Utilização de IA Generativa na Produção de Dados Sintéticos para Análise de Cores da Língua na Medicina Tradicional Chinesa. Um Estudo Exploratório

Ephraim Ferreira Medeiros
Centro de Estudos de Acupuntura e Terapias Alternativas (Ceata)
webmaster@acupunturabrasil.org

Resumo

Este estudo preliminar explora a produção e aplicação de dados sintéticos para a análise de cores da língua, com foco em diferentes categorias de corpo e revestimento na medicina chinesa. Utilizamos como referência dados de aproximadamente 1000 pacientes na China, ampliando para 1.000.000 novas amostras de cores em cada categoria, resultando em um total de 17 milhões de pontos de dados. O objetivo principal foi criar uma visualização detalhada das variações de cor originalmente documentadas, proporcionando uma ferramenta visual robusta para referência futura. O uso desses dados para treinar modelos de aprendizado de máquina é uma possibilidade adicional, mas não o objetivo primário deste trabalho.

Introdução

Os avanços na inteligência artificial (IA) e no aprendizado de máquina têm revolucionado a medicina moderna, proporcionando inovações significativas na análise de dados clínicos. Nosso estudo foca na produção e aplicação de dados sintéticos para a análise de cores da língua, uma prática tradicional na medicina chinesa.

Para a execução do projeto, utilizamos bibliotecas como NumPy e Matplotlib, fundamentais para a manipulação de dados e visualização. O uso de uma GPU NVIDIA® Tesla® P100 foi essencial para acelerar o processamento dos dados e o treinamento dos modelos de IA. Esse processo dura cerca de 20 minutos para ser completado com essa configuração.

É possível adaptar o script para rodar em ambientes gratuitos, como o Google Colab, utilizando menos amostras para reduzir o tempo de execução e os recursos necessários. O Google Colab oferece um ambiente com GPU gratuita, embora as especificações sejam mais limitadas em comparação com a NVIDIA Tesla P100. Seguindo as orientações abaixo, é possível ajustar o script para funcionar no free tier do Google Colab:

1. **Redução do Número de Amostras**: Ajuste o parâmetro `num_samples` para um valor menor, como 100.000, para garantir que o script possa ser executado dentro dos limites de tempo e recursos do Google Colab.
2. **Configuração do Ambiente**: No Google Colab, é possível ativar o uso de GPU gratuitamente:
    - Vá até o menu `Editar` > `Configurações do Notebook`.
    - Na aba `Hardware Acelerado`, selecione `GPU` como tipo de hardware.
3. **Sugestão de Novos Valores**:
    - `num_samples` reduzido para 100.000
    - `sample_indices` para plotagem reduzido para 10.000

Produção de Dados Sintéticos

Os dados sintéticos têm se mostrado uma alternativa segura e eficaz para a preservação da privacidade dos pacientes e a expansão dos conjuntos de dados disponíveis para pesquisa. No nosso estudo, utilizamos as bibliotecas `numpy` e `matplotlib` para gerar e visualizar os dados sintéticos. A técnica empregada envolveu a adição de variabilidade controlada aos valores RGB das cores da língua documentadas em estudos com humanos. Isso permitiu ampliar a amostragem original de 1000 pacientes para 1.000.000 amostras de cores em cada categoria, viabilizando uma análise mais robusta e generalizável, com a segurança de que os valores RGB estavam dentro dos obtidos em um estudo com humanos.

Visualização e Análise dos Dados Sintéticos

Geramos dados sintéticos para as categorias de corpo e revestimento da língua com variabilidade controlada. Utilizamos as seguintes categorias e cores base para o corpo e revestimento da língua, juntamente com as variações em RGB, para criar visualizações detalhadas.

Definição das Cores Base

Utilizamos as cores base fornecidas pelo artigo de referência "Objective Study on Traditional Chinese Medicine Tongue Diagnosis", publicado na revista Engineering Science em 2001. Essas cores foram definidas com valores médios e desvios padrão para as diferentes categorias de cores da língua, tanto para o corpo quanto para o revestimento da língua.

Corpo da Língua

- **Língua Pálida Branca**: Base: (188, 147, 138); Variação: (17.2, 12.5, 15.4)
- **Língua Pálida Vermelha**: Base: (205, 145, 130); Variação: (12.9, 15.4, 14.4)
- **Língua Vermelha Carmezim**: Base: (220, 140, 137); Variação: (20.5, 14.9, 16.8)
- **Língua Vermelha Escura**: Base: (169, 112, 108); Variação: (6.8, 7.8, 11.6)
- **Língua Vermelha Púrpura**: Base: (181, 135, 135); Variação: (12.4, 14.6, 12.9)
- **Língua Púrpura Pálida**: Base: (192, 122, 122); Variação: (14.3, 12.4, 17.1)
- **Língua Púrpura Azulada**: Base: (158, 135, 138); Variação: (15.6, 19.8, 19.3)

