Human Computer Interaction (In English)

Prof. Rajiv Ratn Shah

Department of Computer Science and Engineering
Institute IIT Madras

AI-Powered Tools for Content Generation and Analysis: Firefly, Audino

Hey, everyone. I'm Aman. I'm working as a research assistant under Dr. Rajiv Ratan Shah. So my work focuses on NLP AI, audio AI, and web development. So today I'll show you a few of the examples of the popular domain, such as text to image generation, text to video generation, and audio processing here.

So let's get dive into it. The tools which I'll show you today are not specifically good in their domains, but are very great and doing very well. So let's get started. First of all, what image to text means? Image to text specifically means that you give a model a specific prompt and it will generate an image out of it.

How it works? The model has been traded into billions of images and marked as text by human evaluation. So if you give a prompt, the model will look into the prompt and can feel the context, can understand the context, and it will generate an image. Then it will render the image, then it will go to the second word, then it will render it again, and then the third, and so on. What text to video means? It is actually similar to text to image. What video actually is that it's a combination of images generated by anything, like recorded by a camera, generated by the model, followed by a background music.

What audio processing is? So audio processing involves analyzing and transforming audio data. such as speech to text. If I'm recording this video, my spoken speech have a textual form. If we pass the recorded video or the audio to these kind of models, it will generate the text form of the spoken speech. Then the text has been passed to the LLMs.

The LLM actually summarize like analyze the text. It will summarize that it can like generate a context out of it you can talk to that context you can ask the questions and you can do many more so let's get to the demo part first of all text to image so we will be covering adobe firefly it's actually not open source but adobe are giving few of the credits to the users So yeah, first of all, I'll write a simple prompt. Let's see what the model will generate out of it. A magical forest with sparkling mist and mushrooms in the ground. So the model have given us four of the outputs.

As you can see, it's a forest. The mushrooms are being grown. It's actually a sparkling mist, which is glowing and sunshine is coming. So what do we understand from this? The model have actually taken each of the words and created similar image, adding frames to the image, adding the objects to the image, then rendering it. then it is giving us the output.

It is also giving us four of the options. You can select any one. What happens if we change the prompt? So let's add a water stream flowing from the ground. Let's see if the model can actually generate it or not. As you can see, the model have added a water stream in the forest and the mushrooms are being grown left and right of the water stream.

You can also play with the effects of what Adobe Firefly is giving. As you can see, it's giving a painting effect if you want to make these images as a painting. or a layered paper, a digital art, a hyper realistic or so more. You can also give a reference image if you want to generate a image based on the reference or something you want. So let's see what painting a magical forest with sparkling mist in a painting style looks like.

So this is how the model have generated it. So this is the text to image model generation part. You can also try different tools such as Canva, stability.ai, and so on. I have given the link in the PPT.

You can just check it later. So let's dive into the text to video generation part. For today's demo, I am actually using this InVideo tool, which is a very popular tool. And it is also not free, but it is giving a few of the credits to the users. Go and check that.

Use it. Play with it. For saving the time, I have already made a video using a prompt. As you can see, what this video means, like what prompt I have given. I am shooting a musical video. For the cover of that video, I want a peaceful beach scene.

So this is how it looks like. In video, I have generated this. Beach, waves, and a sunny day. Few people are having their picnic and having the lunch. One person is playing the guitar.

So you can, I'm not sure if you are able to listen to the music or not, but the music is very peaceful and calm. So how do NVIDIA actually generate this? The NVIDIA have taken the prompt and it have rendered frame by frame each images and then combined it with the background music and it had given output as a video. You can go and check how it actually how cool it is. So I have also mentioned few other tools which are popular in the domain for text to video generation. I've also mentioned the link you can go and check.

Let's go to the audio processing part. That is Audino. It's an open source tool. Basically, it takes out the insights from the audios and videos. Actually, the tool specializes in Indian languages.

