Yapay Zeka ile Sesten Yazıya Dönüştürme

Samet Duran

10.03.2024

1 Giriş

Bu projede insanların konuşmalarının metne dönüştürülerek bilgisayar ortamına aktarılması amaçlandı bu yüzden yapay zeka ile sesten yazıya çevirme konusu seçildi. Bu konuda ki örnekler araştırıldı. Python dili kullanılmasına karar verildi. Openai şirketinin açık kaynak kodlu Whisper adlı modeli ve Mozilla şirketinin açık kaynak kodlu Deepspeech adlı modeli incelendi ve araştırıldı. Deepspeech modeli WHisper modeli ile karşılaştırıldı. Whisper modeli kullanılmaya karar verildi. Proje sonunda konuşmayı yazıya çeviren bir uygulama ortaya çıkacak.

2 Literatür Araştırması

OpenAI'nin Whisper ile Mozilla'nın Deepspeech modelleri Konuşmadan Yazıya Dönüştürme konusunda geliştirilmiş modellerdir.

2.1 Genel Bilgi:

Konuşmadan yazıya (STT) dönüştürme, sesli konuşmayı metne dönüştüren bir teknolojidir.

2.2 Whisper:

- Açık kaynaklı bir ASR modelidir.
- Şuanlık gerçek zamanlı şekilde çalışmıyor, ses dosyası yüklenerek çalıştırılabiliyor.
- Çok dilli bir konuşma tanıma sistemi olarak öne çıkar [1].
- 680.000 saatten fazla çok dilli ve webden toplanan veriyle eğitilmiştir [2].
- Bu büyük veri havuzu, benzersiz vurguları, arka plan gürültüsünü ve teknik jargonu daha iyi tanımasına olanak sağlar
- Whisper, hızlı ve doğru sonuçlar elde etmek için tasarlanmıştır.

2.3 DeepSpeech:

DeepSpeech, Baidu'nun Deep Speech araştırma makalesine [3] dayanan makine öğrenimi teknikleriyle eğitilmiş bir model kullanan açık kaynaklı bir Konuşmadan Metne motorudur. DeepSpeech Projesi, uygulamayı kolaylaştırmak için Google'ın TensorFlow 'unu kullanır. TensorFlow Google tarafından geliştirilen uçtan uca açık kaynaklı bir makine öğrenmesi kütüphanesidir.

- Mozilla tarafından geliştirilen bir ASR modelidir.
- Gerçek Zamanlı çalışıyor fakat sadece ingilizce destekliyor.
- Açık kaynaklı ve topluluk tarafından desteklenir.

- Wav2Vec 2.0 gibi unsupervised training temelinde çalışır.
- Raw waveform representation, context part ve linear layer olmak üzere üç bölümden oluşur.
- Wav2Vec 2.0, konuşma tanıma alanında önemli bir başarıya sahiptir

Sonuç olarak Whisper gerçek zamanlı çalışmıyor fakat DeepSpeech modeline göre çok daha doğru sonuçlar veriyor.

3 Sesten Yazıya Çevirmenin Metodolojisi

Whisper'ın mimari tasarımı basit ve etkilidir. Bir encoder-decoder Transformer olarak uygulanmıştır. İşlem adımları şu şekildedir:

- Giriş Sesinin Bölünmesi: Giriş sesi 30 saniyelik parçalara bölünür.
- Log-Mel Spektrogramına Dönüştürme: Her parça bir log-Mel spektrogramına dönüştürülür. Log Mel spektrogramı, ses işleme alanında yaygın olarak kullanılan bir görsel temsil yöntemidir. Bu yöntem, ses sinyallerinin frekans içeriğini zamanla nasıl değiştirdiğini görselleştirmek için kullanılır [4].

  Log Mel Spektogramı hakkında daha ayrıntılı açıklama:

  Ses Sinyali ve Örnekleme: Bir ses sinyali, zaman içinde belirli bir niceliğin değişimini temsil eder. Ses sinyalleri için bu nicelik hava basıncıdır. Ses sinyalini dijital olarak işlemek için zaman içinde hava basıncının örneklerini alabiliriz. Bu örnekleri almak için veri örnekleme hızı genellikle 44.1 kHz'dir (yani saniyede 44,100 örnek). Bu örnekleme işlemiyle, ses sinyalinin dijital bir temsilini elde ederiz.

