Makale: Machine Unlearning of Features and Labels

Warnecke, A., Pirch, L., Wressnegger, C., & Rieck, K. (2021). Machine unlearning of features and labels. arXiv preprint arXiv:2108.11577.

Özet:

Makale, makine öğrenmesi modellerinden eğitimden sonra hassas bilgilerin kaldırılması görevi olan "makine unutma" konusunu ele almaktadır. Mevcut unutma yaklaşımları tek tek veri noktalarını kaldırmaya odaklanmıştır ki bu hassas bilgilerin birçok veri noktasına yayıldığı durumlarda etkisiz hale gelmektedir.

Önerilen çerçeve, model parametrelerine kapalı formlu güncellemeler yaparak modelden tüm öznitelikler ve etiketlerin unutulmasına olanak tanır. Bu da yeniden eğitimden veya parça tabanlı unutmadan çok daha hızlı bir işlemdir.

Temel fikirler şunlardır:

1. Unutmayı, veri noktalarının ağırlıklarını yukarı ve aşağı ayarlayarak model parametrelerine etkilerini güncellemek olarak formüle etmek.
2. Veri değişikliklerini parametre güncellemelerine eşleyen birinci ve ikinci dereceden güncelleme kuralları türetmek.
3. Konveks kayıp fonksiyonları için, diferansiyel gizlilik kavramlarından yararlanarak "sertifikalı unutma" teorik garantileri sağlamak.

Deneysel analiz, yaklaşımın etkinliğini üç pratik senaryoda göstermektedir:

1. Lojistik regresyon spam, malware, nüfus verisi modellerinden hassas özellikleri kaldırmak. İkinci dereceden güncelleme, etkililik, sadakat ve verimlilik arasında en iyi dengeyi sağlar.
2. Dil modellerinden hassas dizelerin istenmeyen ezberletilmelerini kaldırmak. Birinci dereceden güncelleme en etkilidir, yeniden eğitime kıyasla birkac kat hızlıdır.
3. Görüntü sınıflandırıcılarına etiket zehirlenme saldırılarını, zehirli etiketlerin unutulmasıyla hafifletmek. Hem birinci dereceden güncelleme hem de ince ayar iyi çalışır.

Makale, çok büyük unutma isteklerinde etkinliğin azalması, non-konveks modeller için sertifikalı garantilerin olmaması, gizlilik sızıntılarının tespitinin önkoşul olması gibi sınırlamaları da tartışmaktadır. Genel olarak, tek veri noktalarının ötesinde etkili unutmaya olanak veren yeni bir çerçeve sunmaktadır.

Kavramlar:

