Audio Adversarial Examples: Targeted Attacks on Speech-to-Text

Nicholas Carlini, David Wagner
University of California, Berkeley
SYSSEC

PPT by Hanna Kim

Related works

- Attacks on speech recognition
  - Side-Channel → Dolphin Attack, light command
  - Mangled Voice → Hidden Voice Command

Related works

- Dolphin Attack
  - Using the non-linearity of amplifier
  - Playing ultrasonic frequency sound interpreted as baseband frequency sound

  +) Victims can’t hear attack sound!

  -) Need a high-performance device to play the attack sound

- Hidden Voice Command
  - Mangling the voice command so that victims can’t recognize accurate meaning.

  +) Easier to make attack samples

  -) Victims hear the sound of the attack, then feels dangerous.

Related works

- Dolphin Attack:

frequency sound

-

curate meaning.

Introduction

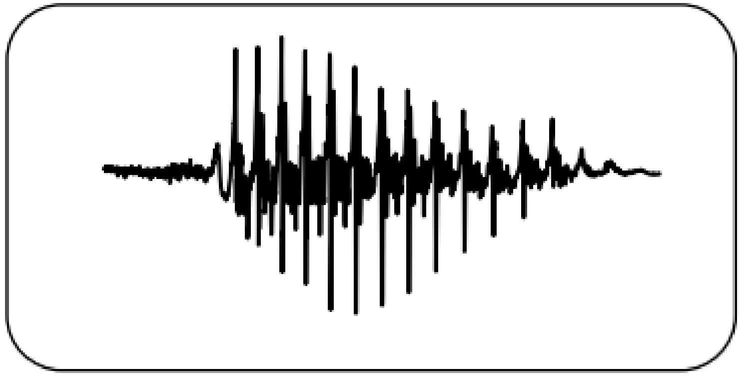
- Attack on speech recognition using adversarial example
- Overcome the limitations
  - Need a high-performance device? No
  - Victims notice the attack? No

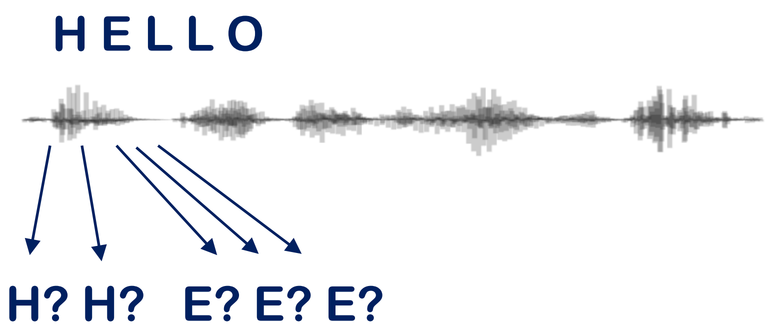
ASR
a technology that converts
speech to text in real time

- Target : DeepSpeech, an automatic speech recognition system

Neural Networks for Automatic Speech Recognition (ASR)

audio sample





Training data

pairs of (audio, text)

