CS 505: Introduction to Natural Language Processing

Wayne Snyder Boston University

Lecture 22 - Conclusions on GPT; Audio Processing for Deep Learning



GPT-3 (2020)

- o 175B parameters,
- o 96 layers
- 96 masked-attention heads in each layer; used a more efficient alternation of dense and sparse attention patterns
- Input width: 2048 tokens
- Used softmax/temperature and Top-p sampling
- Training for GPT-3 was a scaled-up version of GPT-2, including a larger text corpus:
 - Common Crawl of Web
 - Entire English Wikipedia
 - WebText2 (continuation of WebText from GPT-2)
 - Lots of miscellaneous books, technical manuals, encyclopedias, etc. etc. etc.

GPT-3 also used an AI technique called Mixture of Experts (MoE):

- Multiple pretrained networks are used to divide the problem space into separate task regions;
- Expert results are aggregated by a gating network, which combines one or more outputs into an ensemble output:



Figure 2: Mixture of NMT experts used to make context dependent translation prediction.

Examples:

Multi-language models contain pretrained experts for distinct languages:



Examples:

Multi-modal models contains distinct sub-networks for vision, audio, and text:



In the transformer architecture, the experts can also communicate with each other; typically only some layers are built from MoE:



Figure 1: MoE components

GPT-4 (March 2023)

- o Multimodal
- o 1.8 T parameters
- o 120 layers
- Input size of largest version is 32 K tokens
- o Handles 26 languages
- MoE with 16 experts, each with 111 B parameters

Audio Computing: Physical Basis of Sound

Sound is produced by vibrating objects which produce pressure waves in air, traveling at 343.21 meters/sec (768 mph, or a mile in 4.69 seconds), which are sensed by the ear and interpreted by the brain.

These waves are **longitudinal waves** (the motion is along the direction of travel), as opposed to transverse waves (motion is at right angles to the direction).



Physical Basis of Sound

Time Domain Representation: If we record the atmospheric pressure over time, we get a curve with amplitude in pounds per square inch (psi) on the y axis and time in seconds over the x axis in the shape of the sin (or cos) function:



Acoustic Longitudinal Wave

Properties of (Pure) Sine Waves

Wavelength (λ):

Distance between peaks of a wave (affects pitch -- high or low sounds); measured in meters.

Period (p):

The time between peaks, measured in seconds.

Amplitude (A):

magnitude of the wave above the midpoint (x axis).

Frequency (*f*):

the number of times a wave occurs in a second. Measured in cycles per second or Hertz (Hz) or KiloHertz (kHz).

Some important relationships: (where v = 343.21 m/s)

$$f = v / \lambda$$
 or $\lambda = v / f$

Ex: Middle C = 261.63 Hz λ = 343.21 m/s / (261.63 / s) = 1.31 m





10

Sound Wave Properties: Frequency

Frequency is an absolute measure, and is strongly related to but not absolutely identical to the notion of pitch.

Pitch = perceived frequency of a sound

Pitch is subjective, and inaccurate at extremes of frequency or amplitude, and not measured precisely by the ear throughout the range of hearing. However, it is generally perceived in a log scale.

Human: 20Hz – 20 kHz

Dogs: 67 Hz – 44 kHz

Cats: 55 Hz – 79 kHz

Sensitivity of human ear at various frequencies; curves represent impressions of equal loudness at various frequencies:



Sound Wave Properties: Intensity and Loudness

Sound Intensity Notice that sound waves carry energy. We define the intensity I as the rate at which energy E flows through a unit area A perpendicular to the direction of travel of the wave. Intensity = Power / Area I = P / A = E / (At)

Inverse Square Law, General

Any point source which spreads its influence equally in all directions without a limit to its range will obey the inverse square law. This comes from strictly geometrical considerations. The intensity of the influence at any given radius r is the source strength divided by the area of the sphere. Being strictly geometric in its origin, the inverse square law applies to diverse phenomena. Point sources of gravitational force, electric field, light, sound or radiation obey the inverse square law. It is a subject of continuing debate with a source such as a skunk on top of a flag pole; will its smell drop off according to the inverse square law?



Thus, Intensity is proportional to the square of the amplitude:

 $I = c * A^2$ (c is a constant depending on properties of medium)

Since ear drums tend to be similar in area, often Intensity and Power are used as equivalent terms.

Sound Wave Properties: Intensity and Loudness

Loudness is <u>not</u> simply sound intensity!

Loudness of a sound is measured by the logarithm of the intensity.

The threshold of hearing is at an intensity of 10^{-12} W/m².

Sound intensity level is defined by



log here is to base 10

Digital Audio: Analog vs digital signals

An analog signal continually fluctuates in up and down in the real domain (time is a real number and amplitude is a real number).

A digital signal has a discrete number of amplitudes over a discrete number of time steps, T = 0, 1, 2,

Typically, represented as samples taken at a regular sample rate, e.g.,



16-bit integers

44,100 times a second.



