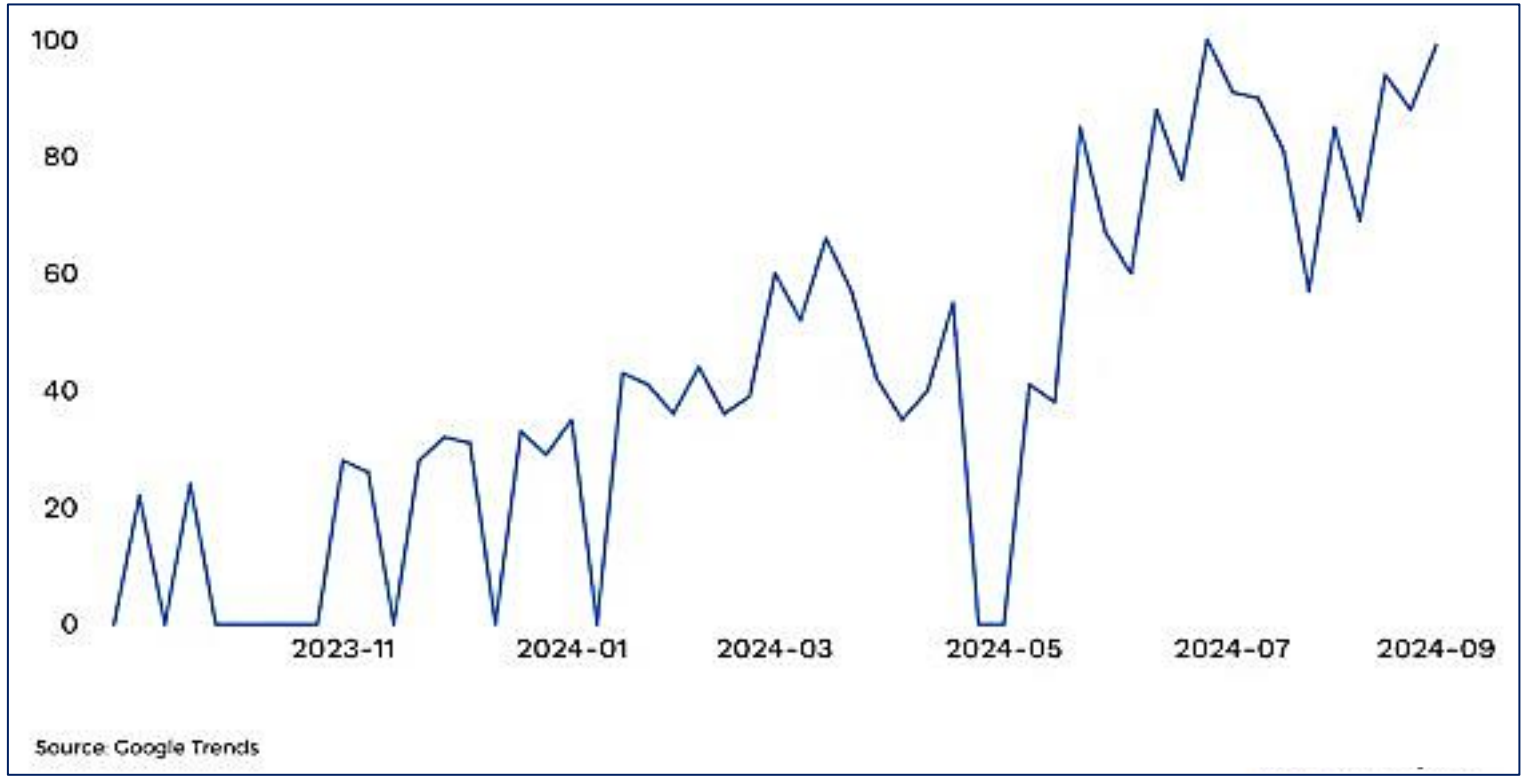
Large Multimodal Models

Achmad Benny Mutiara

September, 2024

Pengantar

- **Model Bahasa Besar** (**LLM**), seperti GPT-4, unggul dalam tugas berbasis teks tetapi ***tidak dapat*** memproses input ***non-teks*** seperti ucapan atau video. Sebaliknya, **Model Multimodal Besar** (**LMM**) muncul untuk menangani berbagai jenis data, seperti teks, gambar, dan audio.
- Namun, kompleksitas dan persyaratan datanya menimbulkan tantangan potensial. Inovasi dalam penelitian AI bertujuan untuk mengatasi tantangan ini.