of variable lengths

with no alignment

Connectionist Temporal Classification, CTC

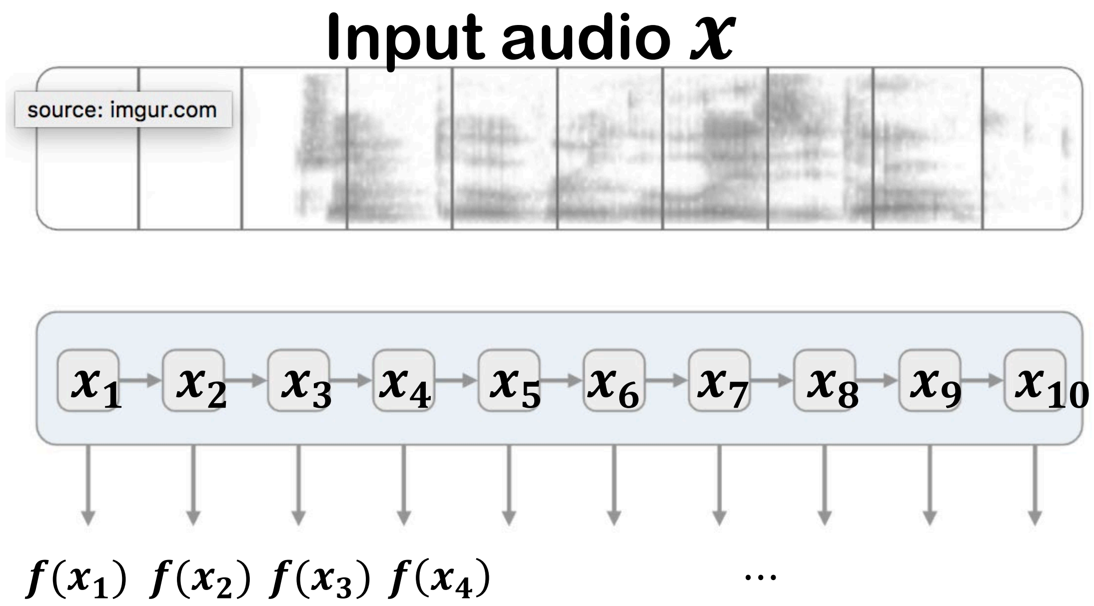
CTC Loss

A differentiable measure of distance
from NN output to the true target phrase.

Training objective

Minimize CTC Loss between
the training audio and corresponding text.

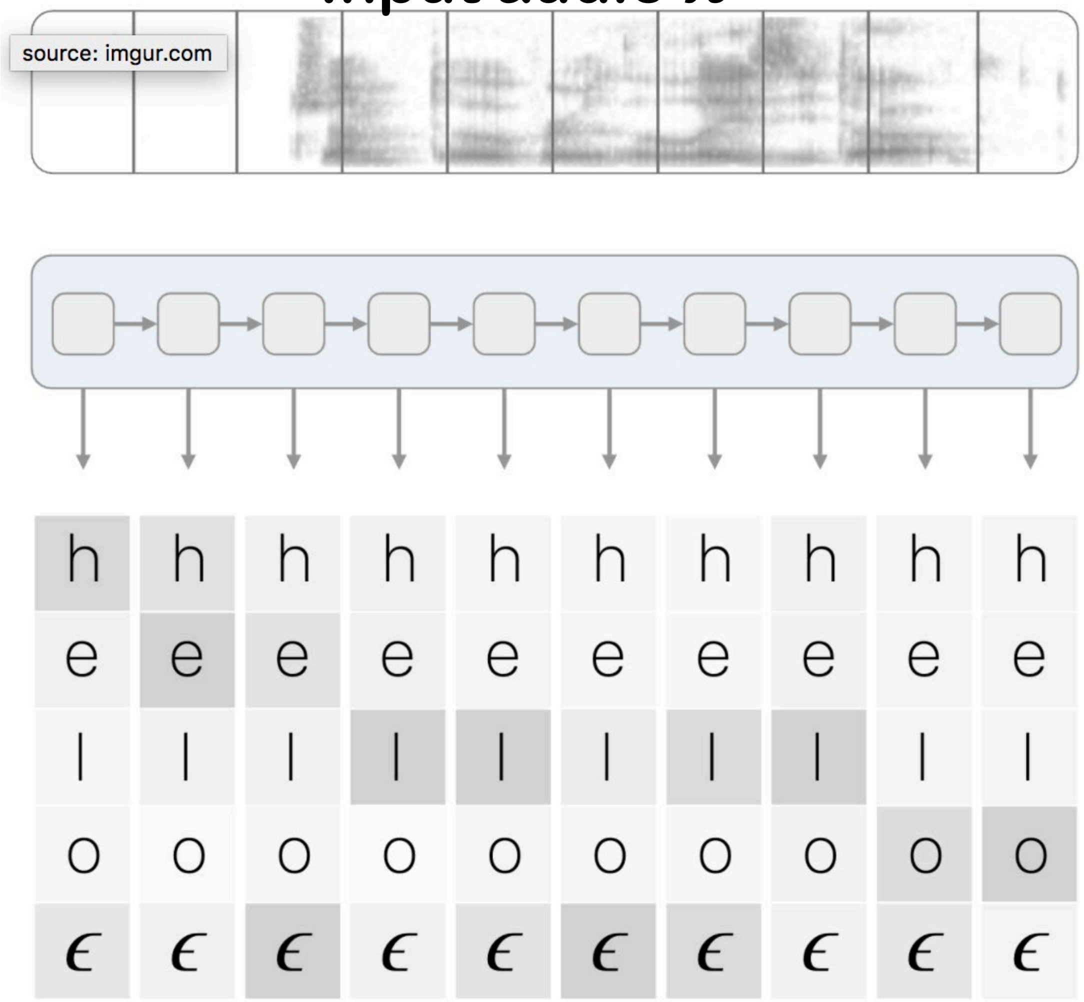
Connectionist Temporal Classification, CTC



