Büyük Ölçekli Zayıf Denetim ile Güçlü Konuşma Tanıma

Alec Radford*1 Jong Wook Kim*1 Tao Xu1 Greg Brockman1 Christine McLeavey1 Ilya Sutskever1

Özet

İnternetteki büyük miktarlarda ses transkriptlerini tahmin etmek için eğitilmiş konuşma işleme sistemlerinin yeteneklerini inceliyoruz. 680.000 saatlik çok dilli ve çok görevli denetim ile ölçeklendirildiğinde, ortaya çıkan modeller standart kıyaslama testlerine iyi genelleme yapar ve herhangi bir ince ayar gerektirmeden sıfır atış transfer ayarında önceki tam denetimli sonuçlarla genellikle rekabet edebilir. İnsanlarla karşılaştırıldığında, modeller onların doğruluğuna ve sağlamlığına yaklaşır. Güçlü konuşma işleme üzerine daha fazla çalışma için temel oluşturmak amacıyla modelleri ve çıkarım kodunu yayınlıyoruz.

1. Giriş

Konuşma tanımadaki ilerleme, Wav2Vec 2.0 (Baevski ve diğerleri, 2020) tarafından örneklenen denetimsiz ön eğitim tekniklerinin geliştirilmesiyle hızlanmıştır. Bu yöntemler insan etiketlerine ihtiyaç duymadan doğrudan ham sesten öğrendikleri için, etiketlenmemiş konuşmanın büyük veri setlerini verimli bir şekilde kullanabilir ve 1.000.000 saatlik eğitim verisine kadar hızla ölçeklendirilmiştir (Zhang ve diğerleri, 2021), bu da tipik bir akademik denetimli veri setinin yaklaşık 1.000 saatinden çok daha fazladır. Standart kıyaslama testlerinde ince ayar yapıldığında, bu yaklaşım özellikle düşük veri ortamında sanatın durumunu iyileştirmiştir.

Bu ön eğitilmiş ses kodlayıcıları konuşmanın yüksek kaliteli temsillerini öğrenir, ancak tamamen denetimsiz oldukları için bu temsilleri kullanılabilir çıktılara eşleyen eşit derecede performanslı bir kod çözücüden yoksundurlar, bu da konuşma tanıma gibi bir görevi gerçekten gerçekleştirmek için bir ince ayar aşaması gerektirir[1]. Bu ne yazık ki kullanışlılıklarını ve etkilerini sınırlar çünkü ince ayar hala yetenekli bir uygulayıcı gerektiren karmaşık bir süreç olabilir. Ince ayar gerektirmenin ek bir riski vardır. Makine öğrenmesi yöntemleri, aynı veri setinden tutulan veriler üzerinde performansı artıran bir eğitim veri seti içindeki kalıpları bulmada son derece yeteneklidir. Ancak, bu kalıpların bazıları kırılgan ve sahte olup diğer veri setlerine ve dağılımlara genelleme yapmaz. Özellikle rahatsız edici bir örnekte, Radford ve diğerleri (2021), ImageNet veri setinde (Russakovsky ve diğerleri, 2015) bir bilgisayar görüşü modelini ince ayar yaparken nesne sınıflandırma doğruluğunda %9,2'lik bir artış belgelemiş, ancak aynı nesneleri yedi diğer doğal görüntü veri setinde sınıflandırırken ortalama doğrulukta herhangi bir iyileşme gözlemlememiştir. Bir veri setinde eğitildiğinde "insanüstü" performans elde eden bir model, başka bir veri setinde değerlendirildiğinde hala birçok temel hata

yapabilir, muhtemelen tam da insanların fark etmediği bu veri setine özgü tuhaflıkları sömürdüğü için (Geirhos ve diğerleri, 2020).

Bu, denetimsiz ön eğitimin ses kodlayıcılarının kalitesini dramatik olarak iyileştirmiş olmasına rağmen, eşit derecede yüksek kaliteli ön eğitilmiş bir kod çözücünün eksikliğinin, veri setine özgü ince ayar önerilen protokolüyle birleştiğinde, kullanışlılıklarını ve sağlamlıklarını sınırlayan kritik bir zayıflık olduğunu önerir. Bir konuşma tanıma sisteminin amacı, her dağıtım dağılımı için bir kod çözücünün denetimli ince ayarını gerektirmeden geniş bir ortam yelpazesinde "kutunun dışında" güvenilir bir şekilde çalışmak olmalıdır.

Narayanan ve diğerleri (2018), Likhomanenko ve diğerleri (2020) ve Chan ve diğerleri (2021) tarafından gösterildiği gibi, birçok veri seti/alan boyunca denetimli bir şekilde ön eğitilmiş konuşma tanıma sistemleri, tek bir kaynakta eğitilmiş modellerden daha yüksek sağlamlık sergiler ve tutulan veri setlerine çok daha etkili bir şekilde genelleme yapar. Bu çalışmalar bunu mümkün olduğunca çok sayıda mevcut yüksek kaliteli konuşma tanıma veri setini birleştirerek başarır. Ancak, kolayca erişilebilir bu verinin hala yalnızca orta miktarda olması vardır. SpeechStew (Chan ve diğerleri, 2021) toplam 5.140 saatlik denetim sağlayan 7 önceden var olan veri setini karıştırır. Önemsiz olmamakla birlikte, bu hala Zhang ve diğerleri (2021)'de kullanılan daha önce bahsedilen 1.000.000 saatlik etiketlenmemiş konuşma verisine kıyasla çok küçüktür.

Mevcut yüksek kaliteli denetimli veri setlerinin sınırlayıcı boyutunu fark eden son çabalar, konuşma tanıma için daha büyük veri setleri oluşturmuştur. Altın standart insan doğrulamalı transkriptler gereksinimini gevşeterek, Chen ve diğerleri (2021) ve Galvez ve diğerleri (2021) zayıf denetimli konuşma tanımayı 10.000 ve 30.000 saatlik daha gürültülü eğitim verisine ölçeklendirmek için sofistike otomatik boru hatları kullanır. Kalite ve miktar arasındaki bu değiş tokuş genellikle doğru çağrıdır. Konuşma tanıma için şimdiye kadar az çalışılmış olmasına rağmen, bilgisayar görüşündeki son çalışmalar, ImageNet (Russakovsky ve diğerleri, 2015) gibi altın standart kitle kaynaklı veri setlerinin ötesine geçerek çok daha büyük ancak zayıf denetimli veri setlerine geçmenin modellerin sağlamlığını ve genellemesini önemli ölçüde iyileştirdiğini göstermiştir (Mahajan ve diğerleri, 2018; Kolesnikov ve diğerleri, 2020).

Yine de bu yeni veri setleri mevcut yüksek kaliteli veri setlerinin toplamından yalnızca birkaç kat daha büyüktür ve önceki denetimsiz çalışmalardan hala çok daha küçüktür. Bu çalışmada bu boşluğu kapatıyor, zayıf denetimli konuşma tanımayı bir sonraki büyüklük mertebesine 680.000 saatlik etiketli ses verisine ölçeklendiriyoruz. Yaklaşımımızı Whisper[2] olarak adlandırıyoruz. Bu ölçekte eğitilmiş modellerin mevcut veri setlerine sıfır atış iyi transfer olduğunu, yüksek kaliteli sonuçlar elde etmek için herhangi bir veri setine özgü ince ayar ihtiyacını ortadan kaldırdığını gösteriyoruz.

Ölçeğe ek olarak, çalışmamız aynı zamanda zayıf denetimli ön eğitimin kapsamını yalnızca İngilizce konuşma tanımanın ötesine genişleterek hem çok dilli hem de çok görevli olmaya odaklanır. Bu 680.000 saatlik sesin 117.000 saati 96 diğer dili kapsar. Veri seti ayrıca 125.000 saatlik X→en çeviri verisini içerir. Yeterince büyük modeller için ortak çok dilli ve çok görevli eğitimde hiçbir dezavantaj olmadığını ve hatta faydalar olduğunu buluyoruz.

Çalışmamız, zayıf denetimli ön eğitimin basit ölçeklendirmesinin konuşma tanıma için şimdiye kadar yeterince takdir edilmediğini önerir. Bu sonuçları, son büyük ölçekli konuşma tanıma çalışmalarının temel taşı olan kendi kendine denetim veya kendi kendine eğitim tekniklerine ihtiyaç duymadan elde ediyoruz. Güçlü konuşma tanıma üzerine daha fazla araştırma için temel oluşturmak amacıyla, aşağıdaki URL'de çıkarım kodu ve modelleri yayınlıyoruz: https://github.com/openai/whisper.

2. Yaklaşım

2.1. Veri İşleme

Makine öğrenmesi sistemlerini eğitmek için internetten web ölçekli metin kullanan son çalışmaların eğilimini takip ederek, veri ön işlemeye minimalist bir yaklaşım benimseriz. Konuşma tanıma üzerine birçok çalışmanın aksine, Whisper modellerini herhangi bir önemli standardizasyon olmadan transkriptlerin ham metnini tahmin etmek için eğitiriz, sıra-sıra modellerinin ifade gücüne güvenerek söylemler ve onların transkript edilmiş formu arasında eşleme öğrenmelerini sağlarız. Bu, doğal transkriptler üretmek için ayrı bir ters metin normalleştirme adımı ihtiyacını ortadan kaldırdığı için konuşma tanıma boru hattını basitleştirir.

Veri setini internette transkriptlerle eşleştirilmiş sesten oluştururuz. Bu, birçok farklı ortam, kayıt düzeneği, konuşmacı ve dilden gelen geniş bir ses dağılımını kapsayan çok çeşitli bir veri seti ile sonuçlanır. Ses kalitesindeki çeşitlilik bir modelin sağlam olmasını eğitmeye yardımcı olabilirken, transkript kalitesindeki çeşitlilik benzer şekilde faydalı değildir. İlk inceleme ham veri setinde büyük miktarda standart altı transkript gösterdi. Bunu ele almak için, transkript kalitesini iyileştirmek için çeşitli otomatik filtreleme yöntemleri geliştirdik.

İnternetteki birçok transkript aslında insan tarafından üretilmemiş, mevcut ASR sistemlerinin çıktısıdır. Son araştırmalar, insan ve makine tarafından üretilmiş verilerin karışık veri setlerinde eğitimin çeviri sistemlerinin performansını önemli ölçüde bozabileceğini göstermiştir (Ghorbani ve diğerleri, 2021). "Transkript-ce" öğrenmekten kaçınmak için, eğitim veri setinden makine tarafından üretilmiş transkriptleri tespit etmek ve kaldırmak için birçok buluşsal yöntem geliştirdik. Mevcut birçok ASR sistemi, yalnızca ses sinyallerinden tahmin etmesi zor olan karmaşık noktalama işaretleri (ünlem işaretleri, virgüller ve soru işaretleri), paragraflar gibi biçimlendirme boşlukları veya büyük harf kullanımı gibi stilistik yönleri kaldıran veya normalleştiren yazılı dilin yalnızca sınırlı bir alt kümesini çıkarır. Tamamen büyük harfli veya tamamen küçük harfli bir transkriptin insan tarafından üretilmiş olması çok düşük olasılıktır. Birçok ASR sistemi bir düzeyde ters metin normalleştirme içerse de, bu genellikle basit veya kural tabanlıdır ve hiçbir zaman virgül içermeme gibi diğer ele alınmamış yönlerden hala tespit edilebilir.