1. **Machine Unlearning (Makine Unutması):** Makine öğrenmesi modellerinden hassas verilerin eğitimden sonra kaldırılması işlemidir. Bu, gizlilik sorunlarını çözmek ve yasal gerekliliklere uymak için gereklidir.
2. **Unlearning Features and Labels (Özelliklerin ve Etiketlerin Unutulması):** Hassas bilgilerin genellikle birçok veri noktasına yayılması nedeniyle, tüm özelliklerin ve etiketlerin unutulmasına olanak tanıyan bir yaklaşımdır. Bu, örnek tabanlı unutmaya göre daha verimli ve esnektir.
3. **Influence Functions (Etki Fonksiyonları):** Bir veri noktasının istatistiksel bir tahminci üzerindeki etkisini ölçen bir tekniktir. Makalede, eğitim verilerinin öğrenme modelleri üzerindeki etkisini değiştirmek için kullanılır.
4. **Closed-Form Updates (Kapalı Formda Güncellemeler):** Eğitim verilerindeki değişiklikleri doğrudan model parametrelerine eşleyen kompakt ve verimli güncellemelerdir. Birinci ve ikinci dereceden türevleri kullanarak, etkilenen verilerin etkisini telafi etmek için model parametrelerini ayarlarlar.
5. **Certified Unlearning (Sertifikalı Unutma):** Bir unutma yönteminin hassas verileri teorik olarak kanıtlanabilir bir şekilde kaldırdığını garanti eden bir kavramdır. Diferansiyel gizlilik ilkelerine dayanır ve unutma işleminin etkinliğini ölçmek için ε ve δ parametrelerini kullanır.
6. **First-Order and Second-Order Updates (Birinci ve İkinci Dereceden Güncellemeler):** Kapalı formda güncelleme yöntemleridir. Birinci dereceden güncelleme, kayıp fonksiyonunun gradyanını kullanırken, ikinci dereceden güncelleme Hessian matrisini kullanır. Her ikisi de hassas verilerin etkisini kaldırmak için model parametrelerini günceller.
7. **Convex Loss Functions (Konveks Kayıp Fonksiyonları):** Optimizasyon problemini basitleştiren ve global bir minimum garanti eden bir kayıp fonksiyonu sınıfıdır. Makalede, konveks kayıp fonksiyonları için sertifikalı unutma sağlanabildiği gösterilmiştir.
8. **Unintended Memorization (İstenmeyen Ezberlenme):** Makine öğrenimi modellerinin eğitim verilerindeki hassas bilgileri yanlışlıkla ezberlemesi ve bu bilgileri kullanıma sunmasıdır. Makale, istenmeyen ezberlemenin dil modellerinden nasıl kaldırılabileceğini gösterir.
9. **Data Poisoning (Veri Zehirleme):** Kötü niyetli bir saldırganın, bir makine öğrenmesi modelinin performansını bozmak için eğitim verilerine kasıtlı olarak hatalı veya yanıltıcı veriler eklemesidir. Makale, zehirlenmiş etiketlerin unutularak bu tür saldırıların etkisinin nasıl azaltılabileceğini gösterir.
10. **Lipschitz Continuity (Lipschitz Sürekliliği):** Bir fonksiyonun, girdilerindeki değişikliklere göre çıktılarının ne kadar hızlı değişebileceğini sınırlayan bir matematiksel özelliktir. Bu kavram, sertifikalı unutma için teorik garantiler sağlamak amacıyla kullanılır.
11. **Differential Privacy (Diferansiyel Gizlilik):** Bir veri kümesindeki bireylere ait bilgilerin gizliliğini korurken, veri kümesinin istatistiksel özelliklerini korumayı amaçlayan bir gizlilik modelidir. Sertifikalı unutma, diferansiyel gizlilik ilkelerine dayanmaktadır.
12. **Exposure Metric (Maruziyet Ölçütü):** Bir dizinin bir dil modeli tarafından üretilme olasılığını ölçen bir metriktir. Bu metrik, bir modeldeki istenmeyen ezberlemeyi tespit etmek ve unutma işleminin etkinliğini değerlendirmek için kullanılır.
13. **Retraining (Yeniden Eğitim):** Hassas verileri kaldırmak için modelin sıfırdan yeniden eğitilmesi işlemidir. Bu, etkili bir yöntemdir ancak hesaplama açısından maliyetli olabilir ve orijinal eğitim verilerinin saklanmasını gerektirir.
14. **Sharding:** Eğitim verilerini bağımsız parçalara (shard) ayırma ve her parça için ayrı bir model eğitme yöntemidir. Unutma işlemi, sadece etkilenen parçaların yeniden

eğitilmesiyle gerçekleştirilir. Bu yöntem, bazı senaryolarda hesaplama verimliliği sağlayabilir.