  Fourier Dönüşümü: Ses sinyali, birçok tek frekanslı ses dalgasından oluşur. Zaman içinde sinyalin örneklerini alırken, sadece sonuç amplitüdlerini yakalarız. Fourier dönüşümü, bir sinyali frekans bileşenlerine ayırarak frekans alanına dönüştüren matematiksel bir formüldür. Bu, sinyali zamandan frekansa dönüştürür ve sonuç spektrum olarak adlandırılır. Her sinyal, orijinal sinyale toplamı veren bir dizi sinüs ve kosinüs dalgasına ayrılabilir. Bu, Fourier teoremi olarak bilinir.

  Hızlı Fourier Dönüşümü (FFT): FFT, Fourier dönüşümünü verimli bir şekilde hesaplayabilen bir algoritmadır ve sinyal işlemede yaygın olarak kullanılır. Örnek ses verisini FFT algoritmasıyla analiz edebiliriz. Bu, sinyalin frekans içeriğini incelememizi sağlar.

  Spektrogram: FFT, bir sinyalin frekans içeriğini analiz etmek için güçlü bir araçtır, ancak sinyalin frekans içeriği zamanla nasıl değişirse? Bu durum, müzik ve konuşma gibi çoğu ses sinyali için geçerlidir. Spektrogram, bu tür sinyallerin zaman içindeki spektrumunu temsil etmek için

kullanılır. Log Mel spektrogramı, önce sinyalin Mel ölçeğindeki logaritmik güç spektrumunu hesaplar ve ardından bu spektrumu görselleştirir. Mel ölçeği, insan işitme algısına daha iyi uyan bir frekans ölçeğidir.

- Kodlayıcı (Encoder): Log-Mel spektrogramı kodlayıcıya iletilir. Kodlayıcı, giriş sesini temsil eden bir vektör oluşturur.
- Dekoder (Decoder): İlgili metin başlığını tahmin etmek üzere eğitilmiş bir dekoder kullanılır. Bu sırada tek bir modelin dil tanımlaması, ifade düzeyinde zaman damgaları, çok dilli konuşma transkripti ve İngilizce'ye çeviri gibi görevleri gerçekleştirmesi için özel belirteçlerle karıştırılır.
- Whisper, geniş ve çeşitli bir veri kümesi üzerinde eğitildiği için, farklı aksanlara, arka plan gürültüsüne ve teknik dil kullanımına daha iyi uyum sağlar. Ayrıca, birden fazla dilde transkript yapabilir ve bu dilleri İngilizce'ye çevirebilir.

3.1 Sesten yazıya çevirme teknolojisinin kullanım alanları:

- Toplantı ve ders notları alma
- Video ve ses dosyalarını metne dönüştürme
- Engelli kişilere erişilebilirlik sağlama
- Müşteri hizmetlerinde ve çağrı merkezlerinde
- Altyazı oluşturma

4 VeriTabanı ve Veriler

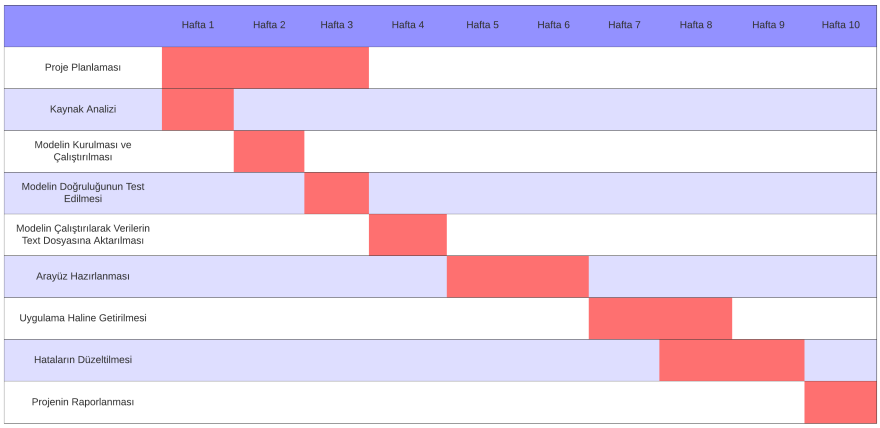
DeepSpeech'in İngilizce modeli, öğrenme için LibriSpeech, Fisher ve Santral projelerinden gelen verileri içerir. Ayrıca, Common Voice adlı bir veri kümesi kullanılmıştır. Bu veri kümesi, yaklaşık 3816 saatlik transkripsiyonlu radyo programı kayıtlarını içerir [5]. Ayrıca, DeepSpeech'in deneysel Mandarin Çince akustik modelleri de bulunmaktadır. Bu modeller, 2000 saatlik okuma metni içeren bir iç veri kümesi üzerinde eğitilmiştir [6]. Bu modeller, bellek haritalı (".pbmm" uzantılı) ve TensorFlow Lite kullanılarak dönüştürülmüş (".tflite" uzantılı) olmak üzere iki farklı formatta sunulmaktadır.