Ayrıca, konuşulan dilin CLD2'ye göre transkriptin diliyle eşleştiğinden emin olmak için veri setinin prototip sürümünde eğitilmiş bir prototip modeli VoxLingua107 (Valk & Aluma¨e, 2021) üzerinde ince ayar yaparak oluşturulan bir ses dil dedektörü kullanırız. İkisi eşleşmezse, (ses, transkript) çiftini veri setinde konuşma tanıma eğitim örneği olarak dahil etmeyiz. Transkript dili İngilizce ise bir

istisna yapar ve bu çiftleri bunun yerine X→en konuşma çevirisi eğitim örnekleri olarak veri setine ekleriz. Eğitim veri setindeki çoğaltma ve otomatik olarak üretilmiş içerik miktarını azaltmak için transkript metinlerinin bulanık çoğaltma kaldırmasını kullanırız.

Ses dosyalarını o zaman dilimi içinde oluşan transkriptin alt kümesiyle eşleştirilmiş 30 saniyelik segmentlere böleriz. Konuşmanın olmadığı segmentler de dahil olmak üzere tüm ses üzerinde eğitim yaparız (ancak alt örneklenmiş olasılıkla) ve bu segmentleri ses etkinliği tespiti için eğitim verisi olarak kullanırız.

Ek bir filtreleme geçişi için, ilk bir model eğittikten sonra eğitim veri kaynakları üzerindeki hata oranı hakkında bilgi topladık ve düşük kaliteli olanları verimli bir şekilde tanımlamak ve kaldırmak için hem yüksek hata oranı hem de veri kaynağı boyutunun bir kombinasyonuna göre sıralayarak bu veri kaynaklarının manuel incelemesini gerçekleştirdik. Bu inceleme, yalnızca kısmen transkript edilmiş veya kötü hizalanmış/yanlış hizalanmış transkriptlerin yanı sıra filtreleme buluşsal yöntemlerinin tespit etmediği kalan düşük kaliteli makine tarafından üretilmiş altyazıların büyük miktarını gösterdi.

Kontaminasyonu önlemek için, eğitim veri seti ile örtüşme riski daha yüksek olduğunu düşündüğümüz değerlendirme veri setleri, yani TED-LIUM 3 (Hernandez ve diğerleri, 2018) arasında transkript düzeyinde çoğaltma kaldırma gerçekleştiririz.

2.2. Model

Çalışmamızın odağı konuşma tanıma için büyük ölçekli denetimli ön eğitimin yeteneklerini incelemek olduğundan, bulgularımızı model iyileştirmeleriyle karıştırmaktan kaçınmak için hazır bir mimari kullanırız. Bu mimari güvenilir bir şekilde ölçeklenmek için iyi doğrulanmış olduğu için bir kodlayıcı-kod çözücü Transformer (Vaswani ve diğerleri, 2017) seçtik. Tüm ses 16.000 Hz'e yeniden örneklenir ve 10 milisaniyelik adımlarla 25 milisaniyelik pencereler üzerinde 80 kanallı log-büyüklük Mel spektrogram temsili hesaplanır. Özellik normalleştirme için, girdiyi ön eğitim veri seti boyunca yaklaşık sıfır ortalama ile -1 ve 1 arasında olacak şekilde global olarak ölçeklendiririz. Kodlayıcı bu girdi temsilini, 3 filtre genişliğine sahip iki evrişim katmanından ve GELU aktivasyon fonksiyonundan (Hendrycks & Gimpel, 2016) oluşan küçük bir gövde ile işler; burada ikinci evrişim katmanı iki adımlıdır. Sinüzoidal konum gömmeleri daha sonra gövdenin çıktısına eklenir ve ardından kodlayıcı Transformer blokları uygulanır. Transformer ön aktivasyon artık blokları (Child ve diğerleri, 2019) kullanır ve kodlayıcı çıktısına son bir katman normalleştirmesi uygulanır. Kod çözücü öğrenilmiş konum gömmeleri ve bağlı girdi-çıktı token temsillerini (Press & Wolf, 2017) kullanır. Kodlayıcı ve kod çözücü aynı genişliğe ve transformer blok sayısına sahiptir. Şekil 1 model mimarisini özetler.

Yalnızca İ̇ngilizce modeller için GPT-2'de (Sennrich ve diğerleri, 2015; Radford ve diğerleri, 2019) kullanılan aynı bayt düzeyinde BPE metin tokenizer'ını kullanırız ve GPT-2 BPE kelime dağarcığı yalnızca İ̇ngilizce olduğu için diğer dillerde aşırı parçalanmayı önlemek için çok dilli modeller için kelime dağarcığını yeniden uyarlarız (ancak aynı boyutu koruruz).

2.3. Çok Görevli Format

Belirli bir ses parçacığında hangi kelimelerin söylendiğini tahmin etmek, tam konuşma tanıma probleminin temel bir parçası ve araştırmada kapsamlı olarak incelenmiş olsa da, tek parça değildir. Tam özellikli bir konuşma tanıma sistemi, ses etkinliği tespiti, konuşmacı diyarizasyonu ve ters metin normalleştirmesi gibi birçok ek bileşen içerebilir. Bu bileşenler genellikle ayrı ayrı ele alınır, bu da temel konuşma tanıma modelinin etrafında nispeten karmaşık bir sistem ile sonuçlanır. Bu karmaşıklığı azaltmak için, yalnızca temel tanıma kısmını değil, tüm konuşma işleme boru hattını gerçekleştiren tek bir modele sahip olmak istiyoruz. Burada önemli bir husus model için arayüzdür. Aynı girdi ses sinyali üzerinde gerçekleştirilebilecek birçok farklı görev vardır: transkripsiyon, çeviri, ses etkinliği tespiti, hizalama ve dil tanımlama bunlardan bazı örneklerdir.

Bu tür bire-çok eşlemenin tek bir modelle çalışması için, bir tür görev belirtimi gereklidir. Tüm görevleri ve koşullandırma bilgilerini kod çözücüye girdi token'ları dizisi olarak belirtmek için basit bir format kullanırız. Kod çözücümüz ses koşullu bir dil modeli olduğu için, belirsiz sesi çözmek için daha uzun menzilli metin bağlamını kullanmayı öğreneceği umuduyla onu transkriptin metin geçmişi üzerinde koşullandırmak için de eğitiriz. Özellikle, bir olasılıkla mevcut ses segmentinden önceki transkript metnini kod çözücünün bağlamına ekleriz. Tahmin başlangıcını bir <|startoftranscript|> token'ı ile belirtiriz.

İlk olarak, eğitim setimizde her dil için benzersiz bir token ile temsil edilen konuşulan dili tahmin ederiz (toplam 99). Bu dil hedefleri yukarıda bahsedilen VoxLingua107 modelinden alınır. Bir ses segmentinde konuşma olmadığı durumda, model bunu belirten bir <|nospeech|> token'ı tahmin etmek için eğitilir. Sonraki token, <|transcribe|> veya <|translate|> token'ı ile görevi (transkripsiyon veya çeviri) belirtir. Bundan sonra, bu durum için bir <|notimestamps|> token'ı dahil ederek zaman damgalarını tahmin edip etmeyeceğimizi belirtiriz. Bu noktada, görev ve istenen format tamamen belirtilmiştir ve çıktı başlar. Zaman damgası tahmini için, mevcut ses segmentine göre zamanı tahmin ederiz, tüm zamanları Whisper modellerinin yerel zaman çözünürlüğüyle eşleşen en yakın 20 milisaniyeye nicelendiririz ve bunların her biri için kelime dağarcığımıza ek token'lar ekleriz. Onların tahminini altyazı token'larıyla serpiştirir: başlangıç zamanı token'ı her altyazının metninden önce tahmin edilir ve bitiş zamanı token'ı sonrasında tahmin edilir. Son bir transkript segmenti mevcut 30 saniyelik ses parçasına yalnızca kısmen dahil edildiğinde, zaman damgası modunda segment için yalnızca başlangıç zamanı token'ını tahmin ederiz, sonraki kod çözmenin o zamanla hizalanmış bir ses penceresi üzerinde gerçekleştirilmesi gerektiğini belirtmek için, aksi takdirde sesi segmenti içermeyecek şekilde keseriz. Son olarak, bir <|endoftranscript|> token'ı ekleriz. Eğitim kaybını yalnızca önceki bağlam metni üzerinde maskeliyoruz ve modeli diğer tüm token'ları tahmin etmek için eğitiyoruz. Formatımızın ve eğitim kurulumumuzun genel bakışı için lütfen Şekil 1'e bakın.

2.4. Eğitim Detayları

Genelleme ve sağlamlığı teşvik etmek için çeşitli boyutlarda modeller eğitiriz. Tam eğitim hiperparametreleri için lütfen Ek F'ye bakın.3

Erken geliştirme ve değerlendirme sırasında, Whisper modellerinin konuşmacıların adları için makul ancak neredeyse her zaman yanlış tahminler yapma eğiliminde olduğunu gözlemledik. Bu, ön eğitim veri setindeki birçok transkriptin konuşan kişinin adını içermesi nedeniyle olur, bu da modeli onları tahmin etmeye teşvik eder, ancak bu bilgi en son 30 saniyelik ses segmentinden nadiren çıkarılabilir.

Modeller, dinamik kayıp ölçeklendirme ve aktivasyon kontrol noktası (Griewank & Walther, 2000; Chen ve diğerleri, 2016) ile FP16 kullanarak hızlandırıcılarda veri paralelliği ile eğitildi. Modeller AdamW (Loshchilov & Hutter, 2017) ve gradyan norm kırpma (Pascanu ve diğerleri, 2013) ile ilk 2048 güncellemeden sonra sıfıra doğrusal bir öğrenme oranı düşüşü ile eğitildi. 256 segmentlik bir parti boyutu kullanıldı ve modeller veri seti üzerinde iki ila üç geçiş arasında olan 220 güncelleme için eğitildi. Sadece birkaç dönem için eğitim yapıldığı için, aşırı uyum büyük bir endişe değildir ve herhangi bir veri artırma veya düzenleme kullanmıyoruz ve bunun yerine böyle büyük bir veri setinde bulunan çeşitliliğe güveniyoruz.

3. Deneyler

3.1. Sıfır Atış Değerlendirme

Whisper'ın amacı, belirli dağılımlarda yüksek kaliteli sonuçlar elde etmek için veri setine özgü ince ayar gerektirmeden güvenilir bir şekilde çalışan tek, sağlam bir konuşma işleme sistemi geliştirmektir. Bu yeteneği incelemek için, Whisper'ın alanlar, görevler ve diller arasında iyi genelleme yapıp yapmadığını kontrol etmek için geniş bir mevcut konuşma işleme veri seti kullanırız. Bu veri setleri için hem eğitim hem de test bölümünü içeren standart değerlendirme protokolünü kullanmak yerine, geniş genellemeyi ölçmek için bu veri setlerinin hiçbir eğitim verisini kullanmadan Whisper'ı sıfır atış ayarında değerlendiririz.

3.2. Değerlendirme Metrikleri

Konuşma tanıma araştırması tipik olarak sistemleri kelime hata oranı (WER) metriğine göre değerlendirir ve karşılaştırır. Ancak, dize düzenleme mesafesine dayanan WER, modelin çıktısı ile referans transkript arasındaki tüm farklılıkları, transkript stilindeki zararsız farklılıklar da dahil olmak üzere cezalandırır. Sonuç olarak, insanlar tarafından doğru olarak değerlendirilecek transkriptler üreten sistemler, küçük biçimlendirme farklılıkları nedeniyle hala büyük bir WER'ye sahip olabilir. Bu, tüm transkriptörler için bir sorun teşkil ederken, belirli veri setlerinin transkript formatlarının hiçbir örneğini gözlemlemeyen Whisper gibi sıfır atış modelleri için özellikle akut bir durumdur.