15. **Fine-tuning (İnce Ayar):** Unutulması istenen verileri içeren örnekler üzerinde modeli eğitmeye devam etme işlemidir. Bu, basit bir unutma stratejisidir ancak hassas bilgileri tamamen kaldırmada her zaman etkili olmayabilir.
16. **Fidelity (Bağlılık):** Unutma işleminden sonra modelin kalitesinin korunmasıdır. Etkili bir unutma yöntemi, hassas verileri kaldırırken modelin performansını mümkün olduğunca orijinal modele yakın tutmalıdır.
17. **Efficiency (Verimlilik):** Unutma yönteminin hesaplama verimliliğini ifade eder. Etkili bir yöntem, yeniden eğitimden önemli ölçüde daha hızlı olmalıdır.
18. **Non-convex Loss Functions (Konveks Olmayan Kayıp Fonksiyonları):** Birden fazla yerel minimum içerebilen ve global minimumu garanti etmeyen kayıp fonksiyonlarıdır. Derin sinir ağları genellikle konveks olmayan kayıp fonksiyonlarına sahiptir, bu da sertifikalı unutmayı zorlaştırır.
19. **Hessian Matrix (Hessian Matrisi):** Bir fonksiyonun ikinci dereceden kısmi türevlerini içeren kare matristir. İkinci dereceden güncelleme yönteminde, model parametrelerindeki güncellemeleri hesaplamak için kullanılır.
20. **Gradient Residual (Gradyan Artığı):** Bir modelin, düzeltilmiş bir veri kümesi üzerinde yeniden eğitilmiş bir modelden ne kadar farklı olduğunun bir ölçüsüdür. Gradyan artığı ne kadar küçükse, unutma işlemi o kadar etkilidir.

Önerilen Yöntem

Önerilen makine unutma yöntemi, eğitim verilerindeki özelliklerin ve etiketlerin etkisini model parametreleri üzerinde kapalı formda güncellemeler yoluyla değiştirmeye odaklanır. Yöntemin temel adımları şunlardır:

1. Unutulması istenen verileri (özellikler veya etiketler) içeren veri noktaları kümesi Z ve bu veri noktalarının düzeltilmiş versiyonları $\tilde{Z}$ belirlenir.

2. Orijinal ve düzeltilmiş veri noktaları arasındaki farkı yakalamak için bir etki fonksiyonu tanımlanır. Bu, veri noktalarının model parametreleri üzerindeki etkisini artırma ve azaltma yoluyla yapılır.

3. Eğitim verileri üzerindeki değişiklikleri model parametrelerindeki güncellemelere eşleyen kapalı formda güncelleme adımları türetilir: a) Birinci dereceden güncelleme: Kayıp fonksiyonunun gradyanını kullanır ve herhangi bir türevlenebilir kayıp fonksiyonuna uygulanabilir. b) İkinci dereceden güncelleme: Kayıp fonksiyonunun Hessian matrisini kullanır ve güçlü konveks, iki kez türevlenebilir kayıp fonksiyonlarını gerektirir.

4. Güncelleme adımları model parametrelerine uygulanarak, etkilenen verilerin etkisi azaltılırken düzeltilmiş verilerin etkisi artırılır. Bu, hassas bilgilerin modelden etkili bir şekilde "unutulmasını" sağlar.

5. Güçlü konveks kayıp fonksiyonları için, diferansiyel gizlilik kavramlarından yararlanarak "sertifikalı unutma" sağlanabilir. Bu, unutma işleminin hassas verileri teorik olarak güvence altına aldığını gösterir.

Önerilen yöntem, hem özellikleri hem de etiketleri etkin bir şekilde unutabilir ve diğer yaklaşımlara kıyasla önemli hesaplama kazançları sağlar. Konveks modeller için teorik garantiler sunarken, konveks olmayan modeller (örn. derin sinir ağları) için de deneysel olarak başarılı sonuçlar verir. Böylece, makine öğrenmesi modellerinden hassas bilgileri kaldırmak için güçlü ve esnek bir çerçeve sağlar.

