

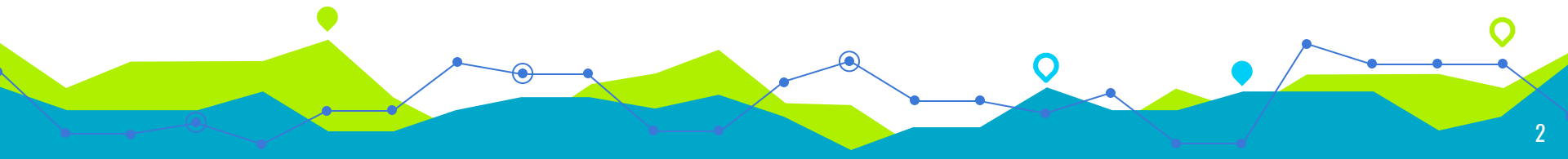
# WIM – What'd I Miss?

Victor Geislinger



# Finding Relevant Info in a Video is Hard

- Videos are informative
- Playlists help contain info
- If organized well, can find relevant info



# Finding Relevant Info in a Video is Hard

- Videos are informative
- Playlists help contain info
- If organized well, can find relevant info
- Titles can only give so much info
- Lots of info to sift through
- No real way to search for related terms/ideas

# Solution: AI

- Can get transcripts (**Whisper**)



# Solution: AI

- Can get transcripts (**Whisper**)
- Can search through related quotes (**BERT/Encoders via transformers**)

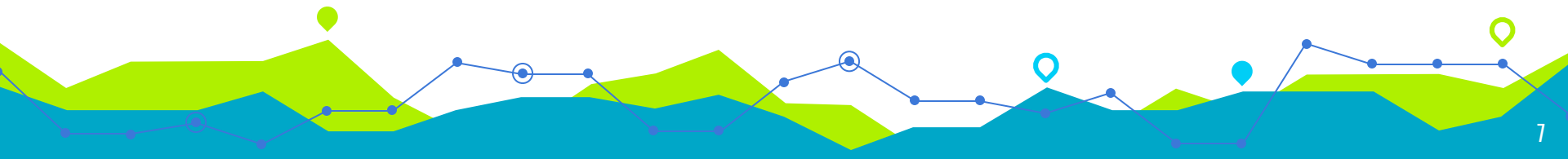
# Solution: AI

- Can get transcripts (**Whisper**)
- Can search through related quotes(**BERT/Encoders via transformers**)
- Can summarize information with generated text (**Anthropic's Claude**)



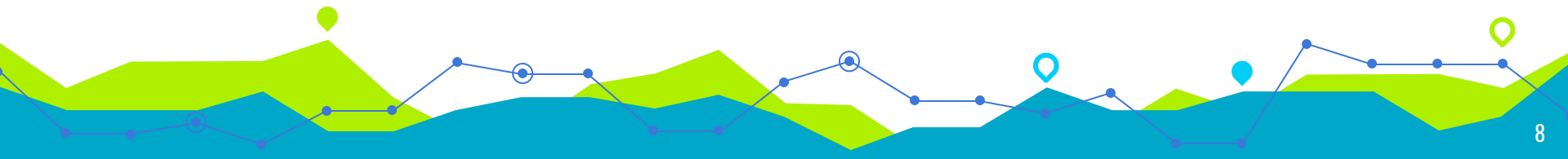
# Solution: AI

WIM



# WIM

Ask pointed questions about a given playlist  
and get back a **summary**, **key points**, and  
related **timestamps** **generated via AI!** 🤖







# Ask pointed questions

## Checkout a Series...

Which Series?

Victor's Deep Learning Lectures

Select all transcripts?

Note: Due to limited demo resources, only up to 50 transcripts can be used

Which episode transcript to search?