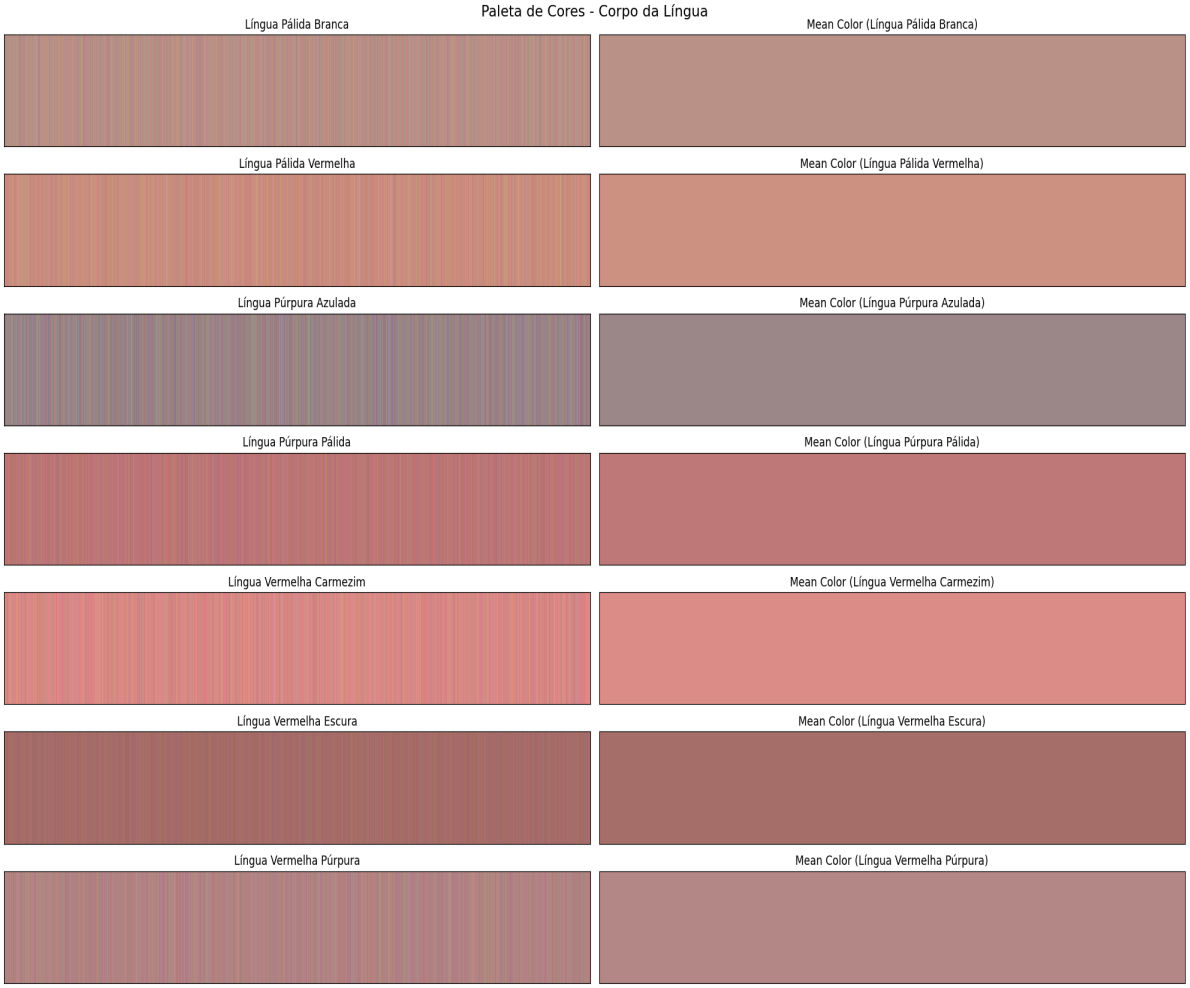
Revestimento da Língua

- **Revestimento Branco Espesso**: Base: (234, 224, 217); Variação: (17.2, 14.1, 13.4)
- **Pegajoso Branco Espesso**: Base: (231, 213, 208); Variação: (15.2, 16.7, 11.2)
- **Revestimento Branco Fino**: Base: (227, 198, 195); Variação: (15.6, 12.5, 18.5)
- **Pegajoso Branco Fino**: Base: (223, 190, 190); Variação: (11.8, 13.7, 10.5)
- **Revestimento Amarelo Fino**: Base: (214, 179, 158); Variação: (11.7, 12.8, 9.9)
- **Pegajoso Amarelo Fino**: Base: (209, 162, 138); Variação: (12.1, 14.8, 13.5)
- **Revestimento Amarelo Espesso**: Base: (202, 157, 129); Variação: (16.4, 10.7, 11.3)
- **Pegajoso Amarelo Espesso**: Base: (194, 148, 129); Variação: (16.7, 11.5, 14.2)
- **Revestimento Amarelo**: Base: (182, 133, 102); Variação: (18.4, 15.5, 22.1)
- **Cinza Preto**: Base: (155, 103, 74); Variação: (22.9, 21.7, 23.1)

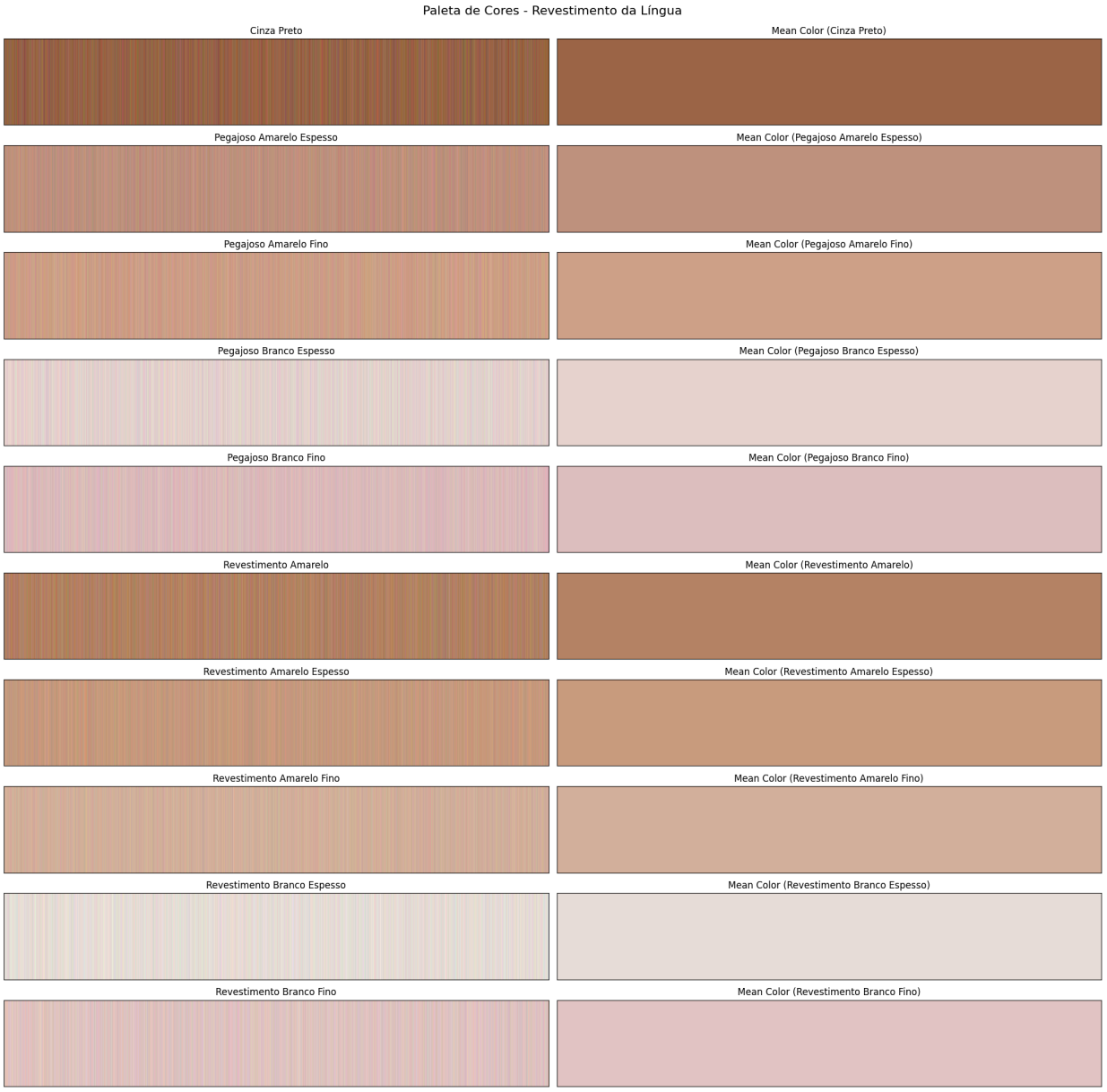
Resultados e Discussão:

As figuras apresentam as paletas de cores geradas pelo processo de sintetização dos dados (lado esquerdo) e as cores médias RGB puras baseadas nos valores de referência do estudo original (lado direito). O lado esquerdo mostra a diversidade de cores obtidas através do processo de geração de dados sintéticos, enquanto o lado direito representa a cor média calculada a partir dos valores de referência fornecidos pelo artigo base que obteve os valores RGB de pacientes reais.

Corpo da Lingua



Revestimento (Saburra)



O total de dados gerados pelo script, considerando cada canal RGB, foi de 51 milhões de pontos de dados. Cada ponto de dado pode ser considerado equivalente a um paciente, e esse número é comparável à população de países como a Colômbia ou a Coreia do Sul, ilustrando a magnitude e a abrangência dos dados gerados. Além disso, cada paleta final mostrada na coluna da esquerda é constituída por 8888 pontos de dados.

Análise Visual

Ao observar a imagem com cuidado, parece que as diferenças entre os revestimentos pegajosos e os revestimentos puros de mesma cor são sutis. Essas diferenças podem estar mais relacionadas ao tom (Shade) do que a grandes variações nos valores RGB. A textura "pegajosa" pode estar afetando a percepção visual das cores, mas sem um tratamento mais específico dessas variações, a distinção baseada apenas na visualização dos códigos RGB pode ser desafiadora. Revestimentos pegajosos podem aparecer mais escuros, criando um contraste visual que é perceptível como um tom mais escuro da mesma cor base. Pequenas variações na intensidade da cor podem criar uma percepção diferente, mesmo que a cor base seja a mesma.

Os revestimentos brancos finos podem parecer rosados nas paletas devido ao contraste com a cor predominante do corpo da língua. Quando um revestimento fino e claro é colocado sobre uma superfície colorida, como a língua, a cor do corpo da língua pode influenciar a percepção da cor do revestimento. Isso ocorre porque a fina camada do revestimento permite que a cor subjacente da língua se misture e modifique a aparência do revestimento, resultando em um tom rosado.

O revestimento pegajoso parece um pouco mais escuro, indicando uma diferença no tom. Entre Revestimento Branco Fino vs. Pegajoso Branco Fino a diferença é mais sutil, mas o revestimento pegajoso pode parecer ligeiramente mais escuro e a principal diferença está na tonalidade (shade) das cores.

Data Augmentation

Para dados que consistem em apenas RGB e não imagens reais , muitas das técnicas tradicionais de aumento de dados realmente não se aplicam diretamente, porque transformações como rotação, inversão, e cortes não fazem sentido para uma imagem que é uniformemente colorida. Por isso, a técnica de adição de ruído controlado se destaca como uma das mais adequadas para este caso específico.

Razões para Escolher a Adição de Ruído Controlado

1. **Preservação da Cor**: Ao adicionar um pequeno ruído aos valores RGB, você continua mantendo os dados dentro daquele intervalo RGB da classe de cor definida nas características de corpo e saburra da língua mas ao mesmo tempo que simula variações naturais que podem ocorrer devido a diferenças de iluminação ou qualidade da câmera.
2. **Robustez do Modelo**: Introduzir essas pequenas variações pode ajudar o modelo a se tornar mais robusto a mudanças sutis nas entradas, preparando-o melhor para aplicações práticas onde tais variações são comuns.