1. **Yöntem: Unutmayı, veri noktalarının ağırlıklarını aşağı ve yukarı ayarlayarak model parametrelerine etkilerini güncellemek olarak formüle etmek.**

Bu fikir şu şekilde açıklanabilir:

Bir makine öğrenmesi modelindeki belirli bir veri noktasının etkisini kaldırmak yerine, önerdikleri yaklaşım veri noktasının modeldeki ağırlığını/etkisini azaltmak ve aynı zamanda bu veri noktasının düzeltilmiş versiyonunun ağırlığını/etkisini artırmaktır.

Yani diyelim ki bir veri noktası (x, y) gizli bir kredi kartı numarası içeriyorsa, bu yaklaşım şunu yapar:

- (x, y) veri noktasının modele olan etkisini azaltır (ağırlığını düşürür)
- (x', y') olarak düzeltilmiş veri noktasının (kredi kartı numarası sansürlenmiş) modele olan etkisini artırır (ağırlığını yükseltir)

Böylece modeldeki parametreler, gizli bilgiyi içeren orijinal veri noktasından uzaklaşır ve düzeltilmiş veri noktasına daha yakın hale gelir.

Bu işlem, tekil veri noktaları için yapılabileceği gibi, bir grup veri noktası için de gerçekleştirilebilir. Dolayısıyla öznitelik ve etiketlerdeki büyük ölçekli değişiklikleri de kapsar.

İlgili denklemi kullanarak, bu ağırlık güncellemelerini kapalı formlu bir parametre güncellemesine dönüştürebilirler. Böylece veriyi açıkça kaldırmak yerine, etkisini azaltıp düzeltilmiş halinin etkisini artırarak model parametrelerini güncellerler.

2. **Yöntem: Veri değişikliklerini parametre güncellemelerine eşleyen birinci ve ikinci dereceden güncelleme kuralları türetmek.**

Bu fikir, makale tarafından önerilen iki farklı yöntemi açıklamaktadır:

a) Birinci Dereceden Güncelleme: Bu yöntem, kayıp fonksiyonunun gradyanını kullanarak bir güncelleme kuralı türetir. Eğer kayıp fonksiyonu türevlenebilir ise, aşağıdaki güncelleme denklemi uygulanabilir:

$\Delta(Z, \tilde{Z}) = -\tau * (\Sigma(\tilde{z} \epsilon \tilde{Z}) \nabla\theta \ell(\tilde{z},\theta^*) - \Sigma(z \epsilon Z) \nabla\theta \ell(z,\theta^*))$

Burada τ bir öğrenme oranı, Z orijinal veri noktaları, $\tilde{Z}$ ise düzeltilmiş veri noktalarıdır. $\Delta(Z, \tilde{Z})$ ise model parametrelerine eklenecek güncellemedir.

Bu güncelleme, kayıp fonksiyonunun gradyanındaki değişikliği yakalamakta ve bu değişikliği parametre güncellemesine dönüştürmektedir. Böylece verinin unutulması, parametre güncellemesi ile gerçekleştirilir.

b) İkinci Dereceden Güncelleme: Bu yöntem ise, eğer kayıp fonksiyonu iki kez türevlenebilir ve kesin konveks ise uygulanabilir. Hessian matrisi ve tersi kullanılarak şu güncelleme elde edilir:

$$\Delta(Z, \tilde{Z}) = -H(\theta^*)^{-1} * (\Sigma(\tilde{z}\epsilon\tilde{Z}) \nabla\theta\ell(\tilde{z},\theta^*) - \Sigma(z\epsilon Z) \nabla\theta\ell(z,\theta^*))$$

Burada $H(\theta^*)$ kayıp fonksiyonunun Hessian matrisidir. Bu güncelleme, öğrenme oranı parametresine ihtiyaç duymaz. Ancak Hessian matrisinin tersinin alınması gerekir ki bu zahmetlidir.

Her iki güncellemede de temel mantık, orijinal ve düzeltilmiş veri arasındaki kayıp/gradyan farkını parametre güncellemesine dönüştürmektir. Böylece verinin etkisi modellerden kaldırılabilir.