Gambar 1: Minat untuk model multimodal besar di Google Trends lebih dari 12 bulan terakhir

LMM Terkemuka

- LMM dengan tanggal peluncurannya:
  - OpenAI GPT-4o (“o” untuk “omni”): Mei, 2024.
  - Gemini 1.5 dai Alphabet: Februari, 2024.
  - Qwen-VL-Plus dan Qwen-VL-Max adalah LMM open source terkemuka yang dirilis oleh Alibaba Cloud: Januari, 2024

Model Multimodal Besar sumber terbuka

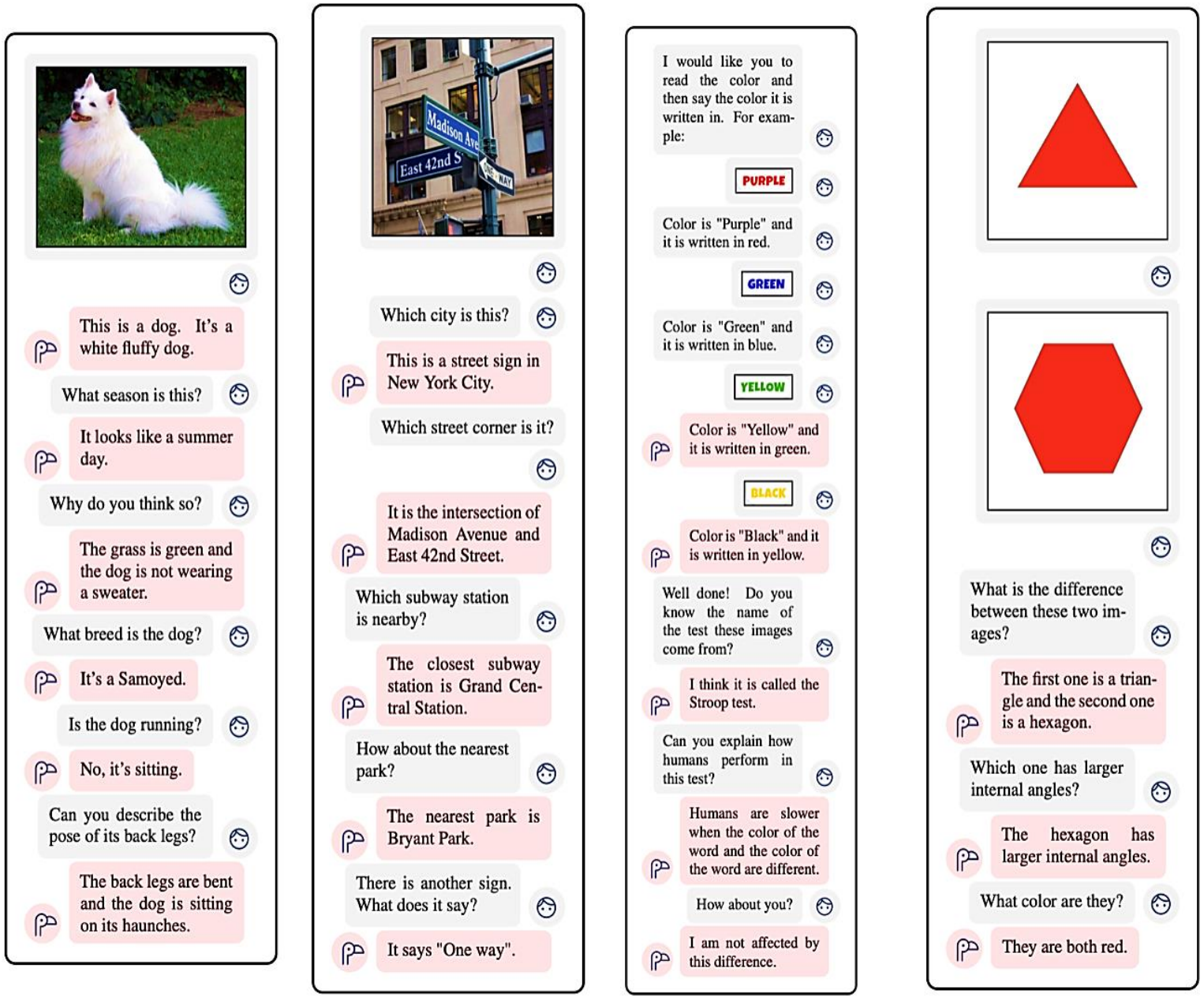
CLIP (Contrastive Language–Image Pretraining) dari OpenAI