a probability distribution over all the possible labels $f(x)$

Connectionist Temporal Classification, CTC

Input audio x



x = Hello

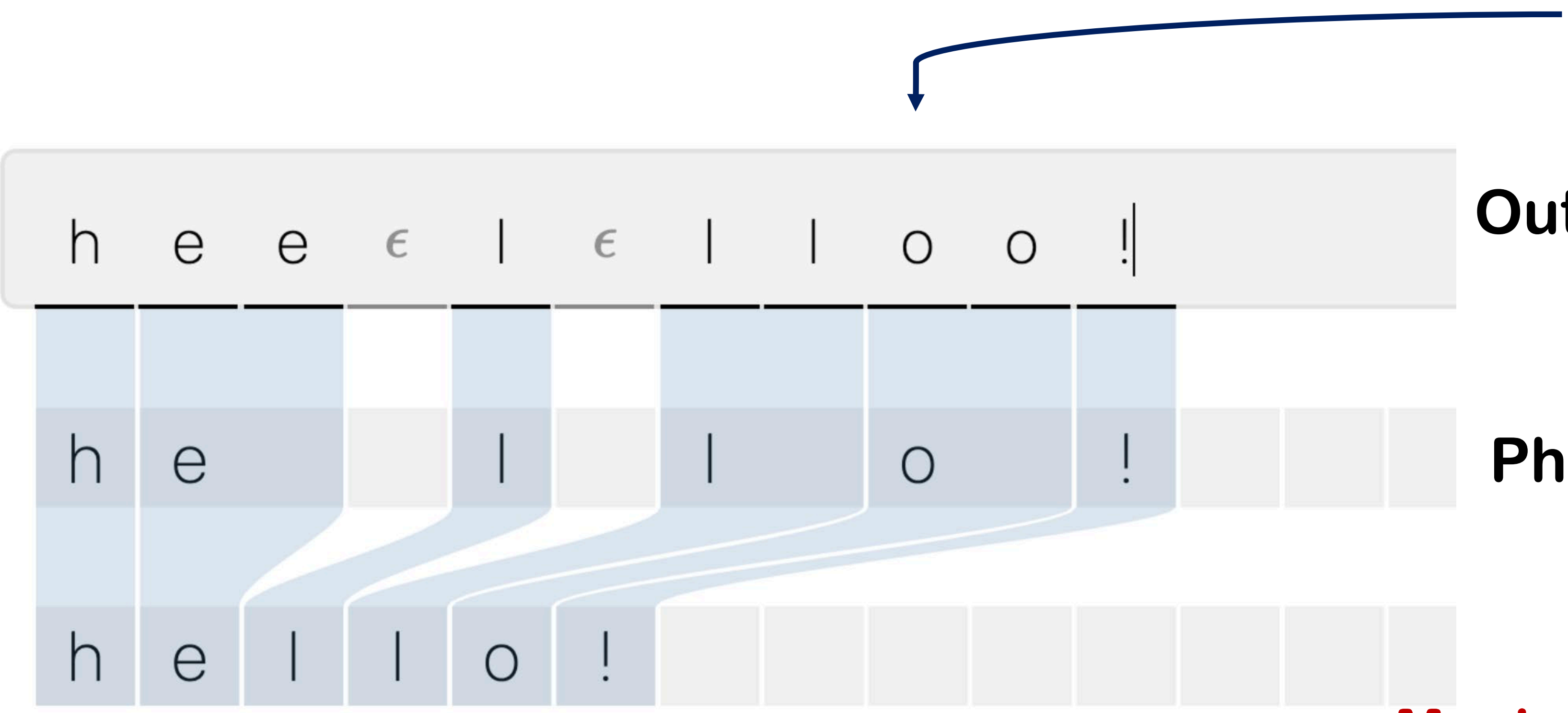
Possible labels
$\{h, e, l, o\} + \{\varepsilon\}$