Outras Técnicas a Considerar

Embora a adição de ruído seja uma técnica primária e bastante útil, outras abordagens podem ser consideradas dependendo de suas necessidades específicas:

- **Ajuste de Luminosidade**: Alterar a luminosidade dos quadradinhos pode simular diferentes condições de iluminação. Isso pode ser feito ajustando os valores RGB uniformemente para cima ou para baixo dentro de um limite que ainda representa a cor original.

Alternativas e Padrões na Produção de Dados Sintéticos

Nossa abordagem é particularmente adequada ao objetivo do estudo, que se concentra na geração de códigos RGB ao invés da manipulação direta de imagens. Essa escolha é reforçada por sua simplicidade e eficiência. A geração de códigos RGB sintéticos através de variabilidade controlada é um método robusto e bem fundamentado, permitindo a criação de grandes volumes de dados de forma rápida e precisa.

Enquanto outras técnicas, como redes adversariais generativas (GANs), podem oferecer vantagens para a geração de imagens sintéticas complexas, elas também introduzem maior complexidade e requisitos computacionais. Dada a natureza do nosso estudo, a abordagem escolhida não só é apropriada, mas também se destaca pela sua eficácia em atender às necessidades específicas da pesquisa.

Importância da Produção de Dados Sintéticos na Medicina

A produção de dados sintéticos na medicina é de grande importância por várias razões:

- **Privacidade e Segurança**: Dados sintéticos permitem que pesquisadores usem informações de saúde sem comprometer a privacidade dos pacientes. Isso é crucial em um campo onde a proteção de dados sensíveis é mandatória .
- **Expansão de Amostras**: Eles permitem a ampliação dos conjuntos de dados, especialmente em situações onde os dados reais são limitados. Isso é essencial para melhorar a robustez dos modelos de aprendizado de máquina e para realizar análises estatísticas mais precisas .
- **Simulações e Previsões**: Dados sintéticos são úteis para realizar simulações e previsões em ambientes controlados, ajudando no desenvolvimento de novos tratamentos e na avaliação de políticas de saúde sem a necessidade de realizar estudos clínicos caros e demorados .
- **Treinamento de Modelos de IA**: A utilização de dados sintéticos é fundamental para treinar modelos de inteligência artificial, garantindo que esses modelos possam generalizar bem em situações reais e detectar padrões que poderiam não ser visíveis em conjuntos de dados menores ou menos diversificados .

Importância da Detecção Precoce de Padrões Emergentes com IA

Capturar sutilezas, como cores com valores RGB não tão dominantes mas presentes, pode ser crucial para a detecção precoce de padrões emergentes na análise da língua, mesmo antes do surgimento de sintomas visíveis. Este conceito está alinhado com a prática de "治未病" (Zhì wèi bìng), que significa "tratar antes da doença", um princípio fundamental na medicina chinesa. A mesma abordagem se aplica ao prognóstico, quando a cor de uma determinada região da língua, da língua como um todo ou ainda padrões de saburra retornam ao padrão considerado normal. Nesse contexto, um histórico de imagens do paciente facilitaria o processo, potencializando as capacidades da IA como um assistente confiável na análise e classificando padrões em imagens.

A Detecção Precoce e Seus Benefícios

1. **Identificação de Anomalias Sutis**: As variações sutis nas cores da língua podem indicar desequilíbrios ou problemas de saúde que ainda não se manifestam em sintomas evidentes. A identificação dessas nuances permite a intervenção antecipada, prevenindo o desenvolvimento de doenças mais graves.
2. **Precisão na Análise**: Utilizando técnicas de inteligência artificial (IA), é possível analisar grandes volumes de dados e identificar padrões que seriam imperceptíveis ao olho humano. A IA pode detectar alterações mínimas nas cores da língua, fornecendo uma base mais precisa para o diagnóstico.
3. **Intervenção Antecipada**: A detecção precoce permite que intervenções sejam feitas em estágios iniciais, muitas vezes antes que os sintomas se tornem aparentes. Isso pode melhorar significativamente os resultados do tratamento e reduzir o risco de complicações.
4. **Alinhamento com Princípios Tradicionais**: "治未病" (Zhì wèi bìng) é um pilar na medicina tradicional chinesa de alto nível, focando na prevenção e na manutenção da saúde. Integrar a IA com esse princípio permite uma abordagem moderna e avançada para um conceito antigo e comprovado, ampliando as possibilidades de tratamento preventivo.

Aplicação Futuras e Segunda Fase do Projeto

O estudo abre caminho para o uso de dados sintéticos em larga escala no treinamento e refinamento de modelos de Machine Learning, utilizando redes neurais convolucionais (CNNs). Essa aplicação será a segunda fase experimental do projeto, permitindo a integração de IA avançada para melhorar a precisão e a eficácia dos diagnósticos baseados na análise de cores da língua.

O Papel das LLMs e da IA Generativa na Aceleração de Projetos de Dados Sintéticos para Pesquisas em Medicina Chinesa

Neste projeto, nenhuma linha de código foi escrita manualmente por humanos. Todo o código foi obtido através de técnicas de Prompt Engineering. Este método inovador demonstra como ferramentas avançadas de IA podem transformar o desenvolvimento de software, automatizando processos que antes exigiam extensa intervenção humana.

O processamento e execução do código foram realizados na plataforma Kaggle, que oferece um plano gratuito generoso, permitindo que pesquisadores e desenvolvedores aproveitem poderosos recursos de computação sem custos elevados. A Kaggle é amplamente reconhecida por fornecer um ambiente acessível e eficiente para projetos de ciência de dados e aprendizado de máquina, facilitando o trabalho com grandes volumes de dados e modelos complexos.

Este estudo abre caminho para o uso de dados sintéticos em larga escala no treinamento e refinamento de modelos de Machine Learning, utilizando redes neurais convolucionais (CNN). Esta abordagem está planejada para a segunda fase experimental do projeto, onde se espera um refinamento ainda maior dos modelos preditivos.

Mesmo especialistas juniores em Machine Learning podem testar, prototipar e incubar ideias e até realizar projetos completos de nível básico em poucas horas, o que antes levaria meses devido à inexperiência em programação. As ferramentas de IA generativa e as plataformas de computação acessíveis democratizam o campo, permitindo que mais indivíduos de diferentes áreas da ciência e com pouco background em programação possam contribuir mais diretamente para avanços tecnológicos e científicos significativos na área de IA.

Explicação Técnica Adicional

Prompt Engineering e Design do Código

A ideia de alto nível e o design da estrutura do código foram concebidos pelo autor, mas a implementação prática foi realizada usando inteligência artificial generativa, especificamente GPT-4o (EUA) e DeepSeek-Coder-V2 (China). As LLMs (Large Language Models) são modelos de inteligência artificial treinados em vastas quantidades de texto, capazes de compreender e gerar linguagem natural de forma impressionante. GPT-4o, desenvolvido pela OpenAI, é conhecido por sua capacidade avançada de gerar texto coerente e relevante, facilitando a criação de códigos complexos a partir de descrições detalhadas. DeepSeek-Coder-V2, uma IA chinesa, complementa essa capacidade com um foco especial em resolver problemas técnicos e fornecer soluções eficientes em programação. Nesse projeto, ambas as LLMs foram utilizadas para produzir todo o código necessário, demonstrando sua eficácia e contribuindo para uma implementação mais rápida e precisa.

As LLMs ajudam a gerar código a partir de descrições detalhadas e objetivos definidos, economizando tempo e reduzindo erros humanos. Elas são capazes de resolver problemas técnicos rapidamente e aprimorar a produtividade dos desenvolvedores ao automatizar partes do processo de codificação, permitindo que os desenvolvedores se concentrem mais na lógica de alto nível e na inovação do projeto. Todo o código gerado neste artigo foi produzido usando inteligência artificial generativa, especificamente GPT-4o e DeepSeek-Coder-V2, exemplificando como essas tecnologias podem acelerar significativamente projetos de dados sintéticos para pesquisas na Medicina Chinesa.

Design e Implementação com Inteligência Artificial - Visão geral da Metodologia de Desenvolvimento Assistido por IA

Inicialmente, o autor define os objetivos do projeto e delineia a lógica necessária para alcançá-los. Este processo inclui a criação de prompts detalhados que orientam a inteligência artificial na geração do código necessário.