- 01-Intro.to.Natur... x
- 02-Big.Data.Intr... x
- 03-Recommend... x
- 04-Intro.to.Neur... x
- 05-Neural.Netwo... x
- 06-Neural.Netwo... x
- 07-How.to.Avoid... x
- 08-Optimizing.Yo... x
- 09-Intro.to.Conv... x
- 10-Visualizing.A... x
- 11-Transfer.Lear... x
- 12-Style.Transfer... x
- 13-Embeddings... x
- 14-Extensions.to... x
- 14-Recurrent.Ne... x

## Ask a Question or Write a Topic

Ask a question or state a topic of interest

What do I need to know about neural networks for machine learning?

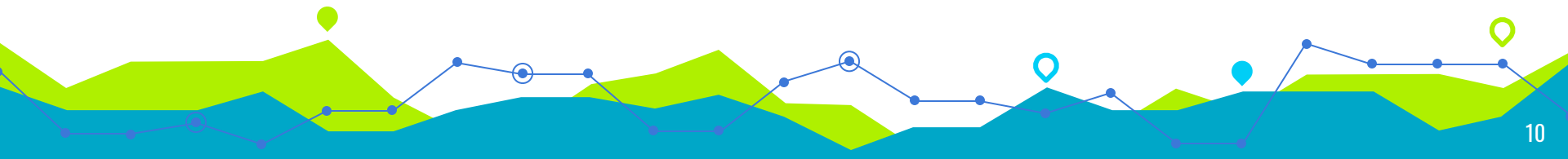
66/500



## summary

### Overall Summary

Neural networks are powerful models that can learn complex functions but require many design choices and hyperparameters to achieve good performance.



# WIM

## key points

### Key Point

Neural networks are complex models that can approximate any function given enough parameters and data.

### Key Point

Neural networks rely on many hyperparameters to work well like activation functions, learning rates, optimizers, and regularization



# WIM

## timestamps

### Key Point

Neural networks are complex models that can approximate any function given enough parameters and data.

### Quotes & Timestamped Links

*"But what's cool about that, that will actually learn this context right here, this embedding layer..."*

<https://youtu.be/np-hv-BkXYA?t=818>

The screenshot shows a YouTube video player with the title "Embeddings for NLP - M4S45 [2019-06-25]". The video content displays a slide with the following sections:

- Male-Female**      **Verb-Tense**      **Country-Capital**
- Convert words into a vector space
  - Mathematical object
  - It's an *abstract* object
  - DISTRIBUTIONAL HYPOTHESIS** [https://en.wikipedia.org/wiki/Distributional\\_hypothesis](https://en.wikipedia.org/wiki/Distributional_hypothesis)
- 2 Resources**
  - Kaggle Tutorial: <https://www.kaggle.com/learn/embeddings>
  - Google's Embedding Class Course: <https://developers.google.com/machine-learning/crash-course/embeddings>
- 3 Word2Vec** [...]
- 4 GloVe - Global Vectors for Word Representation** [...]

The slide also features a diagram of a neural network architecture for word embeddings, with handwritten notes:  $\begin{bmatrix} \text{male} \\ \text{female} \end{bmatrix}$  and  $\text{male} = \text{female} + \text{play} - \text{golf}$ . A red play button icon is overlaid on the diagram.

# WIM

# timestamps

## Key Point

Neural networks rely on many hyperparameters to work well like activation functions, learning rates, optimizers, and regularization

## Quotes & Timestamped Links

"So this brings up the concept of thrashing..." <https://youtu.be/b22dEJbc8b0?t=429>

Big Data Intro & Parallel/Distributed Computing - ...

Copy link

Labels: Numbers Every Programmer Should Know

Branch mispredict: 11.000ns = 37ns

L1 cache reference: 1.000ns = 1µs

L2 cache reference: 1.000ns = 1µs

SSD random read: 11.000ns = 37ns

disk seek: 0.000,000ns = 0ns

Max 1,000,000 bytes sequentially from memory: 1.000,000ns = 1µs

Max 1,000,000 bytes sequentially from disk: 0.000,000ns = 0ns

bucket rounding CA (beta): 0.000,000ns = 0ns

"Sigmoid tends to be not very great at this..." <https://youtu.be/SD8C1bl-hxQ?t=268>

Optimizing Your Neural Network - M4S42 [2019-0...]