3. Konveks kayıp fonksiyonları için, diferansiyel gizlilik kavramlarından yararlanarak "sertifikalı unutma" teorik garantileri sağlamak.

Bu fikir, makine unutma yöntemlerinin güvenilirliğini ve hassas verileri ne kadar iyi kaldırdığını teorik olarak kanıtlamayı amaçlamaktadır. Bunun için diferansiyel gizlilik kavramından yararlanılmaktadır.

Diferansiyel gizlilik, bir algoritmanın çıktılarının, girdi verilerindeki küçük değişikliklere karşı ne kadar duyarlı olduğunu ölçer. Eğer algoritma girdilerdeki değişikliklere karşı duyarlı değilse, bu o algoritmanın gizliliği iyi koruduğu anlamına gelir.

Makalede, veri kaldırma/unutma işleminin başarısını ölçmek için ε-sertifikalı unutma ve (ε,δ)-sertifikalı unutma kavramları tanımlanmaktadır.

ε-sertifikalı unutma, veriyi kaldırmak için kullanılan algoritma A ve unutma yöntemi U için şu koşulu sağlamalıdır:

$$e^{-\varepsilon} \leq P(U(A(D),D,D'))/ P(A(D')) \leq e\varepsilon$$

Yani, D veri seti üzerinde eğitilen A algoritması ile D' düzeltilmiş veri seti üzerinde eğitilmiş A(D') arasındaki olasılık farkı en fazla ε kadardır.

(ε,δ)-sertifikalı unutma ise bu koşulu biraz gevşetir ve δ kadar sapma payı verir.

Makale, ikinci dereceden güncelleme kuralı için bu sertifikalı unutma koşullarının sağlanabildiğini göstermektedir. Bunun için:

1. Kayıp fonksiyonunun kesin konveks ve türevlenebilir olması
2. L2 düzenlileştirme kullanılması
3. Diferansiyel gizlilikte kullanıldığı gibi, bir rassal gürültü terimi b eklenmesi gerekmektedir.

Uygun ε, δ ve b seçimleri ile, ikinci dereceden güncellemenin ε-sertifikalı veya (ε,δ)-sertifikalı unutma koşullarını sağladığı gösterilmiştir.

Böylece konveks kayıp fonksiyonu olan modeller için, hassas verilerin ne kadar iyi kaldırıldığına dair teorik bir garanti elde edilebilmektedir. Bu da makine öğrenmesinde gizliliğin matematiksel olarak analiz edilebilmesini sağlar.

Senaryo 1: Hassas Özelliklerin Unutulması

Bu senaryo, güçlü konveks kayıp fonksiyonuna sahip doğrusal modellerde hassas özelliklerin kaldırılmasına odaklanmaktadır. Deneyler, spam filtreleme, Android kötücül yazılım tespiti, diyabet tahmini ve nüfus verileri kullanılarak gelir tahmini için gerçek dünya verileri üzerinde lojistik regresyon kullanılarak gerçekleştirilmiştir.

Her veri seti, %80 eğitim ve %20 test olacak şekilde ayrılmıştır. Özellik uzayını oluşturmak için sayısal özellikler olduğu gibi kullanılmış veya kelime torbası (bag-of-words) yöntemi uygulanarak özellik vektörleri oluşturulmuştur.

Hassas özellikler şu şekilde belirlenmiştir:

- Yüksek boyutlu ve seyrek veri içeren veri setleri (Spam ve Kötücül Yazılım) için, e-postalardaki kişisel isimler veya Android uygulamalarından çıkarılan URL'ler gibi tüm özellikler kaldırılmıştır.
- Düşük boyutlu ve yoğun veri setleri (Yetişkin ve Diyabet) için, medeni durum, cinsiyet, ırk, yaş, vücut kitle indeksi gibi seçilen özellik değerleri 0 ile değiştirilerek ayrımcı yanlılığın kaldırılması hedeflenmiştir.
- Hassas özellikler belirlendikten sonra, farklı unutma yöntemleri uygulanmıştır. Önerilen ikinci dereceden güncelleme yöntemi, lojistik regresyonun konveks kayıp fonksiyonundan yararlanarak sertifikalı unutmayı mümkün kılmıştır.