a probability distribution over all the possible labels

$f(x)$

Connectionist Temporal Classification, CTC

Probability distribution over all the possible labels $f(x)$



Output sequence

Phrase p

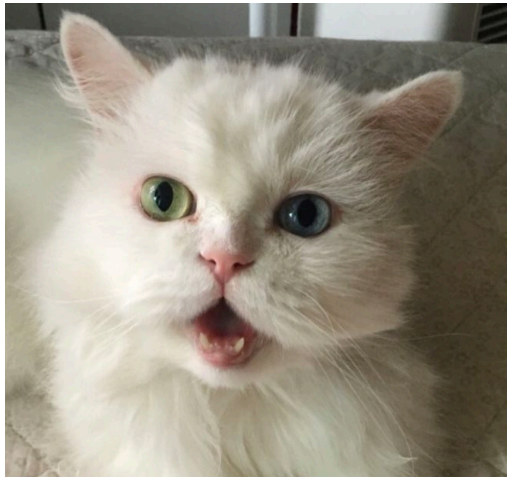
What we want :

Maximize the probability that we will get $f(x)$ that collapses into p

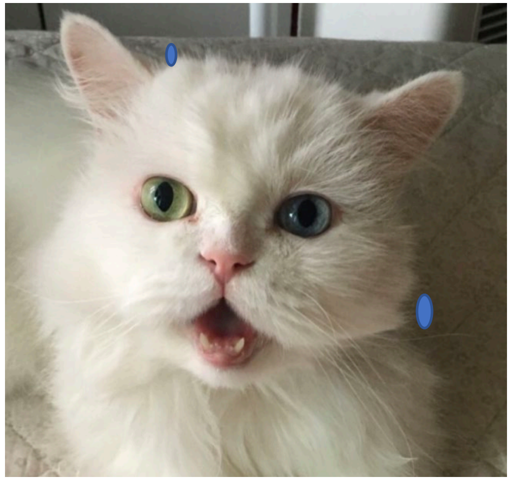
Goal Minimize $CTC - Loss(f(x), p) = -\log Pr(p|f(x))$

Targeted Adversarial Examples

- Given an input x, classified as f(x) = L
- Find the AE x’ close to x
  so that f(x’) = T [for T != L]



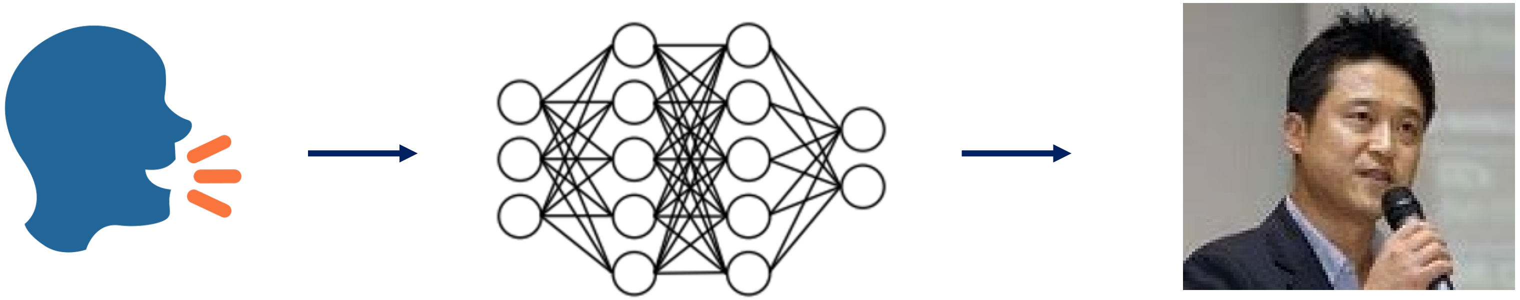
→ cat



→ dog

Targeted Adversarial Examples

- Make the machine recognize my voice as that of Professor Yongdae Kim.