14

Physical Basis of Sound



1 Second

Jean-Bastiste Fourier

Jean-Baptiste Joseph Fourier (1768 – 1830) was a French mathematician and scientist who studied heat transfer in metals (among other things--for example he was the first to explain the Greenhouse Effect). In this investigation he realized the following remarkable fact:



Any periodic function with period P can be constructed by adding together a sequence (possibly infinite) of simple sine waves of various amplitudes and phases, with frequencies that are integer multiples of the fundamental frequency f = 1/P.

A periodic function is one which repeats its values at a period P:



Fourier/Additive Synthesis

Example of Fourier/Additive Synthesis:

A square wave of frequency *f* and amplitude 1.0 can be created from the following infinite *Fourier Series*:

 $(4/\pi)(\sin(2\pi ft) + (1/3)\sin(2\pi 3ft) + (1/5)\sin(2\pi 5ft) +)$

that is:

[(1,1.27,0), (3,0.42,0), (5,0.25,0), (7,0.18,0), (9,0.14,0),]

shown here graphically (duration of 2 sec):







Spectrum: A graph of the amplitude and frequency of the components of a sinusoid is called a **spectrum**; here is a spectrum of a wave consisting of the first ten terms of the Fourier Series for a square wave.



[(1,1.27,0), (3,0.42,0), (5,0.25,0), (7,0.18,0), (9,0.14,0), (11,0.12,0), (13,0.1,0), (15,0.08,0), (17,0.07,0), (19,0.07,0)]

Fourier provided a way of looking at a signal in two equivalent ways, as a graph of amplitude of the signal vs time, the Time Domain, or as a graph of frequencies vs amplitudes (the Frequency Domain):



It can NOT be overemphasized how important these results are for understanding sound (which has three dimensions); these are two different, but completely equivalent ways of viewing the same phenomenon, and the pair of transforms are **lossless** and **very efficient**: O(N log N).



Amplitude vs Time

Frequency Domain: Amplitude vs Frequency

Harmonic Series

Harmonic Series: A series of integer multiples of a particular lowest, fundamental frequency is called a *Harmonic Series* and each component is called a Harmonic or Overtone:

Example (in Hz): 440, 880, 1320, 1760, 2200, 2640, 3080, 3520, 3960,

A periodic signal can always be constructed from such a sequence, although in additive synthesis of musical signals we will not restrict ourselves only to such sequences. Real musical signals always have more complex spectra.

Still, musical sounds are in large part based on such spectra; here is a clarinet playing A 440 Hz:





Spectrogram:

A spectrogram is a 2 D array of spectral data over time; hence it is a 3 D object over frequency, amplitude, and time:





Time

Spectrogram:

A spectrogram is a 2 D array of spectral data over time; hence it is a 3 D object over frequency, amplitude, and time:



Spectrogram:

A spectrogram is a 2 D array of spectral data over time; hence it is a 3 D object over frequency, amplitude, and time:



Spectrograms: Audio software such as Adobe Audition provides for spectrographic views of signals:



Spectrogram:

Viewing 2D data can be done using faux-3D plots:



Spectrogram:

But is more commonly done by "heat-maps" where the amplitude is indicated by greyscale or color:



Spectrogram:

But is more commonly done by "heat-maps" where the amplitude is indicated by greyscale or color:



Spectrogram:

But is more commonly done by "heat-maps" where the amplitude is indicated by greyscale or color:



Spectrogram:

In matplotlib, we have a library function to create spectrograms from a signal directly:

```
specgram(x, NFFT=256, Fs=2, Fc=0, detrend=mlab.detrend_none,
         window=mlab.window_hanning, noverlap=128,
         cmap=None, xextent=None, pad_to=None, sides='default',
         scale_by_freq=None, mode='default', scale='default',
         **kwarqs)
```



1.0

1.5

2.0

2.5

0.5

(spectrum, freqs, t, im) = plt.specgram(X,NFFT=2048, Fs=44100, noverlap=0)

Spectrogram:

By adding a few bells and whistles, we can get log scale and proper axis measurements:



Musical Acoustics: Vocal Tract

Even if you do not play a musical instrument, you use spectra all the time to communicate! The reason you recognize different vowel sounds is that they have different spectra:



MEL Scale

- Humans perceive both loudness and pitch on a log scale;
- For pitch, the relationship between pitch f in Hz, and our human perception m (in Mels) this is called the Mel Scale:

$$m = 2595 \log_{10} \left(1 + rac{f}{700}
ight)$$



MEL Spectrogram

Humans perceive pitch is on a log scale; this is called the Mel Scale:



$$m=2595\log_{10}\left(1+rac{f}{700}
ight)$$

MEL Spectrogram

Therefore, to capture the human experience of sound, we typically use a Mel Spectrogram, where

- Pitch is given in Mels
- Loudness is given in Decibels:

mel_spect = librosa.feature.melspectrogram(y=y, sr=sr, n_fft=2048, hop_length=1024) mel_spect = librosa.power_to_db(spect, ref=np.max) librosa.display.specshow(mel_spect, y_axis='mel', fmax=8000, x_axis='time'); plt.title('Mel Spectrogram');





Human Vocal Signals

Each phoneme in human language has a rather distinct spectrogram:



(Example of spectrogram of human voice: demonstration)