Yöntemlerin etkinliği, bağlılığı ve verimliliği değerlendirilmiştir:

- Etkinlik için gradyan artık normları incelenmiş, ikinci dereceden güncellemenin diğer yöntemlere göre daha iyi performans gösterdiği bulunmuştur.
- Bağlılık için, yeniden eğitim ve unutma arasındaki kayıp farkı ve test doğruluğu ölçülmüştür. İkinci dereceden güncelleme, orijinal modele en yakın sonuçları vermiştir.
- Verimlilik açısından, birinci dereceden güncelleme en hızlı, ikinci dereceden güncelleme ise yeniden eğitimden 4 kat daha hızlı bulunmuştur.

Sonuç olarak, konveks kayıp fonksiyonuna sahip modellerde hassas özellikler teorik garantilerle kaldırılabilir ve ikinci dereceden güncelleme, etkinlik, bağlılık ve verimlilik arasında en iyi dengeyi sağlar.

Senaryo 2: İstenmeyen Ezberlemelerin Unutulması

Bu senaryo, dil modellerindeki istenmeyen ezberlemelerin kaldırılmasına odaklanmaktadır. Araştırmalar, dil modellerinin eğitim verilerindeki nadir girdileri ezberleyebildiğini ve uygulama sırasında bunları tam olarak yeniden üretebildiğini göstermiştir. Üretilen veriler kredi

kartı numaraları veya telefon numaraları gibi özel bilgiler içeriyorsa, bu bir gizlilik sorunu haline gelebilir.

Deneyler, Alice Harikalar Diyarında romanı kullanılarak karakter düzeyinde eğitilen bir LSTM ağı üzerinde gerçekleştirilmiştir. Modelin istenmeyen ezberlemeler yapması için, eğitim verilerine "Telefon numaram (s)! dedi Alice" şeklinde bir cümle (canary) yerleştirilmiştir. Burada "(s)", farklı uzunluklarda (5, 10, 15, 20) olan ve farklı sayıda (200, 500, 1000, 2000) veri noktasını etkileyen rakamlardan oluşan bir dizidir.

Ezberlenmiş dizilerin modelden başarıyla unutulup unutulmadığını değerlendirmek için maruz kalma (exposure) metriği kullanılmıştır. Bu metrik, verilen bir dizinin model tarafından üretilme olasılığını ölçer. Birinci ve ikinci dereceden güncellemeler, tüm dizi uzunlukları için sıfıra yakın maruz kalma değerleri vermiş, böylece hassas bilgilerin çıkarılmasını imkansız kılmıştır.

Bağlılık açısından, küçük değişiklikler için yeniden eğitime yakın sonuçlar elde edilmiştir. Ancak etkilenen veri noktası sayısı arttıkça, birinci ve ikinci dereceden güncellemelerin doğruluğunda kademeli bir düşüş gözlenmiştir.

Verimlilik açısından, birinci dereceden güncelleme en hızlı yaklaşım olup, yeniden eğitime kıyasla üç kat hızlanma sağlamıştır. İkinci dereceden güncelleme, GPU hızlandırması olmamasına rağmen yeniden eğitimden 28 kat daha hızlı olmuştur.

Sonuç olarak, istenmeyen ezberlemeler özellikler ve etiketler unutularak dil modellerinden kaldırılabilir. Birinci dereceden güncelleme, etkinlik, bağlılık ve verimlilik arasında en iyi dengeyi sağlar. Bu, ezberlemelerin dil modellerinde çok derin olmadığını gösterir.