- CLIP dirancang untuk memahami gambar dalam konteks bahasa alami. Ia dapat melakukan tugas-tugas seperti klasifikasi gambar zero-shot, di mana ia dapat secara akurat mengklasifikasikan gambar bahkan dalam kategori yang belum dilatih secara eksplisit, dengan memahami deskripsi teks

Flamingo dari DeepMind

- Flamingo dirancang untuk memanfaatkan kekuatan pemahaman bahasa dan visual, membuatnya mampu melakukan tugas-tugas yang memerlukan penafsiran dan integrasi informasi dari teks dan gambar

Model Multi-modal Besar sumber terbuka



Gambar 2: Contoh yang diambil dari Chip Huyen, *"Multimodality and Large Multimodal Models (LMMs)."*

Pengertian Model Multimodal Besar (LMM)?

- **Model multimodal besar** adalah jenis model kecerdasan buatan canggih yang dapat memproses dan memahami berbagai jenis modalitas data.
  - **Data multimodal** ini dapat mencakup teks, gambar, audio, video, dan berpotensi lainnya.
  - **Fitur utama** dari model multimodal adalah kemampuannya untuk mengintegrasikan dan menafsirkan informasi dari sumber data yang berbeda ini, seringkali secara bersamaan.
- Model dapat dipahami sebagai versi yang lebih canggih dari **model bahasa besar** (**LLM**) yang dapat bekerja **tidak hanya** pada teks tetapi juga beragam tipe data.
- Selain itu, **output model bahasa multimodal** ditargetkan **tidak hanya** tekstual tetapi juga visual, pendengaran, dll.
- **Model bahasa multimodal** dianggap sebagai langkah selanjutnya menuju **Artificial General Intelligence (AGI)**

Perbedaan antara LMM dan LLM

Aspek	Model Multimodal Besar (LMM)	Model Bahasa Besar (LLM)
Modalitas Data	Dapat menangani dan memahami berbagai jenis data, seperti teks, gambar, audio, video, dan terkadang pembacaan sensor.	Hanya berfokus pada teks. Tidak menangani jenis data lain seperti gambar atau audio.
Kemampuan Integrasi	Pandai menggabungkan dan memahami berbagai jenis data sekaligus.	Hanya berfungsi dengan teks dan tidak menggabungkannya dengan jenis data lainnya.
Aplikasi dan Tugas	Digunakan untuk tugas yang perlu memahami beberapa tipe data secara Bersamaan. Misalnya, mengana-lisis artikel berita dengan foto dan video terkait.	Digunakan untuk tugas berbasis teks seperti menulis, menerjemahkan, menjawab pertanyaan, meringkas, dan membuat konten.
Pengumpulan dan Persiapan Data	Melibatkan pengumpulan campuran teks, gambar, audio, dan video. Proses ini kompleks dan membutuhkan anotasi dan normalisasi yang cermat untuk mengintegrasikan data.	Melibatkan pengumpulan teks dalam jumlah besar dari buku, situs web, dan sumber lain, dengan fokus pada berbagai bahasa.