Bu yeni bir gözlem değildir; insan yargısıyla daha iyi korelasyon gösteren değerlendirme metriklerinin geliştirilmesi aktif bir araştırma alanıdır ve bazı umut vadeden yöntemler olsa da, hiçbiri konuşma tanıma için yaygın olarak benimsenmemiştir. Bu sorunu, WER hesaplamasından önce metnin kapsamlı bir şekilde standartlaştırılmasıyla, anlamsal olmayan farklılıkların cezalandırılmasını en aza indirmek için ele almayı tercih ederiz. Metin normalleştiricimiz, saf WER'nin Whisper modellerini zararsız bir farklılık için cezalandırdığı yaygın kalıpları belirlemek için

yinelemeli manuel inceleme yoluyla geliştirilmiştir. Ek C tam ayrıntıları içerir. Birkaç veri seti için, genellikle bir veri setinin referans transkriptlerinin kısaltmaları kelimelerden boşlukla ayırması gibi bir tuhaflık nedeniyle WER'de %50'ye varan düşüşler gözlemleriz. Bu geliştirme prosedürünün, Bölüm 4.4'te incelediğimiz Whisper modellerinin transkripsiyon stiline aşırı uyum riski taşıdığı konusunda uyarıyoruz. Kolay karşılaştırma sağlamak ve diğerlerinin dağıtım dışı ayarlarda konuşma tanıma sistemlerinin performansını incelemesine yardımcı olmak için metin normalleştiricimizin kodunu yayınlıyoruz.

3.3. İngilizce Konuşma Tanıma

2015 yılında, Deep Speech 2 (Amodei ve diğerleri, 2015), LibriSpeech test-clean bölümünü transkript ederken insan düzeyinde performansı eşleştiren bir konuşma tanıma sistemi bildirdi. Analizlerinin bir parçası olarak şu sonuca vardılar: "Bu sonuç göz önüne alındığında, genel bir konuşma sisteminin daha fazla alan adaptasyonu olmadan temiz okunan konuşmada daha fazla iyileşme için çok az yer olduğunu düşünüyoruz." Ancak yedi yıl sonra, LibriSpeech test-clean üzerindeki SOTA WER, %5.3'ten %1.4'e (%73 düşüş) düşmüştür (Zhang ve diğerleri, 2021), bu da bildirilen insan düzeyindeki hata oranı olan %5.8'in çok altındadır. Tutulan ancak dağıtım içi verilerdeki performansta bu büyük ve beklenmedik iyileşmeye rağmen, LibriSpeech üzerinde eğitilmiş konuşma tanıma modelleri, diğer ayarlarda kullanıldığında insan hata oranlarının çok üzerinde kalmaktadır. Dağıtım içi bildirilen insanüstü performans ile dağıtım dışı insan altı performans arasındaki bu boşluğu ne açıklar?

İnsan ve makine davranışları arasındaki bu boşluğun büyük bir kısmının, bir test setinde insan ve makine performansı tarafından ölçülen farklı yeteneklerin karıştırılmasından kaynaklandığından şüpheleniyoruz. Bu iddia ilk başta kafa karıştırıcı görünebilir; hem insanlar hem de makineler aynı testi yapıyorsa, farklı becerilerin nasıl test edildiği nasıl olabilir? Fark, testte değil, bunun için nasıl eğitildiklerinde ortaya çıkar. İnsanlardan genellikle incelenen belirli veri dağılımı üzerinde çok az veya hiç denetim olmadan bir görev yapmaları istenir. Dolayısıyla insan performansı, dağıtım dışı genellemenin bir ölçüsüdür. Ancak makine öğrenimi modelleri genellikle değerlendirme dağılımından büyük miktarda denetimle eğitildikten sonra değerlendirilir, bu da makine performansının bunun yerine dağıtım içi genellemenin bir ölçüsü olduğu anlamına gelir. Hem insanlar hem de makineler aynı test verileri üzerinde değerlendirilse de, eğitim verilerindeki bir farklılık nedeniyle iki oldukça farklı yetenek ölçülmektedir.

Whisper modelleri, geniş ve çeşitli bir ses dağılımı üzerinde eğitilmiş ve sıfır atış ayarında değerlendirilmiş olup, mevcut sistemlerden çok daha iyi insan davranışını eşleştirebilir. Bunun böyle olup olmadığını (veya makine ve insan performansı arasındaki farkın henüz anlaşılamayan faktörlerden kaynaklanıp kaynaklanmadığını) incelemek için Whisper modellerini hem insan performansıyla hem de standart ince ayarlı makine öğrenimi modelleriyle karşılaştırabilir ve hangisine daha çok benzediklerini kontrol edebiliriz.

3.4. Çok Dilli Konuşma Tanıma

Çok dilli konuşma tanıma üzerine önceki çalışmalarla karşılaştırmak için, Tablo 3'te iki düşük veri kıyaslama testi olan Çok Dilli LibriSpeech (MLS) (Pratap ve diğerleri, 2020b) ve VoxPopuli (Wang ve diğerleri, 2021) üzerindeki sonuçları rapor ediyoruz.

Whisper, Çok Dilli LibriSpeech üzerinde iyi performans gösterir, XLS-R (Babu ve diğerleri, 2021), mSLAM (Bapna ve diğerleri, 2021) gibi modelleri geride bırakır.

Tablo 1. Whisper model ailesinin mimari detayları.

Model	Katmanlar	Genişlik	Başlıklar	Parametreler
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

Tablo 2. Çeşitli veri setlerinde etkin sağlamlığın detaylı karşılaştırması. Her iki model de LibriSpeech üzerinde birbirine 0.1% içinde performans gösterse de, sıfır atış Whisper modeli, LibriSpeech performansı için beklenenden çok daha iyi performans gösterir ve ortalama olarak %55.2 daha az hata yapar. Sonuçlar, metin normalleştiricimiz uygulandıktan sonra her iki model için kelime hata oranı (WER) olarak rapor edilmiştir.

Şekil 2. Sıfır atış Whisper modelleri insan sağlamlığına olan boşluğu kapatır. LibriSpeech dev-clean üzerinde bir insanı eşleştirmesine veya geride bırakmasına rağmen, denetimli LibriSpeech modelleri diğer veri setlerinde bir insandan yaklaşık iki kat daha fazla hata yapar, bu da kırılganlıklarını ve sağlamlık eksikliklerini gösterir. Ancak, sıfır atış Whisper modellerinin tahmini sağlamlık sınırı, bu belirli insan için %95 güven aralığını içerir.

Bu farkı nicelendirmek için, hem genel sağlamlığı, yani birçok dağıtım/veri seti üzerindeki ortalama performansı, hem de Taori ve diğerleri (2020) tarafından tanıtılan etkin sağlamlığı inceleriz; bu, genellikle dağıtım içi olan bir referans veri seti ile bir veya daha fazla dağıtım dışı veri seti arasındaki beklenen performans farkını ölçer. Yüksek etkin sağlamlığa sahip bir model, referans veri setindeki performansının bir fonksiyonu olarak dağıtım dışı veri setlerinde beklenenden daha iyi performans gösterir ve tüm veri setlerinde eşit performans idealine yaklaşır. Analizimiz için, modern konuşma tanıma araştırmalarındaki merkezi rolü ve üzerinde eğitilmiş birçok yayınlanmış modelin mevcudiyeti nedeniyle LibriSpeech'i referans veri seti olarak kullanırız, bu da sağlamlık davranışlarını karakterize etmeye olanak tanır. Dağıtım dışı davranışları incelemek için 12 diğer akademik konuşma tanıma veri seti kullanırız. Bu veri setleri hakkında tam ayrıntılar Ek A'da bulunabilir.

Ana bulgularımız Şekil 2 ve Tablo 2'de özetlenmiştir. En iyi sıfır atış Whisper modelinin nispeten sıradan bir LibriSpeech clean-test WER'si 2.5 olmasına rağmen, bu modern denetimli temel veya 2019 ortası sanatın durumu performansına yakındır, sıfır atış Whisper modelleri denetimli LibriSpeech modellerinden çok farklı sağlamlık özelliklerine sahiptir ve diğer veri setlerinde tüm kıyaslanan LibriSpeech modellerini büyük ölçüde geride bırakır. Yalnızca 39 milyon parametreye sahip ve LibriSpeech test-clean üzerinde 6.7 WER'ye sahip en küçük sıfır atış Whisper modeli, diğer veri setlerinde değerlendirildiğinde en iyi denetimli LibriSpeech modeliyle yaklaşık olarak rekabet edebilir. Şekil 2'deki bir insanla karşılaştırıldığında, en iyi sıfır atış Whisper modelleri doğruluklarını ve sağlamlıklarını yaklaşık olarak eşleştirir. Sağlamlıktaki bu büyük iyileşmenin detaylı bir dökümü için, Tablo 2, en iyi sıfır atış Whisper modelinin performansını, LibriSpeech test-clean üzerinde ona en yakın performansa sahip denetimli bir LibriSpeech modeliyle karşılaştırır. Referans dağılımında çok yakın performanslarına rağmen, sıfır atış Whisper modeli, diğer konuşma tanıma veri setlerinde değerlendirildiğinde ortalama %55.2'lik bir göreceli hata azaltma elde eder.

Bu bulgu, özellikle insan performansıyla karşılaştırmaya çalışırken, yanıltıcı karşılaştırmalar nedeniyle makine öğrenimi sistemlerinin yeteneklerini abartmaktan kaçınmak için modellerin sıfır atış ve dağıtım dışı değerlendirmelerine vurgu yapmayı önerir.

Tablo 3. Çok dilli konuşma tanıma performansı. Sıfır atış Whisper, Çok Dilli LibriSpeech (MLS) üzerindeki performansı iyileştirir, ancak VoxPopuli üzerinde hem Maestro, XLS-R hem de mSLAM'in hala önemli ölçüde gerisindedir.

Model	MLS	VoxPopuli
VP-10K + FT	-	15.3
XLS-R (1B)	10.9	10.6
mSLAM-CTC (2B)	9.7	9.1
Maestro	-	8.1
Zero-Shot Whisper	7.3	13.6

Bu iki kıyaslama testi biraz dardır çünkü yalnızca 15 benzersiz dil içerirler ve bunların neredeyse tamamı Hint-Avrupa dil ailesindendir ve birçoğu yüksek kaynaklı dillerdir. Bu kıyaslama testleri, konuşma tanıma için 75 dilde eğitim verisi içeren Whisper modellerinin çok dilli yeteneklerini incelemek için sınırlı kapsama ve alan sağlar. Whisper'ın performansını daha geniş bir şekilde incelemek için Fleurs veri setindeki (Conneau ve diğerleri, 2022) performansı da rapor ediyoruz. Özellikle, belirli bir dil için sahip olduğumuz eğitim verisi miktarı ile o dil için ortaya çıkan sıfır atış performansı arasındaki ilişkiyi incelemekle ilgileniyorduk.

Şekil 3. Ön eğitim denetim miktarının aşağı akış konuşma tanıma performansıyla korelasyonu. Belirli bir dil için ön eğitim konuşma tanıma verisi miktarı, Fleurs'taki o dilde sıfır atış performansını çok iyi tahmin eder.

