sorun bu eğitim sürecinin sanal ortamdakine göre çok yavaş olması. 2017 yılında Alpha Zero isimli bir yapay zekâ uygulaması sadece bir gün içinde Go, satranç ve shogi (bir tür Japon satrancı) öğrenmiş ve bu sırada sanal ortamda 20 milyondan fazla antrenman maçı yapmıştı. Bir robotun gerçek dünyada 20 milyon antrenman maçı yaparak eğitilmesiyse neredeyse imkânsızdır.

Yapay zekâyı gerçek dünyada eğitmek için, tıpkı insanlar gibi, daha az veriyle öğrenmelerini sağlamak gerekiyor. Bir bebeğin, bir hayvanın aslan mı yoksa kaplan mı olduğunu öğrenmesi için milyonlarca aslan ve kaplan fotoğrafı görmesi gerekmez. Sadece birkaç örnek yeterlidir. Çünkü daha önceden pek çok canlı görmüştür ve bir tür canlıyı diğer türdeki canlılardan ayıran özellikleri fark etmekte zorlanmaz.

Yapay zekâyı daha hızlı eğitmek için daha önceki tecrübelerden edindiği bazı ya da bütün bilgileri yeni görevlere aktarması sağlanabilir. Örneğin bir hayvan türünü tanıması için eğitilmiş derin sinir ağındaki gövde biçimini tanıyan kısımlar, başka bir hayvanı tanımak için verilecek yeni bir eğitim sırasında yararlı olabilir. Ancak bu yaklaşım her durumda geçerli değildir. Örneğin bir derin sinir ağını satranç ya da briç oynamak için eğitebilirsiniz. Ancak ikisini aynı anda oynayamaz. Birbirinden tamamen farklı kurallara sahip bu iki oyundan birini öğrenebilmesi için diğerini tamamen unutması gerekir. Yapay sinir ağları, insan beyinleri gibi hafızalara sahip değildir.

Bugün yapay zekâyı eğitmek için kullanılan derin öğrenmeyle ilgili derin sorunlar var. Verilerdeki ufak tefek oynamalar beklenmedik hatalar yapmalarına sebep olabiliyor. Hatta olmayan şeyleri görebiliyor, farklı açılardan çekilmiş fotoğrafları çok yanlış bir biçimde sınıflandırabiliyor, olmayan hastalıkları teşhis edebiliyorlar. Yapay zekâ günlük hayatımızda giderek daha çok yer ediyor. Dolayısıyla bu ve benzeri hatalar gelecekte önemli sorunlara sebep olabilir. Örneğin otonom araçların sabote edilmesi ya da önemli bir görevi üstlenen bir yapay zekâ uygulamasının hacklenmesi mümkün.

Bugün için derin öğrenme ile ilgili sorunların net bir çözümü yok. Bu durumun en önemli nedenlerinden biri konu ile ilgili kuramsal çalışmaların azlığı sebebiyle sorunların kaynağını tespit etmenin zorluğu. Ortaya çıkan sorunları çözmek için, sonucun ne olacağını bilmeden, çeşitli şeyleri denemek gerekiyor.

Gelecekte yapay zekâyı hatalardan arındırmak için yapılabilecek çeşitli şeyler var. Verilerdeki ufak tefek değişikliklerden etkilenmemelerini sağlamak için öğrenme algoritmalarına matematiksel kısıtlar koymak, üç boyutlu dünyayı ve olaylar arasındaki nedensellik ilişkilerini daha iyi anlamaları için gerçek dünyada eğitim vermek, daha az veriyle daha çok şey öğrenmelerini sağlayacak yöntemler bulmak gibi... Gelecekte yapay zekânın gerçekten de insanlar gibi öğrenmesini sağlamak için atılacak önemli bir adım, kendi öğrenme algoritmalarını geliştirmesine imkân vermek olabilir. Bilgisayar bilimciler yıllardır "program sentezi" olarak adlandırılan bilgisayarların otomatik olarak kod geliştirdiği bu alan üzerinde çalışmalar yapıyorlar. Kendi algoritmalarını geliştirme özelliği kazandırıp bir hafızayla donattıktan sonra gerçek dünyayı kendi kendilerine keşfetmelerine izin vermek, yapay zekânın aşina olmadığı durumlarla karşılaştığında basit hatalar yapmasını engellemenin bir yolu olabilir. ■

Kaynaklar

Heaven, Douglas, "Deep trouble for deep learning", *Nature*, Cilt 574, s. 163, 2019.

Szegedy, Christian, ve ark., "Intriguing properties of neural networks", https://arxiv.org/abs/1312.6199v1, 2013.

Nguyen, Anh, ve ark., "Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images", *IEEE Conf. Comp. Vision. Pattern Recog.*, p. 427-436, 2015.

Huang, Sandy, ve ark., "Adversarial Attacks on Neural Network Policies", https://arxiv.org/abs/1702.02284v1, 2017.

Eykholt, Kevin, ve ark., "Robust Physical-World Attacks on Deep Learning Visual Classification", *IEEE/CVF Conf. Comp. Vision. Pattern Recog.*, p. 1625-1634, 2018.

Hendrycks, Dan., ve ark., "Natural Adversarial Examples", https://arxiv.org/abs/1907.07174v2, 2019.