Perbedaan antara LMM dan LLM

Aspek	Model Multimodal Besar (LMM)	Model Bahasa Besar (LLM)
Desain Arsitektur Model	Menggunakan berbagai jenis jaringan saraf, seperti CNN untuk gambar dan transformator untuk teks, dan menggabungkannya untuk menangani berbagai jenis data.	Biasanya menggunakan arsitektur transformator yang dirancang khusus untuk memproses dan menghasilkan teks.
Pre-Training	Pra-pelatihan menggunakan beragam data, termasuk teks, gambar, dan video. Belajar menghubungkan teks dengan data lain, seperti membuat teks untuk gambar.	Pra-pelatihan tentang korpora teks besar menggunakan metode seperti memprediksi kata-kata yang hilang dalam kalimat.
Fine-Tuning	Fine-tuning melibatkan pekerjaan dengan kumpulan data khusus untuk setiap jenis data dan mempelajari bagaimana berbagai jenis data berhubungan satu sama lain.	Fine-tuning menggunakan kumpulan data teks tertentu yang disesuaikan dengan tugas tertentu seperti menjawab pertanyaan atau menerjemahkan bahasa.
Evaluasi dan Iterasi	Dievaluasi berdasarkan penanganan berbagai jenis data, termasuk seberapa baik ia mengenali gambar, memproses audio, dan mengintegrasikan informasi di berbagai jenis.	Dievaluasi berdasarkan kinerja dalam memahami dan menghasilkan teks, dengan fokus pada kefasihan, koherensi, dan relevansi.

Perbedaan antara LMM dan LLM

1. Modalitas Data

- **LMM**:
  - Model dirancang untuk memahami dan memproses berbagai jenis input data, atau modalitas.
  - Termasuk teks, gambar, audio, video, dan terkadang jenis data lainnya seperti data sensorik.
  - Kemampuan utama **LMM** adalah kemampuannya untuk mengintegrasikan dan memahami format data yang berbeda ini, seringkali secara bersamaan.
- **LLM**:
  - Model-model ini mengkhususkan diri dalam memproses dan menghasilkan data tekstual.
  - Model dilatih terutama pada korpora teks yang besar dan mahir memahami dan menghasilkan bahasa manusia dalam berbagai konteks.
  - Model tidak secara inheren memproses data non-tekstual seperti gambar atau audio.

Modalitas data Model Multimodal Besar (LMM)

Teks	Ini termasuk segala bentuk konten tertulis, seperti buku, artikel, halaman web, dan postingan media sosial. Model ini dapat memahami, menafsirkan, dan menghasilkan konten tekstual, termasuk tugas pemrosesan bahasa alami seperti terjemahan, peringkasan, dan menjawab pertanyaan.
Gambar	Model-model ini dapat menganalisis dan menghasilkan data visual. Ini termasuk memahami konten dan konteks foto, ilustrasi, dan representasi grafis lainnya. Tugas seperti klasifikasi gambar, deteksi objek, dan bahkan membuat gambar berdasarkan deskripsi tekstual termasuk dalam kategori ini.
Audio	Mencakup rekaman suara, musik, dan bahasa lisan. Model dapat dilatih untuk mengenali ucapan, musik, suara sekitar, dan input pendengaran lainnya. Mereka dapat mentranskripsikan ucapan, memahami perintah yang diucapkan, dan bahkan menghasilkan ucapan atau musik sintetis.
Video	Menggabungkan elemen visual dan pendengaran, pemrosesan video melibatkan pemahaman gambar bergerak dan suara yang menyertainya. Ini dapat mencakup menganalisis konten video, mengenali tindakan atau peristiwa dalam video, dan membuat klip video.