Bu ilişkiyi Şekil 3'te görselleştiriyoruz. Kelime hata oranının logaritması ile dil başına eğitim verisi miktarının logaritması arasında 0.83'lük güçlü bir kare korelasyon katsayısı buluyoruz. Bu log-log değerlerine doğrusal bir uyum için regresyon katsayısını kontrol etmek, eğitim verisindeki her 16 kat artış için WER'nin yarıya indiği bir tahminle sonuçlanır. Ayrıca, bu eğilime göre beklenenden daha kötü performans açısından en büyük aykırı değerlerin çoğunun, İbranice (HE), Telugu (TE), Çince (ZH) ve Korece (KO) gibi eğitim veri setinin çoğunluğunu oluşturan Hint-Avrupa dillerinden daha uzak ve benzersiz yazım sistemlerine sahip diller olduğunu gözlemledik. Bu farklılıklar, dilsel uzaklık nedeniyle transfer eksikliğinden, bayt düzeyindeki BPE tokenizer'ımızın bu dillere kötü bir eşleşme olmasından veya veri kalitesindeki varyasyonlardan kaynaklanabilir.

Şekil 4. Ön eğitim denetim miktarının aşağı akış çeviri performansıyla korelasyonu. Belirli bir dil için ön eğitim çeviri verisi miktarı, Fleurs'taki o dilde Whisper'ın sıfır atış performansını yalnızca orta derecede tahmin eder.

3.5. Çeviri

Whisper modellerinin çeviri yeteneklerini, CoVoST2 (Wang ve diğerleri, 2020b) X→en alt kümesindeki performanslarını ölçerek inceliyoruz. Maestro, mSLAM ve XLS-R gibi en yüksek performans gösteren önceki çalışmalarla karşılaştırıyoruz. CoVoST2 eğitim verilerinin hiçbirini kullanmadan 29.1 BLEU sıfır atış ile yeni bir sanatın durumunu elde ediyoruz. Bunu, ön eğitim veri setimizdeki bu diller için 68.000 saatlik X→en çeviri verisine bağlıyoruz, bu veri gürültülü olsa da, CoVoST2'deki X→en çeviri için 861 saatlik eğitim verisinden çok daha büyüktür. Whisper değerlendirmesi sıfır atış olduğu için, CoVoST2'nin en düşük kaynak gruplamasında özellikle iyi performans gösterir ve mSLAM'ı 6.7 BLEU ile geride bırakır. Tersine, en iyi Whisper modeli, en yüksek kaynaklı diller için ortalama olarak Maestro ve mSLAM'ı aslında iyileştirmez.

Daha da geniş bir dil kümesi üzerinde ek bir analiz için, bir konuşma tanıma veri seti olan Fleurs'u da çeviri veri seti olarak yeniden kullanıyoruz. Aynı cümleler her dil için transkript edildiğinden, İngilizce transkriptleri referans çeviriler olarak kullanıyoruz. Şekil 4'te, dil başına çeviri eğitim verisi miktarı ile Fleurs'taki ortaya çıkan sıfır atış BLEU puanı arasındaki korelasyonu görselleştiriyoruz. Artan eğitim verisi ile açık bir iyileşme eğilimi olsa da, kare korelasyon katsayısı konuşma tanıma için gözlemlenen 0.83'ten çok daha düşüktür ve sadece 0.24'tür. Bunun kısmen ses dili tanımlamasındaki hatalardan kaynaklanan daha gürültülü eğitim verilerinden kaynaklandığından şüpheleniyoruz. Örneğin, Galce (CY), 9.000 saatlik çeviri verisine sahip olmasına rağmen sadece 13 BLEU ile beklenenden çok daha kötü performans gösteren bir aykırı değerdir. Bu büyük miktardaki Galce çeviri verisi şaşırtıcıdır, genel çeviri verileri arasında 4. sırada yer alır ve Fransızca, İspanyolca ve Rusça gibi dünyanın en çok konuşulan dillerinden bazılarını geride bırakır. İnceleme, sözde Galce çeviri verilerinin çoğunun aslında İngilizce sesli ve İngilizce altyazılı olduğunu, İngilizce sesin dil tanımlama sistemi tarafından yanlışlıkla Galce olarak sınıflandırıldığını ve bu nedenle veri seti oluşturma kurallarımıza göre transkripsiyon verisi yerine çeviri eğitim verisi olarak dahil edildiğini göstermektedir.

Tablo 4. X→en Konuşma çevirisi performansı. Sıfır atış Whisper, genel, orta ve düşük kaynak ayarlarında CoVoST2'deki mevcut modelleri geride bırakır, ancak önceki doğrudan denetimli

çalışmalara kıyasla yüksek kaynaklı dillerde hala orta derecede düşük performans gösterir.

Model	Yüksek	Orta	Düşük	Tümü
XMEF-X	34.2	20.2	5.9	14.7
XLS-R (2B)	36.1	27.7	15.1	22.1
mSLAM-CTC (2B)	37.8	29.6	18.5	24.8
Maestro	38.2	31.3	18.4	25.2
Zero-Shot Whisper	36.2	32.6	25.2	29.1

Şekil 5. Katkılı beyaz gürültü (sol) ve pub gürültüsü (sağ) altında SNR'nin bir fonksiyonu olarak LibriSpeech test-clean üzerindeki WER. LibriSpeech eğitimli modellerin doğruluğu, en iyi Whisper modelinden ((cid:70)) daha hızlı bozulur. NVIDIA STT modelleri (•) düşük gürültü altında en iyi performansı gösterir, ancak yüksek gürültü altında (SNR < 10 dB) Whisper tarafından geride bırakılır. Düşük gürültü altında ikinci en iyi model ((cid:72)) yalnızca LibriSpeech üzerinde ince ayar yapılmıştır ve daha da hızlı bozulur.

3.6. Dil Tanımlama

Dil tanımlamayı değerlendirmek için Fleurs veri setini (Conneau ve diğerleri, 2022) kullanıyoruz. Whisper'ın sıfır atış performansı burada önceki denetimli çalışmalarla rekabetçi değildir ve denetimli SOTA'nın %13.6 gerisinde kalır. Ancak, Whisper, Fleurs'taki 102 dilden 20'si için eğitim verisi içermediği için dil tanımlamada büyük ölçüde dezavantajlıdır, bu da doğruluğu %80.4 ile sınırlar. Çakışan 82 dilde en iyi Whisper modeli %80.3 doğruluk elde eder.

Tablo 5. Dil tanımlama performansı. Sıfır atış Whisper'ın dil tanımlamadaki doğruluğu, Fleurs'taki önceki denetimli sonuçlarla rekabetçi değildir. Bu kısmen, Whisper'ın Fleurs dillerinin 20'si için eğitim verisi olmaması nedeniyle ağır bir şekilde cezalandırılmasından kaynaklanmaktadır.

Model	Fleurs
w2v-bert-51 (0.6B)	71.4
mSLAM-CTC (2B)	77.7
Zero-shot Whisper	64.5

3.7. Katkılı Gürültüye Karşı Sağlamlık

Whisper modellerinin ve 14 LibriSpeech eğitimli modelin gürültü sağlamlığını, sese Audio Degradation Toolbox (Mauch & Ewert, 2013) tarafından beyaz gürültü veya pub gürültüsü eklendiğinde WER ölçerek test ettik. Pub gürültüsü, kalabalık bir restoranda veya pubda tipik olan

ortam gürültüsü ve belirsiz sohbet ile daha doğal gürültülü bir ortamı temsil eder. 14 model arasında, on iki tanesi LibriSpeech üzerinde önceden eğitilmiş ve/veya ince ayar yapılmıştı ve diğer ikisi, LibriSpeech içeren SpeechStew gibi önceki çalışmalara benzer bir karışım veri seti üzerinde eğitilmiş NVIDIA STT modelleridir. Belirli bir sinyal-gürültü oranına (SNR) karşılık gelen katkılı gürültü seviyesi, bireysel örneklerin sinyal gücüne göre hesaplanır. Şekil 5, katkılı gürültü yoğunlaştıkça ASR performansının nasıl bozulduğunu göstermektedir. Düşük gürültü altında (40 dB SNR) sıfır atış performansımızı geride bırakan birçok model vardır, bu modellerin öncelikle LibriSpeech üzerinde eğitildiği göz önüne alındığında şaşırtıcı değildir, ancak tüm modeller gürültü yoğunlaştıkça hızla bozulur ve 10 dB altındaki SNR ile katkılı pub gürültüsü altında Whisper modelinden daha kötü performans gösterir. Bu, Whisper'ın gürültüye karşı sağlamlığını, özellikle pub gürültüsü gibi daha doğal dağıtım kaymaları altında sergiler.

3.8. Uzun Form Transkripsiyon

Whisper modelleri 30 saniyelik ses parçaları üzerinde eğitilmiştir ve daha uzun ses girişlerini bir kerede tüketemez. Bu, çoğu akademik veri setinde kısa ifadelerden oluşan bir sorun değildir, ancak genellikle dakikalar veya saatler süren sesleri transkript etmeyi gerektiren gerçek dünya uygulamalarında zorluklar sunar. Uzun seslerin arabelleğe alınmış transkripsiyonunu gerçekleştirmek için bir strateji geliştirdik; bu, sesin 30 saniyelik segmentlerini ardışık olarak transkript ederek ve pencereyi model tarafından tahmin edilen zaman damgalarına göre kaydırarak yapılır. Uzun sesleri güvenilir bir şekilde transkript etmek için ışın aramasının ve model tahminlerinin tekrarlanabilirliğine ve log olasılığına dayalı sıcaklık zamanlamasının kritik olduğunu gözlemledik. Tam prosedür Bölüm 4.5'te açıklanmıştır.

Uzun form transkripsiyon performansını, çeşitli uzunluklarda ve kayıt koşullarında konuşma kayıtlarından oluşan yedi veri seti üzerinde değerlendiriyoruz, mümkün olduğunca çeşitli bir veri dağılımını kapsamak için. Bunlar arasında, her örneğin tam uzunlukta bir TED konuşması olduğu TED-LIUM3'ün (Hernandez ve diğerleri, 2018) birleştirilmiş uzun form adaptasyonu, Stephen Colbert ile The Late Show'dan (Meanwhile) alınan jargon yüklü segmentler koleksiyonu, çevrimiçi bloglarda ASR kıyaslama testleri olarak kullanılan video/podcast setleri (Rev16 ve Kincaid46), kazanç çağrılarının kayıtları (Del Rio ve diğerleri, 2021) ve Bölgesel Afro-Amerikan Dili Külliyatı'ndan (CORAAL) (Gunter ve diğerleri, 2021) tam uzunluktaki röportajlar bulunmaktadır. Uzun form veri setleri hakkında tam ayrıntılar Ek A'da bulunabilir.

Performansı açık kaynak modellerle ve 4 ticari ASR hizmetiyle karşılaştırıyoruz. Sonuçlar Şekil 6'da özetlenmiştir, Whisper ve 4 ticari ASR hizmetinden kelime hata oranlarının dağılımını göstermektedir, burada giriş uzunlukları birkaç dakikadan birkaç saate kadar değişmektedir. Kutular, örnek başına WER'lerin çeyreklerini göstermektedir ve veri seti başına toplam WER'ler her kutuda belirtilmiştir. Modelimiz, tüm veri setlerinde en iyi açık kaynak modelini (NVIDIA STT) geride bırakır ve çoğu durumda ticari ASR sistemlerini de geride bırakır.