Hindi, English, Marathi, Tamil, Telugu, Bengali, and so on. So what does it do? Arduino actually is divided into four of the verticals. The first vertical is Content AI, the second vertical is Meeting AI, the third is Call Recording and the fourth is Hate Speech. What does it mean? I'll just go one by one. Let's get started with the Content AI.

Suppose you have a very big video, a lecture video. promotional video marketing video upload launch video and so on you don't have the time to actually see what is happening in that four or five hours four or five hours of video what you have to do you just have to upload it using a youtube link using the local or so on what audino model will do it will give you a summary out of it. What summary I mean like it will generate the spoken speech into the text then it will pass that spoken speech into the LLM model. The LLM model then summarize the particular topic. As you can see here it's the features being discussed of Samsung Galaxy S24 Ultra What is the summary of this particular video? Video showcase 17 Samsung Galaxy tips and tricks, including customized notification and security features like lockdown mode and data protection.

So this is a summary which the model have generated using this video. This is the chapter chapters generated by the model by chapters. What I mean, which particular which important topic is being discussed at what time? Suppose you have a long video and you want to go to a specific topic list. I wanted to know in a product launch video, I wanted to know that what was the price of iphone what are the key features what are the differentiation of new iphone to the previous iphone so no need to actually go to each and every part of the video you can just go to a specific topic the time has been mentioned you can click over here and you will just go to that topic again you can also ask a particular type of question if you don't have the time for that As you can see, I have asked one of the question, which device is the video talking about? What is the device which is being used in the particular video? The video is talking about the Samsung Galaxy device, the model I've given that. Based on the context what video is all about.

Similarly, let's go to the meeting AI. I have uploaded a meeting worth 30 minutes approximate. in which few of the items are being discussed. The Ordino's model have generated it and categorized it into these things. What is the agenda of the meeting being discussed in the meeting? What are the key discussion action item taken by the executives, managers, CEOs who are in the meeting? What are the next steps? Meeting sentiments, meeting intents and so on.

As you can see, I have also asked one of the question from the model. It have given me the answer based on the video. Can you list few items for me? From the conversation what model have generated. From the conversation provided, here are few key items related to adding new dependencies and accessing their maturity security. Using the latest version, it is preferred to use the latest version of dependencies considering stability and security risk.

So these are the list of items which the model have generated based on the topics being discussed in the video. Similarly, what call recording is, it is a very useful tool for each and every company. So suppose like every company is having a call center where customer care executive are taking the calls and resolving the customer's query problems, complain and so on. But how the companies are making sure that the customer care executive have solve the problem or not like explain each and every um features of the particular product how the companies make sure that so right now what companies are doing like they are hiring the human evaluators for the same the human evaluators listen to the calls then they pass the judgment but that is too much manpower consuming and a very slow process what you need to do you can you you just have to upload the call recording to the odd news platform It will actually mark it as what was the client's query, client task. What are the actions taken by the customer executive? What is the solution provided by the customer executive? Action items, ensure proper call opening, empathize with the customer and confirm account details.

These are the action items taken by the customer executive. Is there any complaint from the customer or not? What was the intent of the meeting? So these are the things which a customer care executive and a customer is having a conversation about. I've also asked one of the question, what was the client asking? The client was asking about $33 charge in July 2020. So this call is all about resolving an overcharged money to the customer from a company. So the customer have called to the customer care executive and the customer care executive have resolved his problem.

So this is the hate speech AI. Audio-nose hate speech AI basically tags a particular audio video into several categories such as Criminality is being discussed in audio. Any illegal stuff is being discussed in an audio or not. So Audino have categorized into 16 or 17 parameters such as drug abuse, illegal activities, violence, child abuse and so on. So this is a bank of abroad video. As you can see, you can just check it out in this particular video.

Controversiality has been discussed. So the Ordino's model have marked this. The start time is this, the end time is this. Controversiality has been discussed. This is the context of the controversy. And as you can see, these are the parameters which the Ordino have detected of hate speech in this particular video.

You can also chat with the particular same thing. And as I told you before, Audino is actually helpful for analyzing the Indian dialects and languages audios. So you can definitely check this out. Thank you so much for your time.