Perbedaan antara LMM dan LLM

2. Aplikasi dan Tugas

- **LMM**:
  - Karena sifatnya yang multimodal, model ini dapat diterapkan pada tugas-tugas yang memerlukan pemahaman dan integrasi informasi di berbagai jenis data.
  - Misalnya, **LMM** dapat menganalisis artikel berita (teks), foto (gambar) yang menyertainya, dan klip video terkait untuk mendapatkan pemahaman yang komprehensif.
- **LLM**:
  - Aplikasinya berpusat pada tugas-tugas yang melibatkan teks, seperti menulis artikel, menerjemahkan bahasa, menjawab pertanyaan, meringkas dokumen, dan membuat konten berbasis teks.
  - Model **tidak secara inheren** memproses data non-tekstual seperti gambar atau audio.

Perbedaan antara LMM dan LLM

3. Pengumpulan dan Persiapan Data

- **LLM**:
  - Terutama berfokus pada data tekstual.
  - Pengumpulan data melibatkan pengumpulan korpus teks yang luas dari buku, situs web, dan sumber tertulis lainnya.
  - Penekanannya adalah pada keragaman dan keluasan linguistik.
- **LMM**:
  - Selain data tekstual, model ini juga memerlukan gambar, audio, video, dan kemungkinan jenis data lainnya seperti data sensorik.
  - Pengumpulan data lebih kompleks, karena tidak hanya melibatkan berbagai konten tetapi juga format dan modalitas yang berbeda.
  - Anotasi dan normalisasi data sangat penting dalam **LMM** untuk menyelaraskan tipe data yang berbeda ini secara bermakna.

Perbedaan antara LMM dan LLM

4. Desain Arsitektur Model

- **LLM**:
  - Biasanya menggunakan arsitektur seperti transformator yang cocok untuk memproses data berurutan (teks).
  - Fokusnya adalah memahami dan menghasilkan bahasa manusia.
- **LMM**:
  - Arsitektur **LMM** lebih kompleks, karena perlu mengintegrasikan berbagai jenis input data.
  - Sering melibatkan kombinasi jenis jaringan saraf, seperti CNN untuk gambar dan RNN atau transformator untuk teks, bersama dengan mekanisme untuk memadukan modalitas ini secara efektif.

Perbedaan antara LMM dan LLM

5. Pre-Training

- **LLM**:
  - Pra-pelatihan melibatkan penggunaan korpora teks besar.
  - Teknik seperti pemodelan bahasa bertopeng, di mana model memprediksi kata-kata yang hilang dalam kalimat, adalah hal yang umum.
- **LMM**:
  - Pra-pelatihan lebih beragam, karena tidak hanya melibatkan teks tetapi juga modalitas lainnya.
  - Model mungkin belajar untuk menghubungkan teks dengan gambar (misalnya, teks gambar) atau memahami urutan dalam video.

Perbedaan antara LMM dan LLM

6. Fine-Tuning

- **LLM**:
  - **Fine-tuning** dilakukan dengan menggunakan kumpulan data teks yang lebih khusus, sering disesuaikan dengan tugas-tugas tertentu seperti menjawab pertanyaan atau terjemahan.
- **LMM**:
  - **Fine-tuning** tidak hanya melibatkan kumpulan data khusus untuk setiap modalitas tetapi juga kumpulan data yang membantu model mempelajari hubungan lintas modal.
  - Penyesuaian khusus tugas dalam **LMM** lebih kompleks karena beragam tugas yang dirancang untuknya.