- Make the machine recognize the Professor Yongdae Kim's speech as the target phrase.

This Paper/ Overview

Purpose

Propose an targeted white-box adversarial attack method on speech recognition system.

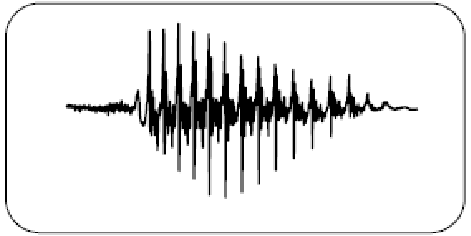
Contribution

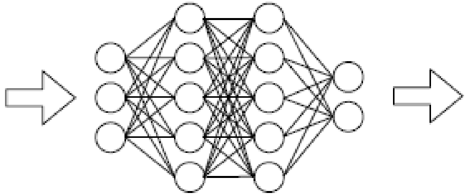
The attack works with 100 % success rate, regardless of the desired transcription or initial source audio sample.

Attack Model

"We have class tomorrow"

Speech





Text

DeepSpeech transcribe as

"We have class tomorrow"

Algorithm

- **Formulation**

  given an original audio sample x,
  find an adversarial example x'

  where

  minimize $d(x, x')$
  such that $f(x') = T$

  ASR model

  Target phrase
  that adversary has choosen

What is the distance metric?

Magnitude of perturbation (in dB) relative to the source audio signal

Algorithm

- **Formulation**

  given x, find x' where

  minimize d(x, x') + l(x', T)

What is the loss function l?

CTC Loss!

- l(x', T) is a loss function measuring how close f(x') is to target T
  - l(x', T) is small, if f(x') = T
  - l(x', T) is large, if f(x') != T

Algorithm

- **Formulation**

  given x, find x' where

  minimize $\|\delta\|_2^{\ 2} + c \cdot \text{CTC-Loss}(x+\delta, T)$

  such that $dB_x(\delta) < \tau$

- **Iterative optimization**

  Repeat until no solution δ can be found by reducing τ.

δ : perturbation, c: regularization coefficient, x : input,

T : target phrase, $\|.\|_2$: 2D p-norm($\sum_{i,j}(\left|\delta_{(i,j)}\right|^p)^{1/p}$), τ: hyper parameter

Evaluation

- Speech to another speech
- Non-speech to speech
- Hide speech

Evaluation (Speech to Speech)

DeepSpeech transcribes it as

“later we simply let life proceed in its own direction toward its own fate”

“the boy looked out at the horizon”

Evaluation (Information Density)

Input waveform is converted into 50 frame/sec

→ theoretical maximum density of audio is 50 character/sec

→ can generate adversarial examples at this maximum rate

→ make short audio clips recognized as longer phrases

Evaluation (Non-Speech to Speech)

DeepSpeech transcribes it as

“speech can be embedded in music”

Evaluation (Speech to Silence)

- You can "Hide" speech by adding adversarial noise.

→DeepSpeech transcribes nothing.

- Targeting silence is easier than targeting a specific phrase.

→Easier to construct adversarial examples
when starting with longer audio waveforms than shorter ones

Conclusion

- Demonstrate targeted audio adversarial examples are effective on automatic speech recognition
- With optimization-based attacks applied end-to-end, turn any audio waveform into any target transcription with 100% success by only adding a slight distortion.