Definição de Objetivos e Estrutura: O autor estabelece claramente os objetivos do código, incluindo a geração de dados sintéticos, a variabilidade dos valores RGB e a visualização das paletas de cores. Esta etapa garante que todos os requisitos funcionais e não funcionais sejam bem compreendidos e documentados, em conformidade com as práticas de engenharia de software.

Criação de Prompts Específicos: Utilizando técnicas de Engenharia de Prompts, o autor elabora instruções detalhadas que especificam cada etapa do código. Esses prompts são então processados pelas IAs GPT-4o e DeepSeek-Coder-V2, resultando na geração do código correspondente. Essa abordagem torna o desenvolvimento de código acessível mesmo para aqueles com conhecimentos básicos em linguagens de programação e codificação.

Geração e Refinamento do Código: As IAs produzem o código com base nos prompts fornecidos. O autor revisa o código gerado, realiza os ajustes necessários e redefine os prompts para aprimorar a precisão e a funcionalidade do código. Esse ciclo de revisão e refinamento é essencial para garantir a qualidade do código, alinhando-se com práticas recomendadas de Continuous Integration (CI) e Continuous Improvement (CI).

Execução e Validação: O código gerado é executado na plataforma Kaggle. A validação é realizada para garantir que o código atende aos objetivos definidos e funciona conforme esperado. Este processo inclui a execução de testes automatizados e manuais para verificar a conformidade com os requisitos, seguindo práticas de validação e verificação (V&V) recomendadas.

Iteração e Melhoria: Com base no feedback e nos resultados da execução, o processo pode ser iterado várias vezes, refinando os prompts e ajustando o código para alcançar a solução ideal. Essa abordagem iterativa está alinhada com o ciclo de desenvolvimento ágil, onde o feedback contínuo é usado para aprimorar o produto final.

Essa metodologia não só acelera o desenvolvimento do projeto, como também torna viável a realização de projetos de alta qualidade, mesmo para desenvolvedores com conhecimentos básicos em linguagens de programação e codificação. As práticas adotadas garantem que o código gerado esteja alinhado com as melhores práticas de desenvolvimento de software e aprendizado de máquina, incluindo design iterativo, integração contínua e melhoria contínua.

Link para o Notebook com o código usado
https://www.kaggle.com/code/ephraimmedeiros/paleta-dados-sinteticos-lingua

Referências

1. Kennedy, S., 2024. Weighing the pros and cons of synthetic healthcare data use. Health Tech Analytics. Available at: https://www.techtarget.com/healthtechanalytics/feature/Weighing-the-pros-and-cons-of-synthetic-healthcare-data-use [Accessed 23 March 2004].
2. Murtaza, H., Ahmed, M., Khan, N.F., Murtaza, G., Zafar, S. and Bano, A., 2023. Synthetic data generation: State of the art in health care domain. Computer Science Review, 48, p.100546. Available at: https://www.sciencedirect.com/science/article/pii/S1574013723000138 [Accessed 7 April 2004].
3. Chen, R.J., Lu, M.Y., Chen, T.Y. et al., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5, pp.493-497. Available at: https://doi.org/10.1038/s41551-021-00751-8 [Accessed 15 February 2004].
4. Arora, A., 2023. Synthetic data: the future of open-access health-care datasets? The Lancet, 401(10381), p.997. Available at: https://doi.org/10.1016/S0140-6736(23)00324-0 [Accessed 12 May 2004].
5. Kamińska, J., 2022. How healthcare enterprises can benefit from synthetic health data, including 3 practical use-cases. Statice. Available at: https://www.statice.ai/post/healthcare-synthetic-health-data-3-use-cases [Accessed 19 January 2004].
6. Shaip, 2024. Synthetic data in healthcare: Definition, Benefits, and Challenges. Shaip. Available at: https://www.shaip.com/blog/synthetic-data-in-healthcare/ [Accessed 28 June 2004].
7. Agarwal, S., 2024. Navigating The Potential And Perils Of Synthetic Data In Healthcare. Forbes. Available at: https://www.forbes.com/sites/shashankagarwal/2024/01/27/navigating-the-potential-and-perils-of-synthetic-data-in-healthcare/ [Accessed 5 February 2004].
8. Madden, B., 2023. Synthetic Data in Healthcare: the Great Data Unlock. Hospitalogy. Available at: https://hospitalogy.com/articles/2023-11-02/synthetic-data-in-healthcare-great-data-unlock/ [Accessed 22 March 2004].
9. Giuffrè, M. and Shung, D.L., 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digital Medicine, 6, p.186. Available at: https://doi.org/10.1038/s41746-023-00927-3 [Accessed 4 June 2004].
10. Tang, Z., Li, P., Zeng, Y., Zhang, Y., Zheng, Y., Li, H., Wei, Q. and Xiao, Z., 2021. Progress in Computer-Aided Tongue Image Analysis Diagnosis [Original title in Chinese: 计算机辅助舌象分析诊断的研究进展]. biomedRxiv. Available at: http://www.biomedrxiv.org.cn/article/doi/bmr.202106.00006 [Accessed 29 April 2004].
11. Chen, J.J., Yeh, S.Y. and Chiang, J.Y., Feature Extraction For The Automatic Tongue Diagnosis [Original title in Chinese: 中醫舌診電腦化之特徵擷取方法]. China Medical College Hospital and National Sun Yat-Sen University. Available at: https://image.cse.nsysu.edu.tw/ServerPaper/24/%A4%A4%C2%E5%A6%DE%B6E

%B9q%B8%A3%A4%C6%A4%A7%AFS%BCx%C2%5E%A8%FA%A4%E8%AAk.pdf [Accessed 13 May 2004].
12. Hui, S.C., He, Y. and Thach, D.T.C., 2008. Machine learning for tongue diagnosis. In: Information, Communications & Signal Processing, 2007 6th International Conference on. IEEE Xplore. Available at: https://doi.org/10.1109/ICICS.2007.4449631 [Accessed 8 March 2004].
13. Xue, W. and Huang, S., 2001. Objective Study on Traditional Chinese Medicine Tongue Diagnosis [Original title in Chinese: 中医舌诊客观化研究]. Engineering Science, 3(1). Available at: https://www.engineering.org.cn/ch/article/21120/detail [Accessed 17 February 2004].
14. Weng, W. and Cao, Y., 1995. Development and Application of a True Color Image System for Traditional Chinese Medicine Tongue Image [Original title in Chinese: 中医舌象真彩色图像系统的研制与应用]. Journal of Practical Traditional Chinese and Western Medicine, (4), p.165. [Accessed 22 January 2004].

Ephraim Ferreira Medeiros

Ephraim Ferreira Medeiros é Acupunturista, Pesquisador em Medicina Chinesa, UX Designer certificado pelo Google e Interaction Designer formado pela Michigan University. Ele também é Especialista certificado nível Junior em Machine Learning pela Deeplearning.ai e é webmaster e fundador do site Acupunturabrasil.org . Ephraim é pesquisador afiliado ao Centro de Estudos de Acupuntura e Terapias Alternativas (CEATA).

ENGLISH

Use of Generative AI in the Production of Synthetic Data for Tongue Color Analysis in Chinese Medicine

Ephraim Ferreira Medeiros Center for the Study of Acupuncture and Alternative Therapies (Ceata) webmaster@acupunturabrasil.org

Abstract This preliminary study explores the production and application of synthetic data for tongue color analysis, focusing on different categories of tongue body and coating in Chinese medicine. Using reference data from approximately 1,000 patients in China, we expanded to 1,000,000 new color samples in each category, resulting in a total of 17 million data points. The primary goal was to create a detailed visualization of the originally documented color variations, providing a robust visual tool for future reference. Using these data to train machine learning models is a secondary possibility but not the primary objective of this work.