Perbedaan antara LMM dan LLM

7. Evaluasi dan Iterasi

- **LLM**:
  - Metrik evaluasi difokuskan pada pemahaman bahasa dan tugas pembuatan, seperti kefasihan, koherensi, dan relevansi.
- **LMM**:
  - Model dievaluasi pada rentang metrik yang lebih luas, karena model harus mahir dalam berbagai domain.
  - Termasuk akurasi pengenalan gambar, kualitas pemrosesan audio, dan kemampuan model untuk mengintegrasikan informasi lintas modalitas.

Melatih Model Multimodal Besar (LMM)

- **Model multimodal besar** mirip dengan model bahasa besar dalam proses pelatihan, desain, dan operasi. Model menggunakan arsitektur transformator dan strategi pelatihan yang sama. **Model multimodal besar** dilatih pada:
  - Data teks
  - Jutaan atau miliaran gambar dengan deskripsi teks
  - Klip video
  - Cuplikan audio
  - Data input lainnya seperti kode

Melatih Model Multimodal Besar (LMM)

- Pelatihan ini melibatkan pembelajaran simultan dari beberapa modalitas data, memungkinkan model untuk:
  - Mengenali foto hewan-hewan
  - Mengidentifikasi woof dalam klip audio
  - Memahami konsep dan detail sensorik di luar teks
- Setelah proses pelatihan, model mungkin memasukkan stereotip yang tidak sehat dan ide-ide beracun. Untuk menyempurnakannya, teknik seperti:
  - Pembelajaran penguatan dengan umpan balik manusia (RLHF), Model AI pengawasan, Red teaming (menguji kekokohan model) dapat digunakan.
- Selain itu, alat tata kelola AI dan platform AI yang bertanggung jawab dapat memastikan kepatuhan AI dan pengoptimalan inventaris AI, membantu mencegah bias AI dan dilema etika lainnya.

Sistem multimodal fungsional

- Tujuannya adalah untuk mengembangkan sistem multimodal fungsional yang mampu menangani:
  - Sintesis teks ke gambar
  - Keterangan gambar
  - Pengambilan gambar berbasis teks
  - Jawaban pertanyaan visual.
- Dengan cara ini, **AI** multimodal dapat mengintegrasikan berbagai modalitas, memberikan kemampuan lanjutan untuk tugas-tugas yang melibatkan bahasa dan penglihatan.

Batasan Model Multimodal Besar (LMM)

Persyaratan dan bias data	Model-model ini membutuhkan kumpulan data yang sangat besar dan beragam untuk pelatihan. Namun, ketersediaan dan kualitas kumpulan data tersebut bisa menjadi tantangan. Selain itu, jika data pelatihan mengandung bias, model cenderung mewarisi dan mungkin memperkuat bias ini, yang mengarah pada hasil yang tidak adil atau tidak etis.
Sumber daya komputasi	Melatih dan menjalankan **model multimodal besar** membutuhkan sumber daya komputasi yang signifikan, membuatnya mahal dan kurang dapat diakses oleh organisasi yang lebih kecil atau peneliti independen.
Interpretabilitas dan eksplainabilitas	Seperti halnya model **AI** yang kompleks, memahami bagaimana model ini membuat keputusan bisa jadi sulit. Kurangnya transparansi ini bisa menjadi masalah kritis, terutama dalam aplikasi sensitif seperti perawatan kesehatan atau penegakan hukum.
Integrasi modalitas	Mengintegrasikan berbagai jenis data secara efektif (seperti teks, gambar, dan audio) dengan cara yang benar-benar memahami nuansa setiap modalitas sangat menantang. Model ini mungkin tidak selalu secara akurat memahami konteks atau seluk-beluk komunikasi manusia yang berasal dari menggabungkan modalitas ini.
Generalisasi dan overfitting	Meskipun model ini dilatih pada kumpulan data yang luas, model mungkin kesulitan dengan generalisasi ke data atau skenario baru yang tidak terlihat yang secara signifikan berbeda dari data pelatihan mereka. Sebaliknya, model mungkin terlalu cocok dengan data pelatihan, menangkap kebisingan dan anomali sebagai pola.

Terima Kasih