Copy link

1 harder, better, faster, stronger

We can optimize our model to run faster and better

1.1 Some useful resources on optimization

- Keras documentation & cheatsheet: <https://keras.io/guides/faq/>
- Excellent blog post on optimization: <https://towardsdatascience.com/optimization-techniques-for-deep-learning-7c4e1e1e1e1e>

2. Vanishing Gradient Problem

When our problems are too small we can run into a few issues when training NN

- Slow training
- Loss reduction problem

2.1 Activation functions (from before)

A simple way to manage the gradient is to change the activation function that produces different gradients

2.2 Random Starts

To avoid local minima, start if you just randomly choose random numbers on the cost function

Go 1 back to get the values inside multiple times, so I can sort of get some different weights and can help avoid getting into a local minimum

2.3 Momentum

3. Speeding Up

# WIM

Ask pointed questions about a given playlist and get back a **summary**, **key points**, and related **timestamps generated via AI!** 🤖



# How Does it Work?

- Transcripts generated via **Whisper**
  - <https://github.com/MrGeislinger/whisper-extract>
  - Technically can be created with any other tool

# How Does it Work?

- Transcripts' sentences compared via embeddings (**BERT**)
  - BERT or other encoding transformer
  - Selects a subset of sentences from given transcripts

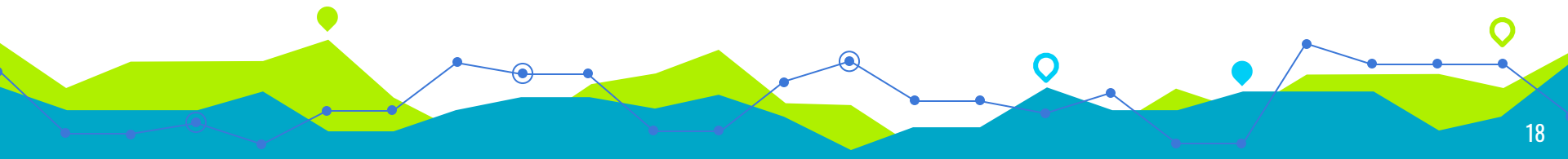


# How Does it Work?

- Subset of transcripts' & user's question fed to AI (**Anthropic's Claude**)
  - Generated summaries and key points
  - Model chooses relevant quotes
  - Quotes cross-checked with subset to provide links w/timestamps

# The Future?

- Deployed with more resources
  - Currently Streamlit deployment limits RAM to 1GB
  - Allows more comparisons (more transcripts)



# The Future?

- Deployed with more resources
  - Currently Streamlit deployment limits RAM to 1GB
  - Allows more comparisons (more transcripts)
- Use vector database (such as Chroma)
  - Currently local
  - Vector database will speed up comparisons on a database server

# The Future?

- Deployed with more resources
  - Currently Streamlit deployment limits RAM to 1GB
  - Allows more comparisons (more transcripts)
- Use vector database (such as Chroma)
  - Currently local
  - Vector database will speed up comparisons on a database server
- Sentence embeddings treated differently between question and transcripts
  - Fine-tuning of model to create embeddings

# The Future?

- Deployed with more resources
  - Currently Streamlit deployment limits RAM to 1GB
  - Allows more comparisons (more transcripts)
- Use vector database (such as Chroma)
  - Currently local
  - Vector database will speed up comparisons on a database server
- Sentence embeddings treated differently between question and transcripts
  - Fine-tuning of model to create embeddings
- Adjust prompting (always different for different LLMs!)

# Demo Time

WIM

[wim.victorsothervector.com](http://wim.victorsothervector.com)