Senaryo 3: Zehirlenmiş Etiketlerin Unutulması

Bu senaryo, bilgisayarlı görü alanında bir zehirleme saldırısını onarmaya odaklanmaktadır. Saldırıda, bir saldırgan eğitim verilerindeki etiketleri kısmen değiştirerek sınıflar arasında geçiş yapmaktadır. Bu saldırı gizliliği etkilemese de, özellikleri değiştirmeden güvenlik tehdidi oluşturduğu için unutma açısından ilginç bir senaryodur.

Deneylerde, CIFAR10 veri kümesi kullanılmış ve üç VGG bloğu ile iki yoğun katmandan oluşan 1,8 milyon parametreye sahip bir evrişimli sinir ağı eğitilmiştir. Ağ, zehirleme olmadan %87 doğruluk oranına ulaşmıştır. Ancak saldırı altında, doğruluk oranı ortalama %10 düşmüştür.

Zehirleme saldırısında, sınıf çiftleri belirlenerek, etiketlerin bir kısmı karşılıklarına çevrilmiştir (örneğin, "kedi"den "kamyon"a). Etiketler, belirli bir bütçeye (zehirlenmiş etiket sayısı) ulaşılana kadar orijinal eğitim verilerinden rastgele örneklenmiştir. Bu strateji, rastgele etiketler eklemekten daha etkili bulunmuştur.

Farklı unutma yöntemleri, rastgele seçilmiş etiketlerle beş deneysel çalışmada uygulanmıştır. Zehirlenmiş etiketlerin tümünü tek bir kapalı formlu güncellemede düzeltmek, bellek kısıtlamaları nedeniyle zor olduğundan, unutma işlemi 512 örnekten oluşan gruplar halinde

gerçekleştirilmiştir. Ayrıca, nihai tahminden esas olarak sorumlu olan tam bağlantılı katmanlar güncellenmiştir.

Sonuçlar şunu göstermiştir:

- Hiçbir yaklaşım zehirleme saldırısının etkisini tamamen ortadan kaldıramamıştır. Ancak birinci ve ikinci dereceden güncellemeler ile ince ayar, 2500 zehirlenmiş etiket için orijinal performansa yaklaşmıştır.
- Etkilenen etiket sayısı arttıkça, tüm yöntemlerde sürekli bir performans düşüşü gözlenmiştir. 10.000 etiketin manipülasyonu hiçbir yöntemle yeterince geri alınamamıştır.
- Birinci dereceden güncelleme ve ince ayar çok verimli olup yaklaşık 10 saniyede hesaplanabilirken, yeniden eğitim 15 dakikadan fazla sürmüştür.
- İkinci dereceden güncelleme biraz daha yavaş olmasına rağmen, yeniden eğitimden 100 kat daha hızlıdır. Ölçeklendirme deneyleri, artan model boyutları için birinci ve ikinci dereceden güncellemelerin doğrusal bir çalışma zamanı gösterdiğini ortaya koymuştur.

Sonuç olarak, etiket zehirlenmesi etiketler unutularak hafifletilebilir. Birinci dereceden güncelleme ve ince ayar, etkinlik (= bağlılık) ve verimlilik arasında en iyi dengeyi sağlar.

Sınırlamalar

Üç ana sınırlama vurgulanmıştır:

1. Unutmanın Sınırları:

   - Unutmanın etkinliği, etkilenen özellik ve etiket sayısı arttıkça azalır. Yüzlerce hassas özellik ve binlerce etiket içeren gizlilik sızıntıları bu yaklaşımla iyi bir şekilde ele alınabilirken, milyonlarca veri noktasını değiştirmek yaklaşımın kapasitesini aşmaktadır.
   - Eğer yaklaşım herhangi büyüklükte bir değişikliği düzeltebilseydi, tüm öğrenme algoritmaları için bir "yerinde değiştirme" olarak kullanılabilirdi ki bu açıkça imkansızdır.
   - Yine de, makul sayıda veri noktasının düzeltilmesi gerektiğinde, yaklaşım yeniden eğitim ve parçalama yöntemlerine göre önemli bir hızlanma sağlar.