Şekil 6. Whisper, uzun form transkripsiyonda son teknoloji ticari ve açık kaynak ASR sistemleriyle rekabetçidir. Yedi uzun form veri setinde altı ASR sisteminden kelime hata oranlarının dağılımı karşılaştırılmıştır, burada giriş uzunlukları birkaç dakikadan birkaç saate kadar

değişmektedir. Kutular, örnek başına WER'lerin çeyreklerini göstermektedir ve veri seti başına toplam WER'ler her kutuda belirtilmiştir. Modelimiz, tüm veri setlerinde en iyi açık kaynak modelini (NVIDIA STT) geride bırakır ve çoğu durumda ticari ASR sistemlerini de geride bırakır.

ve ayrıca NeMo araç setinden (Kuchaiev ve diğerleri, 2019) NVIDIA STT Conformer-CTC Large modeli, açık kaynak modeller arasında en iyi performansı gösterdi. Tüm ticari ASR hizmetleri, 1 Eylül 2022 itibarıyla varsayılan İngilizce transkripsiyon ayarları kullanılarak sorgulanır ve NVIDIA STT modeli için uzun form transkripsiyonu etkinleştirmek üzere FrameBatchASR sınıfındaki arabelleğe alınmış çıkarım uygulamalarını kullandık. Sonuçlar, Whisper'ın karşılaştırılan modellerden çoğu veri setinde, özellikle de nadir kelimelerle dolu Meanwhile veri setinde daha iyi performans gösterdiğini göstermektedir. Ayrıca, bazı ticari ASR sistemlerinin bu halka açık veri setlerinden bazıları üzerinde eğitilmiş olabileceği olasılığını da belirtmek gerekir ve bu nedenle bu sonuçlar sistemlerin göreceli sağlamlığını doğru bir şekilde yansıtmayabilir.

3.9. İnsan Performansıyla Karşılaştırma

Belirsiz veya anlaşılmaz konuşma ile etiketleme hataları nedeniyle, her veri setinde farklı düzeylerde indirgenemez hata vardır ve yalnızca ASR sistemlerinden gelen WER metrikleriyle her veri setinde ne kadar iyileşme alanı olduğunu anlamak zordur. Whisper'ın performansının insan performansına ne kadar yakın olduğunu nicelendirmek için, Kincaid46 veri setinden 25 kayıt seçtik ve profesyonel transkriptörler tarafından üretilen transkriptleri elde etmek için 5 hizmet kullandık; bunlardan biri bilgisayar destekli transkripsiyon sağlarken, diğer dördü tamamen insan tarafından transkript edilmiştir. Ses seçimi, senaryolu ve senaryosuz yayın, telefon ve VoIP aramaları ve toplantılar gibi çeşitli kayıt koşullarını kapsar. Şekil 7, 25 kayıt boyunca örnek başına WER'lerin ve toplam WER'in dağılımını göstermektedir; burada bilgisayar destekli hizmet, Whisper'ınkinden 1.15% puan daha iyi olan en düşük toplam WER'ye sahiptir ve saf insan performansı, Whisper'ınkinden sadece yüzde bir puanın bir kısmı kadar daha iyidir. Bu sonuçlar, Whisper'ın İngilizce ASR performansının mükemmel olmadığını ancak insan düzeyindeki doğruluğa çok yakın olduğunu göstermektedir.

Şekil 7. Whisper'ın performansı profesyonel insan transkriptörlerininkine yakındır. Bu grafik, Kincaid46 veri setinden Whisper, Şekil 6'daki aynı 4 ticari ASR sistemi (A-D), bir bilgisayar destekli insan transkripsiyon hizmeti (E) ve 4 insan transkripsiyon hizmeti (F-I) tarafından transkript edilen 25 kaydın WER dağılımlarını göstermektedir. Kutu grafiği, bireysel kayıtlardaki WER'leri gösteren noktalarla üst üste bindirilmiştir ve 25 kayıt üzerindeki toplam WER'ler her kutuda belirtilmiştir.

4. Analiz ve Ablasyonlar

4.1. Model Ölçeklendirme

Zayıf denetimli eğitim yaklaşımlarının vaatlerinin büyük bir kısmı, geleneksel denetimli öğrenmedekilerden çok daha büyük veri setlerini kullanma potansiyelleridir. Ancak, bu, altın standart denetimden muhtemelen çok daha gürültülü ve düşük kaliteli verileri kullanma maliyetiyle

birlikte gelir. Bu yaklaşımla ilgili bir endişe, başlangıçta umut vadedici görünse de, bu tür veriler üzerinde eğitilmiş modellerin performansının, insan düzeyinin çok altında olabilecek veri setinin doğal kalite seviyesinde doygunluğa ulaşabilmesidir. Ilgili bir endişe, veri seti üzerinde eğitim için harcanan kapasite ve hesaplama arttıkça, modellerin veri setinin kendine özgü özelliklerini sömürmeyi öğrenebilmesi ve dağıtım dışı verilere sağlam bir şekilde genelleme yeteneklerinin bile bozulabilmesidir.

Bunun böyle olup olmadığını kontrol etmek için, Whisper modellerinin sıfır atış genellemesini model boyutunun bir fonksiyonu olarak inceliyoruz. Analizimiz Şekil 8'de özetlenmiştir. Ingilizce konuşma tanıma hariç, çok dilli konuşma tanıma, konuşma çevirisi ve dil tanımlama genelinde model boyutuyla performans artmaya devam etmektedir. Ingilizce konuşma tanıma için azalan getiriler, Bölüm 3.9'daki analizin önerdiği gibi insan düzeyindeki performansa yaklaşmaktan kaynaklanan doygunluk etkilerinden kaynaklanabilir.

4.2. Veri Seti Ölçeklendirme

680.000 saatlik etiketli ses ile Whisper veri seti, denetimli konuşma tanımada şimdiye kadar oluşturulmuş en büyük veri setlerinden biridir. Ham veri setinin boyutu Whisper'ın performansı için tam olarak ne kadar önemlidir? Bunu incelemek için, veri setinin %0.5, %1, %2, %4 ve %8'i olan alt örneklenmiş versiyonları üzerinde bir dizi orta boyutlu model eğittik ve performanslarını tüm veri seti üzerinde eğitilmiş aynı orta boyutlu modelle karşılaştırdık. Doğrulama kaybına dayalı erken durdurma, her veri seti boyutu için model kontrol noktalarını seçmek için kullanıldı. Değerlendirme, erken durdurma nedeniyle alt örneklenmiş veri setleri üzerinde eğitilmiş modeller için öğrenme oranının tamamen sıfıra düşmemesinin etkisini azaltmaya yardımcı olmak için 0.9999'luk bir düzeltme oranı kullanılarak parametrelerin üstel hareketli ortalama tahmini üzerinde gerçekleştirildi. Ingilizce ve çok dilli konuşma tanıma ve X→en çeviri performansı Tablo 6'da rapor edilmiştir.

Şekil 8. Sıfır atış Whisper performansı, artan model boyutuyla görevler ve diller arasında güvenilir bir şekilde ölçeklenir. Açık gölgeli çizgiler, bireysel veri setlerini veya dilleri temsil eder ve performansın, toplu performanstaki düzgün eğilimlerden daha çeşitli olduğunu gösterir. Large V2, bu analizdeki daha küçük modellerde bulunmayan birkaç değişiklik içerdiğinden kesikli turuncu bir çizgiyle ayırt edilmiştir.

Tablo 6. Performans, artan veri seti boyutuyla iyileşir. Ingilizce konuşma tanıma performansı 12 veri seti üzerindeki ortalamayı ifade ederken, Çok dilli konuşma tanıma, Fleurs'taki dillerin çakışan alt kümesindeki performansı rapor eder ve X→en çeviri, CoVoST2 üzerindeki ortalama BLEU'yu rapor eder. Veri seti boyutu saat olarak rapor edilmiştir.

Veri Seti Boyutu (saat)	İngilizce WER (↓)	Çok Dilli WER (↓)	X→En BLEU (↑)
3405	30.5	92.4	0.2
6811	19.6	72.7	1.7
13621	14.4	56.6	7.9
27243	12.3	45.0	13.9
54486	10.9	36.4	19.2
681070	9.9	29.2	24.8

Veri seti boyutundaki tüm artışlar, tüm görevlerde performansın iyileşmesine neden olur, ancak görevler ve boyutlar arasında iyileşme oranlarında önemli değişkenlik görüyoruz. İngilizce konuşma tanımada performans 3.000'den 13.000 saate hızla iyileşir ve ardından 13.000 ile 54.000 saat arasında belirgin şekilde yavaşlar. Boyutta 12.5 kat daha fazla artışa karşılık gelen tam veri setini kullanmak, WER'de sadece 1 puanlık bir düşüşle sonuçlanır. Bu, İngilizce konuşma tanıma için model boyutu ölçeklendirmesiyle gözlemlenen azalan getirileri yansıtır ve benzer şekilde insan düzeyindeki performansa yaklaşmaktan kaynaklanan doygunluk etkileriyle açıklanabilir.

Çok dilli konuşma tanıma için WER'deki iyileşmeler 54.000 saate kadar bir güç yasası eğilimini takip eder ve ardından bu eğilimden sapar, tam veri seti boyutuna artırıldığında sadece 7 puan daha iyileşir. X→en çeviri için, 7.000 saat veya daha az ses üzerinde eğitim yapıldığında performans neredeyse sıfırdır ve ardından 54.000 saate kadar kabaca log-doğrusal bir iyileşme eğilimini takip eder ve ardından azalan getiriler göstermeye başlar.

4.3. Çok Görevli ve Çok Dilli Transfer

Birçok görev ve dilde tek bir modeli birlikte eğitmenin potansiyel bir endişesi, birkaç görevin öğrenilmesi arasındaki etkileşimin, yalnızca tek bir görev veya dilde eğitimle elde edilecek performanstan daha kötü bir performansla sonuçlandığı negatif transfer olasılığıdır. Bunun olup olmadığını araştırmak için, yalnızca İngilizce konuşma tanıma üzerinde eğitilmiş modellerin performansını standart çok görevli ve çok dilli eğitim kurulumumuzla karşılaştırdık ve sıfır atış İngilizce konuşma tanıma kıyaslama testlerimizdeki ortalama performanslarını ölçtük. Ortak bir eğitim kurulumunda hesaplamanın sadece %65'i bu göreve harcandığı için, İngilizce konuşma tanıma görevinde harcanan FLOPs miktarını ayarlıyoruz; aksi takdirde analiz, aynı boyuttaki yalnızca İngilizce bir modelle karşılaştırıldığında görevde yetersiz eğitimle karıştırılabilirdi.

Şekil 9'da görselleştirilen sonuçlarımız, orta miktarda hesaplama ile eğitilmiş küçük modeller için görevler ve diller arasında gerçekten negatif transfer olduğunu göstermektedir: ortak modeller, aynı miktarda hesaplama için eğitilmiş yalnızca İngilizce modellerden daha düşük performans gösterir. Ancak, çok görevli ve çok dilli

modeller daha iyi ölçeklenir ve en büyük deneylerimizde yalnızca İngilizce modellerden daha iyi performans gösterir, bu da diğer görevlerden pozitif transfer olduğunu gösterir. En büyük

deneylerimiz için, ortak modeller, görev başına harcanan hesaplama ayarlanmadığında bile yalnızca İngilizce modellerden biraz daha iyi performans gösterir.

Şekil 9. Çok görevli ve çok dilli transfer ölçekle iyileşir. Küçük modeller için, çok görevli ve çok dilli bir kurulumda birlikte eğitildiğinde İngilizce konuşma tanıma performansı düşer. Ancak, çok dilli ve çok görevli modeller ölçekten daha fazla fayda sağlar ve sonunda yalnızca İngilizce veriler üzerinde eğitilmiş modellerden daha iyi performans gösterir. %95 bootstrap tahmini güven aralıkları gösterilmiştir.