Introduction Advances in artificial intelligence (AI) and machine learning have revolutionized modern medicine, bringing significant innovations in clinical data analysis. Our study focuses on the production and application of synthetic data for tongue color analysis, a traditional practice in Chinese medicine. For this project, we used libraries such as NumPy and Matplotlib, essential for data manipulation and visualization. The use of an NVIDIA® Tesla® P100 GPU was crucial for accelerating data processing and AI model training, a process that takes about 20 minutes with this configuration. The script can be adapted to run in free environments like Google Colab, using fewer samples to reduce execution time and required resources. Google Colab offers a free GPU environment, although its specifications are more limited compared to the NVIDIA Tesla P100. Following the guidelines below, the script can be adjusted to work on Google Colab's free tier:

1. **Reducing the Number of Samples**: Adjust the num_samples parameter to a lower value, such as 100,000, to ensure the script can run within Google Colab's time and resource limits.
2. **Environment Setup**: In Google Colab, you can enable GPU usage for free:
    - Go to Edit > Notebook Settings.
    - In the Hardware Accelerator tab, select GPU as the hardware type.
3. **Suggested New Values**:
    - num_samples reduced to 100,000
    - sample_indices for plotting reduced to 10,000

Production of Synthetic Data Synthetic data has proven to be a safe and effective alternative for preserving patient privacy and expanding the datasets available for research. In our study, we used the numpy and matplotlib libraries to generate and visualize synthetic data. The technique employed involved adding controlled variability to the RGB values of tongue colors documented in human studies. This allowed us to expand the original sample of 1,000 patients to 1,000,000 color samples in each category, enabling a more robust and generalizable analysis, ensuring that the RGB values were within those obtained in human studies.

Visualization and Analysis of Synthetic Data We generated synthetic data for the tongue body and coating categories with controlled variability. We used the following base colors and variations in RGB for the tongue body and coating to create detailed visualizations.

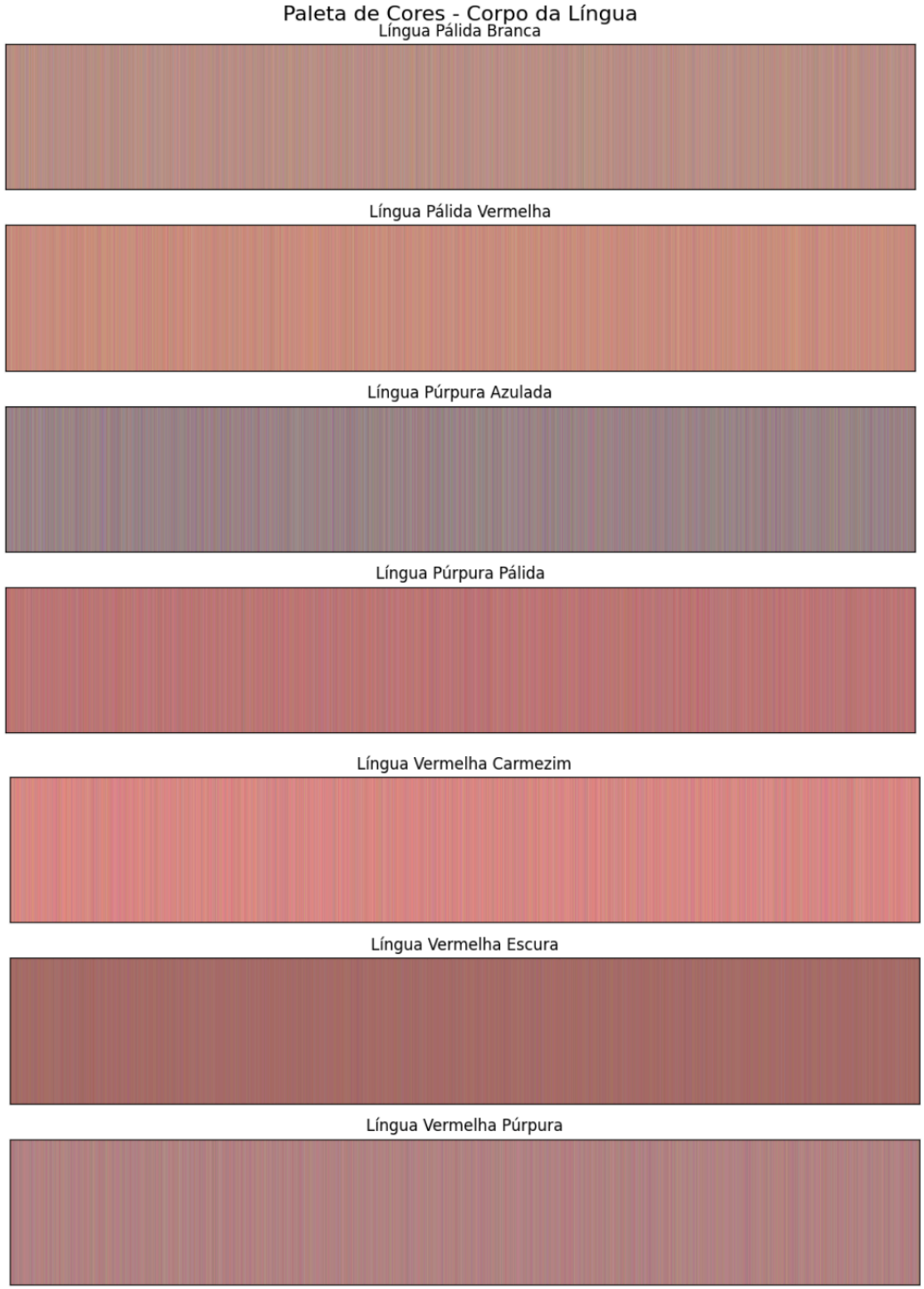
Definition of Base Colors We used the base colors provided by the reference article "Objective Study on Traditional Chinese Medicine Tongue Diagnosis," published in the journal Engineering Science in 2001. These colors were defined with average values and standard deviations for different tongue color categories, both for the body and coating of the tongue. **Tongue Body Colors**

- Pale White Tongue: Base: (188, 147, 138); Variation: (17.2, 12.5, 15.4)
- Pale Red Tongue: Base: (205, 145, 130); Variation: (12.9, 15.4, 14.4)
- Crimson Red Tongue: Base: (220, 140, 137); Variation: (20.5, 14.9, 16.8)
- Dark Red Tongue: Base: (169, 112, 108); Variation: (6.8, 7.8, 11.6)
- Purple Red Tongue: Base: (181, 135, 135); Variation: (12.4, 14.6, 12.9)
- Pale Purple Tongue: Base: (192, 122, 122); Variation: (14.3, 12.4, 17.1)
- Bluish Purple Tongue: Base: (158, 135, 138); Variation: (15.6, 19.8, 19.3)

Tongue Coating Colors

- Thick White Coating: Base: (234, 224, 217); Variation: (17.2, 14.1, 13.4)
- Sticky Thick White Coating: Base: (231, 213, 208); Variation: (15.2, 16.7, 11.2)
- Thin White Coating: Base: (227, 198, 195); Variation: (15.6, 12.5, 18.5)
- Sticky Thin White Coating: Base: (223, 190, 190); Variation: (11.8, 13.7, 10.5)
- Thin Yellow Coating: Base: (214, 179, 158); Variation: (11.7, 12.8, 9.9)
- Sticky Thin Yellow Coating: Base: (209, 162, 138); Variation: (12.1, 14.8, 13.5)
- Thick Yellow Coating: Base: (202, 157, 129); Variation: (16.4, 10.7, 11.3)
- Sticky Thick Yellow Coating: Base: (194, 148, 129); Variation: (16.7, 11.5, 14.2)
- Yellow Coating: Base: (182, 133, 102); Variation: (18.4, 15.5, 22.1)
- Gray Black Coating: Base: (155, 103, 74); Variation: (22.9, 21.7, 23.1)