2. Konveks Olmayan Kayıp Fonksiyonları:

   - Önerilen yaklaşım, yalnızca güçlü konveks kayıp fonksiyonları ve Lipschitz sürekli gradyanları olan fonksiyonlar için sertifikalı unutma garantisi verebilir.
   - Her iki güncelleme adımı da konveks olmayan kayıp fonksiyonlarına sahip sinir ağları için iyi çalışsa da (deneysel değerlendirmede gösterildiği gibi), unutma başarısını doğrulamak için ek bir ölçüme ihtiyaç duyarlar.
   - Neyse ki, bu tür harici ölçümler genellikle mevcuttur, çünkü tipik olarak kaldırılmadan önce veri sızıntısını karakterize etmenin temelini oluştururlar. Benzer şekilde, bir zehirleme saldırısının nasıl düzeltildiğini ölçmek için bağlılık ölçütü kullanılmıştır.

3. Unutma Tespiti Gerektirir:

    - Önerilen yöntem, kaldırılacak verilerin bilinmesini gerektirir. Öğrenme modellerindeki gizlilik sızıntılarını tespit etmek, bu çalışmanın kapsamı dışında olan zor bir problemdir.
    - Gizlilik sızıntılarının doğası, ele alınan verilere, öğrenme modellerine ve uygulamaya bağlıdır. Örneğin, Carlini ve diğerlerinin analizi üretici öğrenme modellerindeki sıralı verilere odaklanmaktadır ve diğer öğrenme modellerine veya görüntü verilerine kolayca aktarılamaz.
    - Sonuç olarak, bu yöntem sızıntıları bulmaktan ziyade onarmakla sınırlıdır.

Bu sınırlamalar, önerilen makine unutma yaklaşımının uygulanabilirliğini ve geçerliliğini etkileyebilecek önemli pratik hususları vurgulamaktadır. Bununla birlikte, yazarlar yaklaşımlarının hala birçok yaygın senaryoda etkili olduğunu ve gelecekteki araştırmalar için sağlam bir temel sağladığını öne sürmektedirler.

Sonuç

Makale, örnekler üzerinde çalışan unutmaya odaklanan mevcut yaklaşımların sınırlamalarını ele almak için bir çözüm olarak özellikler ve etiketler için yeni bir makine unutma çerçevesi önermektedir. Bu çerçeve, eğitim verilerinin model parametreleri üzerindeki etkisini, etkisi fonksiyonları kavramına dayanarak kapalı formda güncellemelerle yakalar. Bu da diğer yaklaşımlara göre önemli hız artışları sağlar.

Önerilen yaklaşımın etkinliği teorik ve deneysel olarak analiz edilmiştir:

- Diferansiyel gizlilik kavramına dayanarak, güçlü konveks kayıp fonksiyonuna sahip modeller için çerçevenin sertifikalı unutmayı sağladığı kanıtlanmıştır.
- Üç pratik senaryoda (hassas özelliklerin kaldırılması, istenmeyen ezberlemelerin unutulması ve zehirlenmiş etiketlerin düzeltilmesi) unutma stratejisinin faydaları değerlendirilmiştir.
- Özellikle, üretken dil modelleri için, modellerin işlevselliğini koruyarak istenmeyen ezberlemelerin kaldırılabildiği gösterilmiştir.

Yazarlar, bu çalışmanın makine unutma üzerine yapılacak gelecek araştırmaları teşvik etmesini ve makine öğrenmesinde gizliliğe ilişkin teorik sınırları keskinleştirmesini ummaktadır. Bu gelişmeyi desteklemek için tüm uygulamalarını ve veri kümelerini açık kaynak olarak sundukları bir GitHub deposu paylaşmışlardır.

Github linki: GitHub - alewarne/MachineUnlearning: Code related to the paper "Machine Unlearning of Features and Labels"