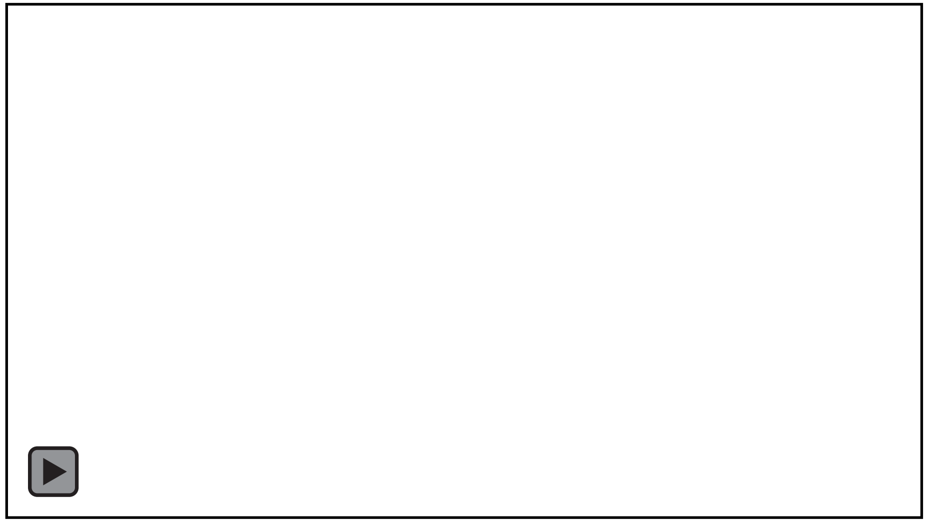
- Open Questions
  - Are audio adversarial examples transferable?
  - Which existing defenses can be applied audio?

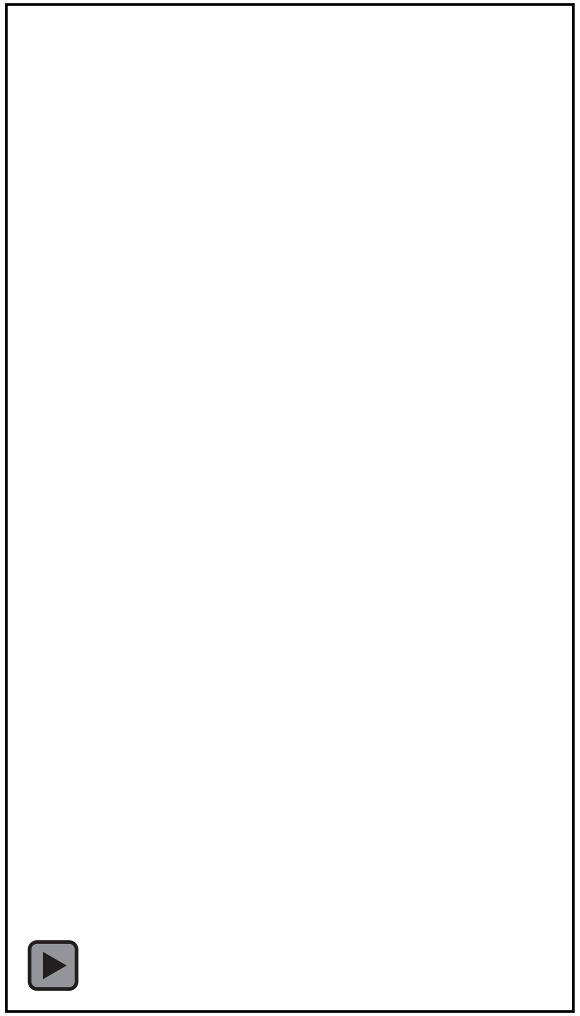
Follow-up works

- Yuan, Xuejing, et al. " **CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition."** (USENIX 2018)
  - Integrate the commands into a song in a way that can be effectively recognized by ASR through the air, in the presence of background noise, while not being detected by a human listener.

- Chen, Yuxuan, et al. "Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices." (USENIX 2020)
  - Enhance a simple local model roughly approximating the target black-box platform with a white-box model that is more advanced yet unrelated to the target

Follow-up works





Questions

Question 1) Over-the-air Attack

From 김광민 (Best question)

Q: The attack cannot be played over-the-air. So, isn't there a lack of practicality in the attack? Has there been any subsequent research on such attacks being performed over-the-air?

- Yuan, Xuejing, et al. " **CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition.**" (USENIX 2018)

Questions

Question 2) Adversarial training

From 정기원 (Best question)

Q: Do this adversarial examples can make model more robust with adversarial training?

Questions

Question 3) Black-box

From 박상민 (Best question)

Q: The attack was measured under the assumption that the attacker knew both the information and parameter values of the model and that classification occurred directly without any noise.

However, it seems difficult to satisfy these assumptions when they are attacked in the real world.

Was there any research after this paper that weakened these assumptions a little?

- Yuan, Xuejing, et al. " **CommanderSong: A Systematic Approach for Practical Adversarial Voice Recognition.**" (USENIX 2018)

Thank you