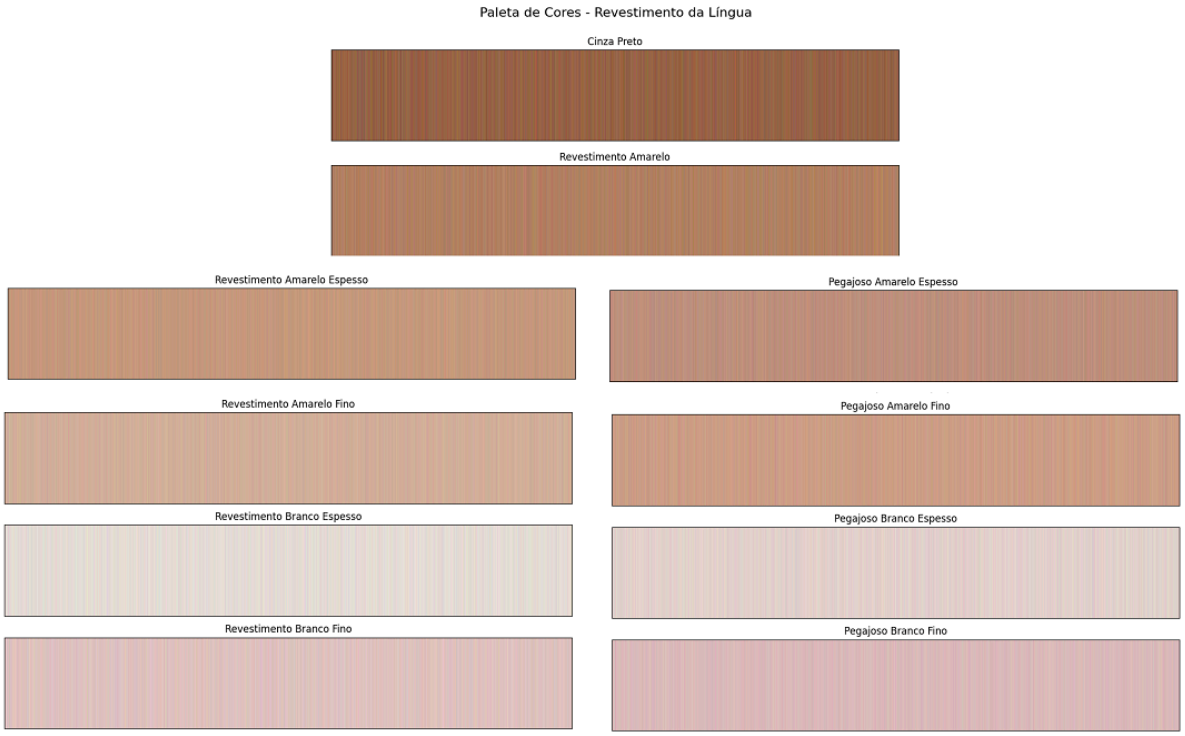
Results and Discussion
Tongue Body



Color Palette - Tongue Body (top to bottom)

1. Pale White Tongue
2. Pale Red Tongue
3. Bluish Purple Tongue
4. Pale Purple Tongue
5. Crimson Red Tongue
6. Dark Red Tongue
7. Purple Red Tongue

Tongue Coating



Color Palette - Tongue Coating

- Top (two items):
    - Left: **Gray Black**
    - Right: **Yellow Coating**
- Middle (four items):
    - Left: **Thick Yellow Coating**
    - Right: **Sticky Thick Yellow Coating**
    - Left: **Thin Yellow Coating**
    - Right: **Sticky Thin Yellow Coating**
- Bottom (four items):
    - Left: **Thick White Coating**
    - Right: **Sticky Thick White Coating**
    - Left: **Thin White Coating**
    - Right: **Sticky Thin White Coating**

The total data generated by the script, considering each RGB channel, amounted to 51 million data points. Each data point can be considered equivalent to a patient, comparable to the population of countries like Colombia or South Korea, illustrating the magnitude and scope of the data generated. Additionally, each final palette consists of 30,000 data points, and considering 17 categories, we have a total of 510,000 data points.

Observing the image carefully, the differences between sticky coatings and pure coatings of the same color are subtle. These differences may be more related to the shade than to significant variations in RGB values. The "sticky" texture may affect the visual perception of

colors, but without more specific treatment of these variations, distinguishing based only on the visualization of RGB codes can be challenging. Sticky coatings may appear darker, creating a visual contrast perceived as a darker shade of the same base color. Small variations in color intensity can create a different perception, even if the base color is the same.

Visual Analysis The sticky coating appears slightly darker, indicating a difference in tone. Between Thin White Coating vs. Sticky Thin White Coating, the difference is more subtle, but the sticky coating may appear slightly darker, and the main difference lies in the shade of the colors.

Data Augmentation For data consisting of only RGB and not actual images, many traditional data augmentation techniques do not directly apply, as transformations such as rotation, flipping, and cropping do not make sense for a uniformly colored image. Therefore, the technique of adding controlled noise stands out as one of the most suitable for this specific case.

Reasons for Choosing Controlled Noise Addition

1. **Color Preservation**: By adding small noise to the RGB values, you continue to keep the data within the RGB range of the defined tongue body and coating color classes while simulating natural variations that may occur due to differences in lighting or camera quality.
2. **Model Robustness**: Introducing these small variations can help the model become more robust to subtle changes in inputs, better preparing it for practical applications where such variations are common.

Other Techniques to Consider Although noise addition is a primary and quite useful technique, other approaches can be considered depending on specific needs:

- **Brightness Adjustment**: Altering the brightness of the squares can simulate different lighting conditions. This can be done by adjusting the RGB values uniformly up or down within a limit that still represents the original color.

Alternatives and Standards in Synthetic Data Production Our approach is particularly suited to the study's objective, focusing on generating RGB codes rather than direct image manipulation. This choice is reinforced by its simplicity and efficiency. The generation of synthetic RGB codes through controlled variability is a robust and well-founded method, allowing the creation of large volumes of data quickly and accurately. While other techniques, such as Generative Adversarial Networks (GANs), may offer advantages for generating complex synthetic images, they also introduce greater complexity and computational requirements. Given the nature of our study, the chosen approach is not only appropriate but also stands out for its effectiveness in meeting the research's specific needs.

Importance of Synthetic Data Production in Medicine The production of synthetic data in medicine is essential for several reasons:

- **Privacy and Security**: Synthetic data allows researchers to use health information without compromising patient privacy. This is mandatory in a field where the protection of sensitive data is crucial.
- **Sample Expansion**: They enable the expansion of datasets, especially in situations where real data is limited. This is essential for improving the robustness of machine learning models and conducting more accurate statistical analyses.
- **Simulations and Predictions**: Synthetic data is useful for conducting simulations and predictions in controlled environments, aiding in the development of new treatments and the evaluation of health policies without the need for costly and time-consuming clinical studies.
- **AI Model Training**: The use of synthetic data is fundamental for training artificial intelligence models, ensuring that these models can generalize well in real situations and detect patterns that may not be visible in smaller or less diverse datasets.

Importance of Early Detection of Emerging Patterns with AI Capturing subtleties, such as colors with not-so-dominant but present RGB values, can be essential for the early detection of emerging patterns in tongue analysis, even before visible symptoms appear. This concept aligns with the practice of "治未病" (Zhì wèi bìng), which means "treat before illness," a fundamental principle in Chinese medicine. The same approach applies to prognosis when the color of a specific region of the tongue, the tongue as a whole, or coating patterns return to the normal standard. In this context, a patient's image history would facilitate the process, enhancing AI's capabilities as a reliable assistant in analyzing and classifying patterns in images.

Early Detection and Its Benefits

1. **Identification of Subtle Anomalies**: Subtle variations in tongue colors can indicate imbalances or health issues that have not yet manifested in obvious symptoms. Identifying these nuances allows for early intervention, preventing the development of more severe diseases.
2. **Accuracy in Analysis**: Using artificial intelligence (AI) techniques, it is possible to analyze large volumes of data and identify patterns that would be imperceptible to the human eye. AI can detect minute changes in tongue colors, providing a more accurate basis for diagnosis.
3. **Early Intervention**: Early detection allows for interventions to be made at initial stages, often before symptoms become apparent. This can significantly improve treatment outcomes and reduce the risk of complications.
4. **Alignment with Traditional Principles**: "治未病" (Zhì wèi bìng) is a high-level pillar in traditional Chinese medicine, focusing on prevention and health maintenance. Integrating AI with this principle allows for a modern and advanced approach to an ancient and proven concept, expanding the possibilities for preventive treatment.

Future Applications and Second Phase of the Project The study paves the way for the large-scale use of synthetic data in training and refining Machine Learning models, using convolutional neural networks (CNNs). This application will be the second experimental phase of the project, allowing the integration of advanced AI to improve the accuracy and effectiveness of diagnostics based on tongue color analysis.

The Role of LLMs and Generative AI in Accelerating Synthetic Data Projects for Chinese Medicine Research In this project, no line of code was written manually by humans. All code was obtained through Prompt Engineering techniques. This innovative method demonstrates how advanced AI tools can transform software development, automating processes that previously required extensive human intervention. The code processing and execution were performed on the Kaggle platform, which offers a generous free plan, allowing researchers and developers to leverage powerful computing resources at no high costs. Kaggle is widely recognized for providing an accessible and efficient environment for data science and machine learning projects, facilitating work with large volumes of data and complex models. This study paves the way for large-scale use of synthetic data in training and refining Machine Learning models using convolutional neural networks (CNN). This approach is planned for the second experimental phase of the project, where an even greater refinement of predictive models is expected. Even junior Machine Learning specialists can test, prototype, and incubate ideas and even complete basic-level projects in a few hours, which would have taken months due to inexperience in programming. Generative AI tools and accessible computing platforms democratize the field, allowing more individuals from different science areas and with little programming background to contribute more directly to significant technological and scientific advances in AI.