4.4. Metin Normalleştirme

Metin normalleştirmemizi, zararsız kelime hatalarını göz ardı etmek için Whisper ile birlikte geliştirdiğimizden, normalleştiricimizin genel transkripsiyon varyasyonunu ele almak yerine Whisper'ın tuhaflıklarını düzeltmeye aşırı uyum sağlama riski vardır. Bunu kontrol etmek için, normalleştiricimizi kullanarak Whisper'ın performansını, FairSpeech projesinden (Koenecke ve diğerleri, 2020) bağımsız olarak geliştirilmiş bir normalleştiriciyle karşılaştırdık. Şekil 10'da farklılıkları görselleştiriyoruz. Çoğu veri setinde, iki normalleştirici benzer şekilde performans gösterir, Whisper ve karşılaştırılan açık kaynak modeller arasında WER azaltmada önemli farklılıklar yoktur, ancak bazı veri setlerinde, özellikle WSJ, CallHome ve Switchboard'da, normalleştiricimiz Whisper modellerinin WER'sini önemli ölçüde daha fazla azaltır. Azaltmadaki farklılıklar, gerçek veriler tarafından kullanılan farklı formatlara ve iki normalleştiricinin bunları nasıl cezalandırdığına kadar izlenebilir. Örneğin, CallHome ve Switchboard'da, standardizasyon aracımız "you're" yerine "you are" gibi yaygın İngilizce kısaltmalardaki farklılıkları cezalandırmadı ve WSJ'de, normalleştiricimiz "sixty-eight million dollars" yerine "$68 million" gibi yazılı ve konuşulan sayısal ve parasal ifadeleri standartlaştırdı.

Şekil 10. Çoğu veri setinde, metin normalleştiricimiz, Whisper modelleri ve diğer açık kaynak modeller arasında WER'leri azaltmada FairSpeech'in normalleştiricisine benzer etkiye sahiptir. Her veri seti için, kutu grafiği, değerlendirme süitimizdeki farklı modeller arasında göreceli WER azaltma dağılımını gösterir ve metin normalleştiricimizi kullanmanın genellikle FairSpeech'inkinden daha düşük WER'lerle sonuçlandığını gösterir. Bazı veri setlerinde normalleştiricimiz WER'yi önemli ölçüde azaltır ve CallHome ve Switchboard gibi gerçek verilerde birçok kısaltma içeren ve WSJ gibi birçok sayısal ifade içeren Whisper modelleri için daha da fazla azaltır.

4.5. Güvenilir Uzun Form Transkripsiyon Stratejileri

Whisper kullanarak uzun form ses transkripsiyonu, modelin 30 saniyelik ses bağlam penceresini ne kadar kaydıracağını belirlemek için zaman damgası token'larının doğru tahminine dayanır ve bir penceredeki yanlış transkripsiyon, sonraki pencerelerdeki transkripsiyonu olumsuz etkileyebilir. Uzun form transkripsiyonunun başarısızlık durumlarını önlemeye yardımcı olan bir dizi buluşsal yöntem geliştirdik, bu yöntemler Bölüm 3.8 ve 3.9'da rapor edilen sonuçlarda uygulanmıştır. İlk olarak, açgözlü kod çözmede daha sık meydana gelen tekrar döngüsünü azaltmak için puan

fonksiyonu olarak log olasılığını kullanarak 5 ışınlı ışın aramasını kullanırız. Sıcaklığı 0 ile başlatırız, yani her zaman en yüksek olasılığa sahip token'ları seçeriz ve üretilen token'ların ortalama log olasılığı -1'den düşük olduğunda veya üretilen metnin gzip sıkıştırma oranı 2.4'ten yüksek olduğunda sıcaklığı 0.2'ye kadar 1.0'a yükseltiriz. Uygulanan sıcaklık 0.5'in altında olduğunda önceki pencereden transkript edilmiş metni önceki metin koşullandırması olarak sağlamak performansı daha da iyileştirir. Yalnızca <|nospeech|> token'ının olasılığının yeterli olmadığını bulduk.

Tablo 7. Ek kod çözme buluşsal yöntemleri kullanıldıkça uzun form transkripsiyon performansı artar. Her müdahalenin ayrıntıları Bölüm 4.5'te açıklanmıştır.

	TED-LIUM3	Meanwhile	Kincaid46	Rev16	Earnings-21	Earnings-22	CORAAL	Ortalama
Greedy decoding only	3.95	5.16	9.69	11.7	10.7	14.0	22.0	11.0
+Beam search	4.16	5.71	9.42	11.5	10.2	13.4	20.0	10.6
+Temperature fallback	4.16	5.71	9.42	11.5	10.2	13.4	20.0	10.6
+Voice activity detection	3.56	4.61	9.45	11.4	10.1	13.2	19.4	10.2
+Previous text conditioning	3.42	6.16	8.72	11.0	9.63	13.3	18.1	10.0
+Initial timestamp constraint	3.51	5.26	8.41	11.5	9.73	12.6	19.1	10.0

Konuşma olmayan bir segmenti ayırt etmek için <|nospeech|> token'ının olasılığının tek başına yeterli olmadığını, ancak 0.6'lık konuşma olmayan olasılık eşiği ile -1'lik ortalama log-olasılık eşiğinin birleştirilmesinin Whisper'ın ses etkinliği tespitini daha güvenilir hale getirdiğini bulduk. Son olarak, modelin girişteki ilk birkaç kelimeyi göz ardı ettiği bir hata modunu önlemek için, başlangıç zaman damgası token'ını 0.0 ile 1.0 saniye arasında olacak şekilde kısıtladık. Tablo 7, yukarıdaki müdahalelerin her birinin WER'yi genel olarak kademeli olarak azalttığını, ancak veri seti genelinde eşit olmadığını göstermektedir. Bu buluşsal yöntemler, modelin gürültülü tahminleri için bir geçici çözüm olarak hizmet eder ve uzun form kod çözmenin güvenilirliğini daha da iyileştirmek için daha fazla araştırmaya ihtiyaç duyulacaktır.

5. Iℓgili Çalışmalar

Konuşma Tanımayı Ölçeklendirme Konuşma tanıma araştırmalarında tutarlı bir tema, hesaplama, modeller ve veri setlerini ölçeklendirmenin faydalarını belgelemek olmuştur. Derin öğrenmeyi konuşma tanımaya uygulayan erken çalışmalar, model derinliği ve boyutuyla iyileştirilmiş performans buldu ve bu daha büyük modelleri eğitilebilir kılmak için GPU hızlandırmasından yararlandı (Mohamed ve diğerleri, 2009). Daha fazla araştırma, derin öğrenme yaklaşımlarının konuşma tanımaya faydasının veri seti boyutuyla arttığını, telefon tanıma için sadece 3 saatlik TIMIT eğitim verisi kullanıldığında önceki GMM-HMM sistemleriyle sadece rekabetçi olmaktan, 2.000 saatlik Switchboard veri setinde eğitildiğinde %30 kelime hata oranı azalması elde etmeye kadar iyileştiğini gösterdi (Seide ve diğerleri, 2011). Liao ve diğerleri (2013), derin öğrenmeye dayalı bir konuşma tanıma veri setinin boyutunu 1.000 saatin üzerinde artırmak için zayıf denetimli öğrenmeden yararlanan erken bir örnektir. Bu eğilimler, Deep Speech 2 (Amodei ve diğerleri, 2015) ile devam etti; bu, 16 GPU üzerinde yüksek verimli dağıtılmış eğitimi geliştiren ve 12.000 saatlik eğitim verisine ölçeklenen ve bu ölçekte sürekli iyileşmeler gösteren dikkate değer bir sistemdir. Yarı denetimli ön eğitimden yararlanarak, Narayanan ve diğerleri (2018) veri seti boyutunu çok daha fazla artırabildi ve 162.000 saatlik etiketli ses üzerinde eğitimi inceledi. Daha yeni çalışmalar, milyar parametreli modelleri (Zhang ve diğerleri, 2020) ve 1.000.000 saate kadar eğitim verisi (Zhang ve diğerleri, 2021) kullanmayı araştırmıştır.

Çok Görevli Öğrenme Çok görevli öğrenme (Caruana, 1997) uzun zamandır incelenmektedir. Konuşma tanımada, çok dilli modeller on yıldan fazla bir süredir araştırılmaktadır (Schultz & Kirchhoff, 2006). NLP'de çok görevli öğrenmeyi tek bir modelle keşfeden ilham verici ve temel bir çalışma Collobert ve diğerleri (2011)'dir. Birden fazla kodlayıcı ve kod çözücü kullanarak sıra-sıra çerçevesinde (Sutskever ve diğerleri, 2014) çok görevli öğrenme Luong ve diğerleri (2015) tarafından araştırılmıştır. Paylaşılan bir kodlayıcı/kod çözücü mimarisi ile dil kodlarının kullanımı ilk olarak Johnson ve diğerleri (2017) tarafından makine çevirisi için gösterilmiş, ayrı kodlayıcı ve kod çözücülere olan ihtiyacı ortadan kaldırmıştır. Bu yaklaşım, McCann ve diğerleri (2018) tarafından "metinden metne" çerçevesine daha da basitleştirilmiş ve Radford ve diğerleri (2019) ve Raffel ve diğerleri (2020) çalışmalarında büyük transformer dil modelleriyle başarısıyla popüler hale gelmiştir. Toshniwal ve diğerleri (2018), modern bir derin öğrenme konuşma tanıma sistemini tek bir modelle birkaç dilde birlikte eğitmeyi göstermiş ve Pratap ve diğerleri (2020a) bu çalışma hattını milyar parametreli bir modelle 50 dile önemli ölçüde ölçeklendirmiştir. MUTE (Wang ve diğerleri, 2020c) ve mSLAM (Bapna ve diğerleri, 2022), hem metin hem de konuşma dili görevleri üzerinde ortak eğitimi incelemiş, aralarında transfer olduğunu göstermiştir.

Sağlamlık Modellerin ne kadar etkili bir şekilde transfer olduğu ve dağıtım kaymasına ve diğer pertürbasyon türlerine karşı ne kadar sağlam oldukları sorusu uzun zamandır incelenmekte ve makine öğreniminin birçok alanında aktif olarak araştırılmaktadır. Torralba & Efros (2011), on yıldan fazla bir süre önce makine öğrenimi modellerinin veri setleri arasındaki genelleme eksikliğini vurguladı. Birçok başka çalışma, IID test setlerinde yüksek performansa rağmen, makine öğrenimi modellerinin hafifçe farklı ayarlarda değerlendirildiğinde bile birçok hata yapabileceğini göstermiş ve sürekli olarak yinelemiştir (Lake ve diğerleri, 2017; Jia & Liang, 2017; Alcorn ve diğerleri, 2019; Barbu ve diğerleri, 2019; Recht ve diğerleri, 2019). Daha yakın zamanda, Taori ve diğerleri (2020)

görüntü sınıflandırma modellerinin sağlamlığını incelemiş ve Miller ve diğerleri (2020) bunu soru-cevap modelleri için araştırmıştır. Temel bir bulgu, Giriş bölümünde tartışıldığı gibi çok alanlı eğitimin sağlamlığı ve genellemeyi artırdığı olmuştur. Bu bulgu, NLP (Hendrycks ve diğerleri, 2020) ve bilgisayar görüşü (Radford ve diğerleri, 2021) dahil olmak üzere konuşma tanıma dışındaki birçok alanda tekrarlanmıştır.

6. Sınırlamalar ve Gelecek Çalışmalar

Deneysel sonuçlarımızdan, analizlerimizden ve ablasyonlarımızdan, birkaç sınırlama ve gelecek çalışma alanı belirledik.