Mevcut diğer yaklaşımlar sıklıkla daha küçük, daha yakın eşleştirilmiş sesmetin eğitim veri kümeleri kullanmakta, [7], [8], [9] veya geniş ancak denetimsiz ses ön eğitimi kullanmaktadır. [10], [11] Whisper büyük ve çeşitli bir veri kümesi üzerinde eğitildiği ve herhangi bir veri kümesine göre ince ayar yapılmadığı için, konuşma tanıma alanında rekabetçi bir ölçüt olarak bilinen LibriSpeech performansında uzmanlaşmış modelleri geçememektedir. Bununla birlikte, Whisper'ın sıfır atış performansını birçok farklı veri kümesinde ölçtüğümüzde, çok daha sağlam olduğunu ve bu modellerden %50 daha az hata yaptığını görüyoruz.

Whisper'ın ses veri kümesinin yaklaşık üçte biri İngilizce değildir ve dönüşümlü olarak orijinal dilde yazıya dökme veya İngilizceye çevirme görevi verilir. Bu yaklaşımın özellikle konuşmadan metne çeviriyi öğrenmede etkili olduğunu ve CoVoST2'deki denetimli SOTA'dan İngilizce çeviriye sıfır atışta daha iyi performans gösterdiğini görüyoruz.

5 Proje planlama Gantt Çizelgesi

Şekil 1: GANTT ÇİZELGESİ



Kaynakça

[1] openai, "Whisper," Nov 17, 2023.

[2] openai, "Research introducing whisper," September 21, 2022.

[3] A. Y. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *CoRR*, vol. abs/1412.5567, 2014.

[4] L. Roberts, "Understanding the mel spectrogram," Mar 6, 2020.

[5] mozilla, "Deepspeech: Mozilla'nın konuşma tanıma motoru,"

[6] mozilla, "Project deepspeech," Dec 10, 2020.

[7] W. Chan, D. S. Park, C. A. Lee, Y. Zhang, Q. V. Le, and M. Norouzi, "Speechstew: Simply mix all available speech recognition data to train one large neural network," *CoRR*, vol. abs/2104.02133, 2021.

[8] D. Galvez, G. Diamos, J. Ciro, J. F. Cerón, K. Achorn, A. Gopi, D. Kanter, M. Lam, M. Mazumder, and V. J. Reddi, "The people's speech: A large-scale diverse english speech recognition dataset for commercial usage," *CoRR*, vol. abs/2111.09344, 2021.

[9] G. Chen, S. Chai, G. Wang, J. Du, W. Zhang, C. Weng, D. Su, D. Povey, J. Trmal, J. Zhang, M. Jin, S. Khudanpur, S. Watanabe, S. Zhao, W. Zou, X. Li, X. Yao, Y. Wang, Y. Wang, Z. You, and Z. Yan, "Gigaspeech: An evolving, multi-domain ASR corpus with 10, 000 hours of transcribed audio," *CoRR*, vol. abs/2106.06909, 2021.

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020.

[11] Y. Zhang, D. S. Park, W. Han, J. Qin, A. Gulati, J. Shor, A. Jansen, Y. Xu, Y. Huang, S. Wang, Z. Zhou, B. Li, M. Ma, W. Chan, J. Yu, Y. Wang, L. Cao, K. C. Sim, B. Ramabhadran, T. N. Sainath, F. Beaufays, Z. Chen, Q. V. Le, C.-C. Chiu, R. Pang, and Y. Wu, "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, p. 1519–1532, Oct. 2022.