Technical Explanation Prompt Engineering and Code Design The high-level idea and code structure design were conceived by the author, but the practical implementation was carried out using generative artificial intelligence, specifically GPT-4o (USA) and DeepSeek-Coder-V2 (China). LLMs (Large Language Models) are AI models trained on vast amounts of text, capable of understanding and generating natural language impressively. GPT-4o, developed by OpenAI, is known for its advanced ability to generate coherent and relevant text, facilitating the creation of complex code from detailed descriptions. DeepSeek-Coder-V2, a Chinese AI, complements this capability with a special focus on solving technical problems and providing efficient programming solutions. In this project, both LLMs were used to produce all the necessary code, demonstrating their effectiveness and contributing to a faster and more accurate implementation. LLMs help generate code from detailed descriptions and defined objectives, saving time and reducing human errors. They can quickly solve technical problems and enhance developers' productivity by automating parts of the coding process, allowing developers to focus more on high-level logic and project innovation. All code generated in this article was produced using generative artificial intelligence, specifically GPT-4o and DeepSeek-Coder-V2, exemplifying how these technologies can significantly accelerate synthetic data projects for Chinese Medicine research.

Design and Implementation with Artificial Intelligence - Overview of AI-Assisted Development Methodology Initially, the author defines the project objectives and outlines the logic needed to achieve them. This process includes creating detailed prompts that guide the AI in generating the necessary code.

1. **Objective and Structure Definition**: The author clearly establishes the code's objectives, including synthetic data generation, RGB value variability, and color palette visualization. This step ensures that all functional and non-functional

requirements are well understood and documented, following software engineering practices.

2. **Creation of Specific Prompts**: Using Prompt Engineering techniques, the author creates detailed instructions that specify each code step. These prompts are then processed by the GPT-4o and DeepSeek-Coder-V2 AIs, resulting in the corresponding code generation. This approach makes code development accessible even to those with basic knowledge of programming languages and coding.
3. **Code Generation and Refinement**: The AIs produce the code based on the provided prompts. The author reviews the generated code, makes necessary adjustments, and refines the prompts to enhance the code's accuracy and functionality. This review and refinement cycle is essential to ensure code quality, aligning with Continuous Integration (CI) and Continuous Improvement (CI) practices.
4. **Execution and Validation**: The generated code is executed on the Kaggle platform. Validation is performed to ensure the code meets the defined objectives and functions as expected. This process includes automated and manual testing to verify compliance with requirements, following recommended validation and verification (V&V) practices.
5. **Iteration and Improvement**: Based on feedback and execution results, the process can be iterated several times, refining the prompts and adjusting the code to achieve the ideal solution. This iterative approach aligns with the agile development cycle, where continuous feedback is used to improve the final product.

This methodology not only accelerates the project's development but also makes it possible to achieve high-quality projects, even for developers with basic programming and coding knowledge. The adopted practices ensure that the generated code aligns with the best software development and machine learning practices, including iterative design, continuous integration, and continuous improvement.

Link to the Notebook with the Used Code Link to the Notebook

References

1. Kennedy, S., 2024. Weighing the pros and cons of synthetic healthcare data use. Health Tech Analytics. Available at: https://www.techtarget.com/healthtechanalytics/feature/Weighing-the-pros-and-cons-of-synthetic-healthcare-data-use [Accessed 23 March 2004].
2. Murtaza, H., Ahmed, M., Khan, N.F., Murtaza, G., Zafar, S. and Bano, A., 2023. Synthetic data generation: State of the art in the healthcare domain. Computer Science Review, 48, p.100546. Available at: https://www.sciencedirect.com/science/article/pii/S1574013723000138 [Accessed 7 April 2004].
3. Chen, R.J., Lu, M.Y., Chen, T.Y. et al., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5, pp.493-497. Available at: https://doi.org/10.1038/s41551-021-00751-8 [Accessed 15 February 2004].
4. Arora, A., 2023. Synthetic data: the future of open-access health-care datasets? The Lancet, 401(10381), p.997. Available at: https://doi.org/10.1016/S0140-6736(23)00324-0 [Accessed 12 May 2004].

5. Kamińska, J., 2022. How healthcare enterprises can benefit from synthetic health data, including 3 practical use-cases. Statice. Available at: https://www.statice.ai/post/healthcare-synthetic-health-data-3-use-cases [Accessed 19 January 2004].
6. Shaip, 2024. Synthetic data in healthcare: Definition, Benefits, and Challenges. Shaip. Available at: https://www.shaip.com/blog/synthetic-data-in-healthcare/ [Accessed 28 June 2004].
7. Agarwal, S., 2024. Navigating The Potential And Perils Of Synthetic Data In Healthcare. Forbes. Available at: https://www.forbes.com/sites/shashankagarwal/2024/01/27/navigating-the-potential-and-perils-of-synthetic-data-in-healthcare/ [Accessed 5 February 2004].
8. Madden, B., 2023. Synthetic Data in Healthcare: the Great Data Unlock. Hospitalogy. Available at: https://hospitalogy.com/articles/2023-11-02/synthetic-data-in-healthcare-great-data-unlock/ [Accessed 22 March 2004].
9. Giuffrè, M. and Shung, D.L., 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digital Medicine, 6, p.186. Available at: https://doi.org/10.1038/s41746-023-00927-3 [Accessed 4 June 2004].
10. Tang, Z., Li, P., Zeng, Y., Zhang, Y., Zheng, Y., Li, H., Wei, Q. and Xiao, Z., 2021. Progress in Computer-Aided Tongue Image Analysis Diagnosis [Original title in Chinese: 计算机辅助舌象分析诊断的研究进展]. biomedRxiv. Available at: http://www.biomedrxiv.org.cn/article/doi/bmr.202106.00006 [Accessed 29 April 2004].
11. Chen, J.J., Yeh, S.Y. and Chiang, J.Y., Feature Extraction For The Automatic Tongue Diagnosis [Original title in Chinese: 中醫舌診電腦化之特徵擷取方法]. China Medical College Hospital and National Sun Yat-Sen University. Available at: https://image.cse.nsysu.edu.tw/ServerPaper/24/%A4%A4%C2%E5%A6%DE%B6E%B9q%B8%A3%A4%C6%A4%A7%AFS%BCx%C2%5E%A8%FA%A4%E8%AAk.pdf [Accessed 13 May 2004].
12. Hui, S.C., He, Y. and Thach, D.T.C., 2008. Machine learning for tongue diagnosis. In: Information, Communications & Signal Processing, 2007 6th International Conference on. IEEE Xplore. Available at: https://doi.org/10.1109/ICICS.2007.4449631 [Accessed 8 March 2004].
13. Xue, W. and Huang, S., 2001. Objective Study on Traditional Chinese Medicine Tongue Diagnosis [Original title in Chinese: 中医舌诊客观化研究]. Engineering Science, 3(1). Available at: https://www.engineering.org.cn/ch/article/21120/detail [Accessed 17 February 2004].
14. Weng, W. and Cao, Y., 1995. Development and Application of a True Color Image System for Traditional Chinese Medicine Tongue Image [Original title in Chinese: 中医舌象真彩色图像系统的研制与应用]. Journal of Practical Traditional Chinese and Western Medicine, (4), p.165. [Accessed 22 January 2004].

Ephraim Ferreira Medeiros Ephraim Ferreira Medeiros is an Acupuncturist, Researcher in Chinese Medicine, certified UX Designer by Google, and Interaction Designer from Michigan University. He is also a Junior level certified Machine Learning Specialist by Deeplearning.ai and webmaster and founder of the site Acupunturabrasil.org. Ephraim is an affiliated researcher at the Center for the Study of Acupuncture and Alternative Therapies (CEATA).

Korean

인공 지능을 사용한 중국 의학에서의 혀 색상 분석을 위한 합성 데이터 생성

Ephraim Ferreira Medeiros 침술 및 대체 요법 연구 센터 **(Ceata)**
webmaster@acupunturabrasil.org

요약 이 예비 연구는 혀 색상 분석을 위한 합성 데이터의 생성 및 응용을 탐구하며, 중국 의학의 혀 본체 및 코팅의 다양한 범주에 중점을 둡니다. 우리는 중국의 약 1,000명의 환자 데이터를 참조로 사용하여 각 범주에서 1,000,000개의 새로운 색상 샘플로 확장하여 총 1,700만 개의 데이터 포인트를 생성했습니다. 주요 목표는 원래 문서화된 색상 변형의 세부 시각화를 생성하여 향후 참조를 위한 견고한 시각적 도구를 제공하는 것이었습니다. 이러한 데이터를 사용하여 기계 학습 모델을 훈련하는 것은 부가적인 가능성이지만 이 작업의 주요 목적은 아닙니다.