Iẏileştirilmiş kod çözme stratejileri. Whisper'ı ölçeklendirdikçe, daha büyük modellerin benzer sesli kelimeleri karıştırma gibi algılamayla ilgili hataları azaltmada istikrarlı ve güvenilir ilerleme kaydettiğini gözlemledik. Özellikle uzun form transkripsiyonda kalan birçok hata, doğası gereği daha inatçı ve kesinlikle insan dışı/algısal görünmektedir. Bunlar, tekrar döngülerine takılma, bir ses segmentinin ilk veya son birkaç kelimesini transkript etmeme veya modelin gerçek sesle tamamen ilgisiz bir transkript çıkaracağı tam halüsinasyon gibi seq2seq modellerinin, dil modellerinin ve metin-ses hizalamasının başarısızlık modlarının bir kombinasyonudur. Bölüm 4.5'te tartışılan kod çözme detayları önemli ölçüde yardımcı olsa da, Whisper modellerini yüksek kaliteli denetimli bir veri seti üzerinde ince ayar yapmanın ve/veya kod çözme performansını daha doğrudan optimize etmek için pekiştirmeli öğrenme kullanmanın bu hataları daha da azaltmaya yardımcı olabileceğini düşünüyoruz.

Daha Az Kaynaklı Diller Için Eğitim Verilerini Artırma Şekil 3'ün gösterdiği gibi, Whisper'ın konuşma tanıma performansı birçok dilde hala oldukça zayıftır. Aynı analiz, performansın dil için eğitim verisi miktarı tarafından çok iyi tahmin edildiği için açık bir iyileşme yolu önermektedir. Ön eğitim veri setimiz şu anda veri toplama boru hattımızın önyargıları nedeniyle çok Iṅgilizce ağırlıklı olduğundan, internetin Iṅgilizce merkezli kısımlarından kaynaklanan, çoğu dilde 1000 saatten az eğitim verisi vardır. Bu nadir diller için veri miktarını artırmaya yönelik hedeflenmiş bir çaba, genel eğitim veri seti boyutumuzda küçük bir artışla bile ortalama konuşma tanıma performansında büyük bir iyileşmeyle sonuçlanabilir.

Iṅce Ayarı Iṅceleme Bu çalışmada, konuşma işleme sistemlerinin sağlamlık özelliklerine odaklandık ve sonuç olarak yalnızca Whisper'ın sıfır atış transfer performansını inceledik. Bu, genel güvenilirliği temsil ettiği için incelenmesi gereken kritik bir ayar olsa da, yüksek kaliteli denetimli konuşma verilerinin mevcut olduğu birçok alanda, ince ayar yaparak sonuçların daha da iyileştirilmesi muhtemeldir. Iṅce ayarı incelemenin ek bir faydası, daha yaygın bir değerlendirme ayarı olduğu için önceki çalışmalarla doğrudan karşılaştırmalara olanak sağlamasıdır.

Dil Modellerinin Sağlamlık Üzerindeki Etkisini Iṅceleme Giriş bölümünde tartışıldığı gibi, Whisper'ın sağlamlığının kısmen güçlü kod çözücüsünden, yani ses koşullu bir dil modelinden kaynaklandığından şüpheleniyoruz. Whisper'ın faydalarının kodlayıcısını, kod çözücüsünü veya her ikisini eğitmekten ne ölçüde kaynaklandığı şu anda belirsizdir. Bu, Whisper'ın çeşitli tasarım bileşenlerini ablasyon yaparak, örneğin kod çözücüsüz bir CTC modeli eğiterek veya wav2vec 2.0

gibi mevcut konuşma tanıma kodlayıcılarının bir dil modeliyle birlikte kullanıldığında performansının nasıl değiştiğini inceleyerek incelenebilir.

7. Sonuç

Whisper, zayıf denetimli ön eğitimin konuşma tanıma araştırmalarında şimdiye kadar yeterince takdir edilmediğini göstermektedir. Son büyük ölçekli konuşma tanıma çalışmalarının temel taşı olan kendi kendine denetim ve kendi kendine eğitim tekniklerine ihtiyaç duymadan sonuçlarımızı elde ediyoruz ve büyük ve çeşitli bir denetimli veri seti üzerinde basitçe eğitimin ve sıfır atış transferine odaklanmanın bir konuşma tanıma sisteminin sağlamlığını önemli ölçüde nasıl iyileştirebileceğini gösteriyoruz.

TEŞEKKÜRLER

Whisper tarafından kullanılan verilerin oluşturulmasında yer alan milyonlarca insana teşekkür etmek istiyoruz. Ayrıca, bu projeye ilham veren şelale yürüyüşündeki sohbet için Nick Ryder, Will Zhuk ve Andrew Carr' a teşekkür etmek istiyoruz. Ayrıca, bu projenin kullandığı yazılım ve donanım altyapısı üzerindeki kritik çalışmaları için OpenAI' deki Hızlandırma ve Süper Bilgisayar ekiplerine minnettarız. Projeye politika perspektifinden danışmanlık yaptığı için Pamela Mishkin' e de teşekkür etmek istiyoruz. Son olarak, bu proje boyunca kullanılan birçok yazılım paketinin geliştiricilerine, Numpy (Harris ve diğerleri, 2020), SciPy (Virtanen ve diğerleri, 2020), ftfy (Speer, 2019), PyTorch (Paszke ve diğerleri, 2019), pandas (pandas geliştirme ekibi, 2020) ve scikit-learn (Pedregosa ve diğerleri, 2011) dahil ancak bunlarla sınırlı olmamak üzere minnettarız.

Referanslar

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., ve Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. IEEE/CVF Bilgisayar Görüşü ve Örüntü Tanıma Konferansı Bildirileri, ss. 4845–4854, 2019.

Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., ve diğerleri. Deep speech 2: end-to-end speech recognition in english and mandarin. arxiv. arXiv ön baskısı arXiv:1512.02595, 2015.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M.,

ve Weber, G. Common voice: Büyük ölçekli çok dilli bir konuşma külliyatı. arXiv ön baskısı arXiv:1912.06670, 2019.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., ve diğerleri. XLS-R: Ölçekte kendi kendine denetimli çapraz dilli konuşma temsili öğrenimi. arXiv ön baskısı arXiv:2111.09296, 2021.

Baevski, A., Zhou, H., Mohamed, A., ve Auli, M. wav2vec 2.0: Konuşma temsillerinin kendi kendine denetimli öğrenimi için bir çerçeve. arXiv ön baskısı arXiv:2006.11477, 2020.

Baevski, A., Hsu, W.-N., Conneau, A., ve Auli, M. Denetimsiz konuşma tanıma. Neural Information Processing Systems Gelişmeleri, 34:27826–27839, 2021.

Bapna, A., Cherry, C., Zhang, Y., Jia, Y., Johnson, M., Cheng, Y., Khanuja, S., Riesa, J., ve Conneau, A. mslam: Konuşma ve metin için büyük ölçekli çok dilli ortak ön eğitim. arXiv ön baskısı arXiv:2202.01374, 2022.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., ve Katz, B. Objectnet: Nesne tanıma modellerinin sınırlarını zorlamak için büyük ölçekli önyargı kontrollü bir veri seti. Neural information processing systems gelişmeleri, 32, 2019.

Caruana, R. Çok görevli öğrenme. Machine learning, 28(1):41–75, 1997.

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., ve Norouzi, M. SpeechStew: Tek bir büyük sinir ağı eğitmek için mevcut tüm konuşma tanıma verilerini basitçe karıştırın. arXiv ön baskısı arXiv:2104.02133, 2021.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., ve diğerleri. Gigaspeech: 10.000 saat transkript edilmiş ses ile gelişen, çok alanlı bir asr külliyatı. arXiv ön baskısı arXiv:2106.06909, 2021.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., ve diğerleri. Unispeech-sat: Konuşmacı farkındalıklı ön eğitim ile evrensel konuşma temsili öğrenimi. ICASSP 2022-2022 IEEE Uluslararası Akustik, Konuşma ve Sinyal İşleme Konferansı (ICASSP), ss. 6152–6156. IEEE, 2022a.

Chen, T., Xu, B., Zhang, C., ve Guestrin, C. Derin ağları alt doğrusal bellek maliyetiyle eğitme. arXiv ön baskısı arXiv:1604.06174, 2016.

Chen, Z., Zhang, Y., Rosenberg, A., Ramabhadran, B., Moreno, P., Bapna, A., ve Zen, H. Maestro: Modalite eşleştirmesi yoluyla eşleşen konuşma metin temsilleri. arXiv ön baskısı arXiv:2204.03409, 2022b.

Child, R., Gray, S., Radford, A., ve Sutskever, I. Seyrek transformer' larla uzun diziler oluşturma. arXiv ön baskısı arXiv:1904.10509, 2019.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., ve Kuksa, P. Doğal dil işleme (neredeyse) sıfırdan. Journal of machine learning research, 12(ARTICLE):2493–2537, 2011.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., ve Bapna, A. Fleurs: Konuşmanın evrensel temsillerinin birkaç atışlı öğrenme değerlendirmesi. arXiv ön baskısı arXiv:2205.12446, 2022.

Del Rio, M., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., McNamara, Q., Dong, J., Zelasko, P., ve Jette´, M. Earnings-21: Vahşi doğada asr için pratik bir kıyaslama. arXiv ön

baskısı arXiv:2104.11348, 2021.

Galvez, D., Diamos, G., Torres, J. M. C., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., ve Reddi, V. J. The people' s speech: Ticari kullanım için büyük ölçekli çeşitli İngilizce konuşma tanıma veri seti. arXiv ön baskısı arXiv:2111.09344, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., ve Wichmann, F. A. Derin sinir ağlarında kısayol öğrenimi. Nature Machine Intelligence, 2(11):665–673, 2020.

Caruana, R. Çok görevli öğrenme. Machine learning, 28(1):41–75, 1997.

Chan, W., Park, D., Lee, C., Zhang, Y., Le, Q., ve Norouzi, M. SpeechStew: Tek bir büyük sinir ağı eğitmek için mevcut tüm konuşma tanıma verilerini basitçe karıştırın. arXiv ön baskısı arXiv:2104.02133, 2021.

Chen, G., Chai, S., Wang, G., Du, J., Zhang, W.-Q., Weng, C., Su, D., Povey, D., Trmal, J., Zhang, J., ve diğerleri. Gigaspeech: 10.000 saat transkript edilmiş ses ile gelişen, çok alanlı bir asr külliyatı. arXiv ön baskısı arXiv:2106.06909, 2021.

Chen, S., Wu, Y., Wang, C., Chen, Z., Chen, Z., Liu, S., Wu, J., Qian, Y., Wei, F., Li, J., ve diğerleri. Unispeech-sat: Konuşmacı farkındalıklı ön eğitim ile evrensel konuşma temsili öğrenimi. ICASSP 2022-2022 IEEE Uluslararası Akustik, Konuşma ve Sinyal İşleme Konferansı (ICASSP), ss. 6152–6156. IEEE, 2022a.

Chen, T., Xu, B., Zhang, C., ve Guestrin, C. Derin ağları alt doğrusal bellek maliyetiyle eğitme. arXiv ön baskısı arXiv:1604.06174, 2016.

Chen, Z., Zhang, Y., Rosenberg, A., Ramabhadran, B., Moreno, P., Bapna, A., ve Zen, H. Maestro: Modalite eşleştirmesi yoluyla eşleşen konuşma metin temsilleri. arXiv ön baskısı arXiv:2204.03409, 2022b.

Child, R., Gray, S., Radford, A., ve Sutskever, I. Seyrek transformer' larla uzun diziler oluşturma. arXiv ön baskısı arXiv:1904.10509, 2019.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., ve Kuksa, P. Doğal dil işleme (neredeyse) sıfırdan. Journal of machine learning research, 12(ARTICLE):2493–2537, 2011.

Conneau, A., Ma, M., Khanuja, S., Zhang, Y., Axelrod, V., Dalmia, S., Riesa, J., Rivera, C., ve Bapna, A. Fleurs: Konuşmanın evrensel temsillerinin birkaç atışlı öğrenme değerlendirmesi. arXiv ön baskısı arXiv:2205.12446, 2022.

Del Rio, M., Delworth, N., Westerman, R., Huang, M., Bhandari, N., Palakapilly, J., McNamara, Q., Dong, J., Zelasko, P., ve Jette´, M. Earnings-21: Vahşi doğada asr için pratik bir kıyaslama. arXiv ön baskısı arXiv:2104.11348, 2021.

Galvez, D., Diamos, G., Torres, J. M. C., Achorn, K., Gopi, A., Kanter, D., Lam, M., Mazumder, M., ve Reddi, V. J. The people' s speech: Ticari kullanım için büyük ölçekli çeşitli İngilizce konuşma tanıma veri seti. arXiv ön baskısı arXiv:2111.09344, 2021.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., ve Wichmann, F. A. Derin sinir ağlarında kısayol öğrenimi. Nature Machine Intelligence, 2(11):665–673, 2020.

Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., ve Cherry, C. Sinirsel makine çevirisi için ölçekleme yasaları. arXiv ön baskısı arXiv:2109.07740, 2021.

Griewank, A. ve Walther, A. Algorithm 799: revolve: hesaplamalı farklılaşmanın ters veya ek modunda kontrol noktası oluşturmanın bir uygulaması. ACM Transactions on Mathematical Software (TOMS), 26(1):19–45, 2000.

Gunter, K., Vaughn, C., ve Kendall, T. Contextualizing/s/retraction: Washington dc Afrika Amerikan dilinde sibilant varyasyonu ve değişimi. Language Variation and Change, 33(3):331–357, 2021.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Ferna´ndez del R´ıo, J., Wiebe, M., Peterson, P., Ge´rard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., ve Oliphant, T. E. NumPy ile dizi programlama. Nature, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

Hendrycks, D. ve Gimpel, K. Gaussian error linear units (gelus). arXiv ön baskısı arXiv:1606.08415, 2016.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., ve Song, D. Önceden eğitilmiş transformer’ lar dağıtım dışı sağlamlığı iyileştirir. arXiv ön baskısı arXiv:2004.06100, 2020.

Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N.A., ve Este`ve, Y. Ted-lium 3: Konuşmacı adaptasyonu deneyleri için iki kat daha fazla veri ve külliyat yeniden bölümlemesi. SPECOM, 2018.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., ve Mohamed, A. Hubert: Gizli birimlerin maskeli tahmini ile kendi kendine denetimli konuşma temsili öğrenimi. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021a.

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., ve diğerleri. Sağlam wav2vec 2.0: Kendi kendine denetimli ön eğitimde alan kaymasını analiz etme. arXiv ön baskısı arXiv:2104.01027, 2021b.

Huang, G., Sun, Y., Liu, Z., Sedra, D., ve Weinberger, K. Q. Stokastik derinlikli derin ağlar. European conference on computer vision, ss. 646–661. Springer, 2016.

Jia, R. ve Liang, P. Okuma anlama sistemlerini değerlendirmek için düşmanca örnekler. arXiv ön baskısı arXiv:1707.07328, 2017.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Vie´gas, F., Wattenberg, M., Corrado, G., ve diğerleri. Google’ ın çok dilli sinirsel makine çevirisi sistemi: Sıfır atış çeviriyi etkinleştirme. Transactions of the Association for Computational Linguistics, 5:339–351, 2017.

Kendall, T. ve Farrington, C. Bölgesel Afro-Amerikan dili külliyatı. Sürüm 2021.07. Eugene, OR: The Online Resources for African American Language Project. http://oraal.uoregon.edu/coraal, 2021.

Erişim tarihi: 2022-09-01.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., ve Goel, S. Otomatik konuşma tanımada ırksal eşitsizlikler. Proceedings of the National Academy of Sciences, 117(14):7684–7689, 2020.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., ve Houlsby, N. Büyük transfer (bit): Genel görsel temsil öğrenimi. European conference on computer vision, ss. 491–507. Springer, 2020.

Kuchaiev, O., Li, J., Nguyen, H., Hrinchuk, O., Leary, R., Ginsburg, B., Kriman, S., Beliaev, S., Lavrukhin, V., Cook, J., ve diğerleri. Nemo: AI uygulamaları oluşturmak için bir araç seti. arXiv ön baskısı arXiv:1909.09577, 2019.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., ve Gershman, S. J. İnsanlar gibi öğrenen ve düşünen makineler inşa etmek. Behavioral and brain sciences, 40, 2017.

Liao, H., McDermott, E., ve Senior, A. Youtube video transkripsiyonu için yarı denetimli eğitim verileriyle büyük ölçekli derin sinir ağı akustik modellemesi. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, ss. 368–373. IEEE, 2013.

Likhomanenko, T., Xu, Q., Pratap, V., Tomasello, P., Kahn, J., Avidov, G., Collobert, R., ve Synnaeve, G. ASR' de değerlendirmeyi yeniden düşünmek: Modellerimiz yeterince sağlam mı? arXiv ön baskısı arXiv:2010.11745, 2020.

Loshchilov, I. ve Hutter, F. Ayrık ağırlık düşüşü düzenlemesi. arXiv ön baskısı arXiv:1711.05101, 2017.

Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., ve Kaiser, L. Çok görevli sıra-sıra öğrenme. arXiv ön baskısı arXiv:1511.06114, 2015.

Mahajan, D., Girshick, R., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., ve Van Der Maaten, L. Zayıf denetimli ön eğitimin sınırlarını keşfetme. European conference on computer vision (ECCV) Bildirileri, ss. 181–196, 2018.

Mauch, M. ve Ewert, S. Ses bozulma araç kutusu ve sağlamlık değerlendirmesine uygulaması. 14. Uluslararası Müzik Bilgi Edinme Topluluğu Konferansı (ISMIR 2013), Curitiba, Brezilya, 2013 Bildirileri. kabul edildi.

McCann, B., Keskar, N. S., Xiong, C., ve Socher, R. Doğal dil dekatlonu: Soru yanıtlama olarak çok görevli öğrenme. arXiv ön baskısı arXiv:1806.08730, 2018.

Meyer, J., Rauchenstein, L., Eisenberg, J. D., ve Howell, N. Artie bias külliyatı: Konuşma uygulamalarında demografik önyargıyı tespit etmek için açık bir veri seti. 12. Dil Kaynakları ve Değerlendirme Konferansı Bildirileri, ss. 6462–6468, Marsilya, Fransa, Mayıs 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.796.

Miller, J., Krauth, K., Recht, B., ve Schmidt, L. Doğal dağıtım kaymasının soru yanıtlama modelleri üzerindeki etkisi. ICML, 2020.

Mohamed, A.-r., Dahl, G., Hinton, G., ve diğerleri. Telefon tanıma için derin inanç ağları. Nips workshop on deep learning for speech recognition and related applications, cilt 1, ss. 39, 2009.

Narayanan, A., Misra, A., Sim, K. C., Pundak, G., Tripathi, A., Elfeky, M., Haghani, P., Strohman, T., ve Bacchiani, M. Büyük ölçekli eğitim yoluyla alan invariant konuşma tanımaya doğru. 2018 IEEE Spoken Language Technology Workshop (SLT), ss. 441–447. IEEE, 2018.

Panayotov, V., Chen, G., Povey, D., ve Khudanpur, S. Librispeech: Kamu malı sesli kitaplara dayalı bir asr külliyatı. 2015 IEEE uluslararası akustik, konuşma ve sinyal işleme konferansı (ICASSP), ss. 5206–5210. IEEE, 2015.

pandas geliştirme ekibi, T. pandas-dev/pandas: Pandas, Şubat 2020. URL https://doi.org/10.5281/zenodo.3509134.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., ve Le, Q. V. SpecAugment: Otomatik konuşma tanıma için basit bir veri artırma yöntemi. arXiv ön baskısı arXiv:1904.08779, 2019.

Pascanu, R., Mikolov, T., ve Bengio, Y. Tekrarlayan sinir ağlarını eğitmenin zorluğu üzerine. International conference on machine learning, ss. 1310–1318. PMLR, 2013.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., ve Chintala, S. Pytorch: Zorunlu bir stil, yüksek performanslı derin öğrenme kütüphanesi. Advances in Neural Information Processing Systems 32, ss. 8024–8035, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., ve Duchesnay, E. Scikit-learn: Python' da makine öğrenimi. Journal of Machine Learning Research, 12:2825–2830, 2011.

Polyak, B. T. ve Juditsky, A. B. Ortalama alma yoluyla stokastik yaklaşımın hızlandırılması. SIAM journal on control and optimization, 30(4):838–855, 1992.

Pratap, V., Sriram, A., Tomasello, P., Hannun, A. Y., Liptchinsky, V., Synnaeve, G., ve Collobert, R. Büyük ölçekli çok dilli asr: 50 dil, 1 model, 1 milyar parametre. ArXiv, abs/2007.03001, 2020a.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., ve Collobert, R. Mls: Konuşma araştırması için büyük ölçekli çok dilli bir veri seti. arXiv ön baskısı arXiv:2012.03411, 2020b.

Press, O. ve Wolf, L. Dil modellerini iyileştirmek için çıktı gömme kullanma. 15. Avrupa Hesaplamalı Dilbilim Derneği Bölümü Konferansı Bildirileri: Cilt 2, Kısa Makaleler, ss. 157–163, Valensiya, İspanya, Nisan 2017. Hesaplamalı Dilbilim Derneği. URL https://aclanthology.org/E17-2025.

Provilkov, I., Emelianenko, D., ve Voita, E. Bpe-dropout: Basit ve etkili alt kelime düzenlemesi. arXiv ön baskısı arXiv:1910.13267, 2019.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., ve Sutskever, I. Dil modelleri denetimsiz çok görevli öğrenicilerdir. 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., ve Sutskever, I. Doğal dil denetiminden aktarılabilir görsel modeller öğrenme. arXiv ön baskısı arXiv:2103.00020, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., ve diğerleri. Birleşik bir metinden metne transformer ile transfer öğreniminin sınırlarını keşfetme. J. Mach. Learn. Res., 21(140):1–67, 2020.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., ve Bengio, Y. SpeechBrain: Genel amaçlı bir konuşma araç seti, 2021. arXiv:2106.04624.

Recht, B., Roelofs, R., Schmidt, L., ve Shankar, V. ImageNet sınıflandırıcıları ImageNet'e genelleme yapar mı? Chaudhuri, K. ve Salakhutdinov, R. (ed.), Proceedings of the 36th International Conference on Machine Learning, cilt 97, Proceedings of Machine Learning Research, ss. 5389–5400. PMLR, 09–15 Haziran 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., ve diğerleri. Imagenet büyük ölçekli görsel tanıma yarışması. International journal of computer vision, 115(3):211–252, 2015.

Schultz, T. ve Kirchhoff, K. Çok dilli konuşma işleme. Elsevier, 2006.

Seide, F., Li, G., Chen, X., ve Yu, D. Konuşma transkripsiyonu için bağlama bağlı derin sinir ağlarında özellik mühendisliği. 2011 IEEE Workshop on Automatic Speech Recognition & Understanding, ss. 24–29. IEEE, 2011.