서론 인공지능(AI) 및 기계 학습의 발전은 임상 데이터 분석에서 중요한 혁신을 가져오며 현대 의학을 혁신했습니다. 우리의 연구는 중국 의학에서 전통적으로 사용되는 혀 색상 분석을 위해 합성 데이터를 생성하고 응용하는 데 중점을 둡니다. 이 프로젝트를 위해 데이터 조작 및 시각화에 필수적인 NumPy 및 Matplotlib 라이브러리를 사용했습니다. NVIDIA® Tesla® P100 GPU를 사용하여 데이터 처리 및 AI 모델 훈련을 가속화하는 것이 중요했으며, 이 프로세스는 이 구성으로 약 20분이 소요됩니다. 스크립트는 Google Colab과 같은 무료 환경에서 실행되도록 조정할 수 있으며, 샘플 수를 줄여 실행 시간과 필요한 리소스를 줄일 수 있습니다. Google Colab은 무료 GPU 환경을 제공하지만 NVIDIA Tesla P100에 비해 사양이 제한적입니다. 아래 지침에 따라 스크립트를 Google Colab의 무료 티어에서 실행하도록 조정할 수 있습니다:

1. 샘플 수 줄이기: num_samples 매개변수를 100,000과 같이 낮은 값으로 조정하여 스크립트가 Google Colab의 시간 및 리소스 제한 내에서 실행될 수 있도록 합니다.
2. 환경 설정: Google Colab에서 GPU 사용을 무료로 활성화할 수 있습니다:
    - 메뉴에서 편집 > 노트북 설정으로 이동합니다.
    - 하드웨어 가속 탭에서 GPU를 하드웨어 유형으로 선택합니다.
3. 새로운 값 제안:
    - num_samples를 100,000으로 줄임
    - 샘플 플로팅을 위한 sample_indices를 10,000으로 줄임

합성 데이터 생성 합성 데이터는 환자 프라이버시를 보호하고 연구에 사용할 수 있는 데이터 세트를 확장하는 안전하고 효과적인 대안으로 입증되었습니다. 우리의 연구에서는 numpy 및 matplotlib 라이브러리를 사용하여 합성 데이터를 생성하고 시각화했습니다. 사용된 기술은 인간 연구에서 문서화된 혀 색상의 RGB 값에 제어된 변동성을 추가하는 것이었습니다. 이를 통해 원래 1,000명의 환자 샘플을 각 범주에서 1,000,000개의 색상 샘플로 확장하여 보다 견고하고 일반화 가능한 분석을 가능하게 했습니다.

합성 데이터의 시각화 및 분석 우리는 제어된 변동성으로 혀 본체 및 코팅 범주에 대한 합성 데이터를 생성했습니다. 기본 색상과 RGB의 변동성을 사용하여 세부적인 시각화를 생성했습니다.

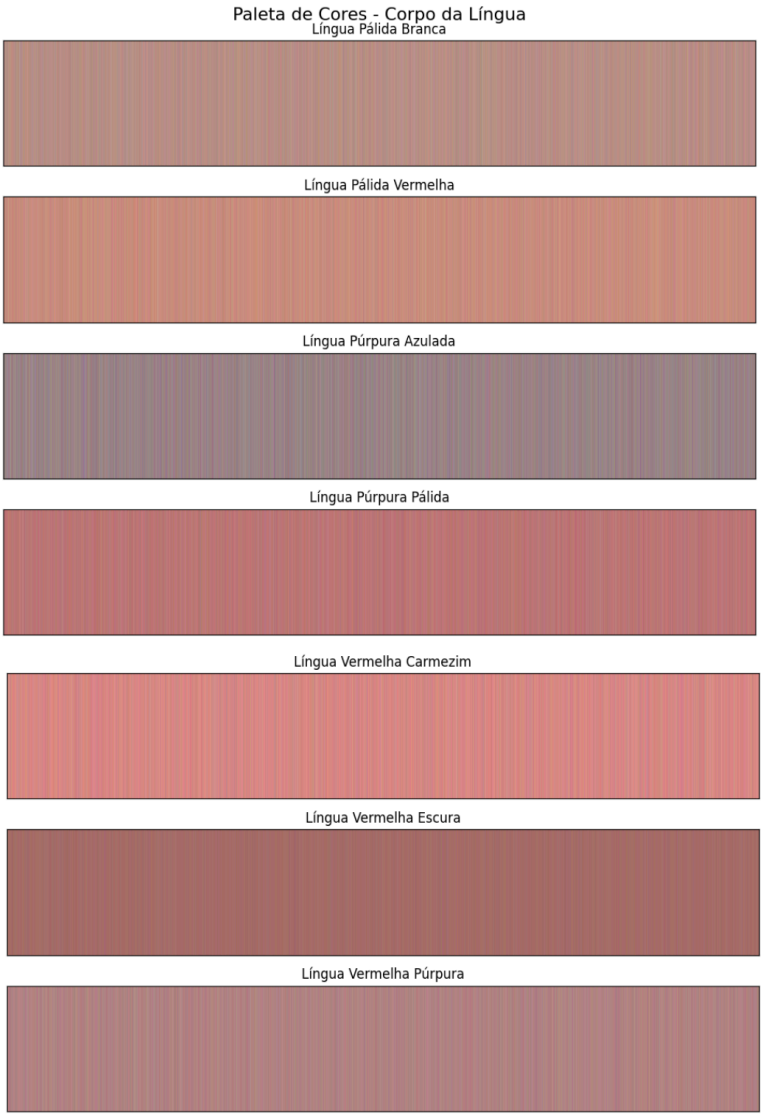
기본 색상의 정의 우리는 2001년 Engineering Science 저널에 발표된 "Objective Study on Traditional Chinese Medicine Tongue Diagnosis" 참조 논문에서 제공한 기본 색상을 사용했습니다. 이 색상은 혀 본체 및 코팅의 다양한 색상 범주에 대해 평균값과 표준 편차로 정의되었습니다. 혀 본체 색상

- 창백한 흰색 혀: 기본: (188, 147, 138); 변동성: (17.2, 12.5, 15.4)
- 창백한 빨간색 혀: 기본: (205, 145, 130); 변동성: (12.9, 15.4, 14.4)
- 진홍색 혀: 기본: (220, 140, 137); 변동성: (20.5, 14.9, 16.8)
- 짙은 빨간색 혀: 기본: (169, 112, 108); 변동성: (6.8, 7.8, 11.6)
- 보라색 빨간색 혀: 기본: (181, 135, 135); 변동성: (12.4, 14.6, 12.9)
- 창백한 보라색 혀: 기본: (192, 122, 122); 변동성: (14.3, 12.4, 17.1)
- 푸르스름한 보라색 혀: 기본: (158, 135, 138); 변동성: (15.6, 19.8, 19.3)

혀 코팅 색상

- 두꺼운 흰색 코팅: 기본: (234, 224, 217); 변동성: (17.2, 14.1, 13.4)
- 끈적한 두꺼운 흰색 코팅: 기본: (231, 213, 208); 변동성: (15.2, 16.7, 11.2)
- 얇은 흰색 코팅: 기본: (227, 198, 195); 변동성: (15.6, 12.5, 18.5)
- 끈적한 얇은 흰색 코팅: 기본: (223, 190, 190); 변동성: (11.8, 13.7, 10.5)
- 얇은 노란색 코팅: 기본: (214, 179, 158); 변동성: (11.7, 12.8, 9.9)
- 끈적한 얇은 노란색 코팅: 기본: (209, 162, 138); 변동성: (12.1, 14.8, 13.5)
- 두꺼운 노란색 코팅: 기본: (202, 157, 129); 변동성: (16.4, 10.7, 11.3)
- 끈적한 두꺼운 노란색 코팅: 기본: (194, 148, 129); 변동성: (16.7, 11.5, 14.2)
- 노란색 코팅: 기본: (182, 133, 102); 변동성: (18.4, 15.5, 22.1)
- 회색 검은색 코팅: 기본: (155, 103, 74); 변동성: (22.9, 21.7, 23.1)

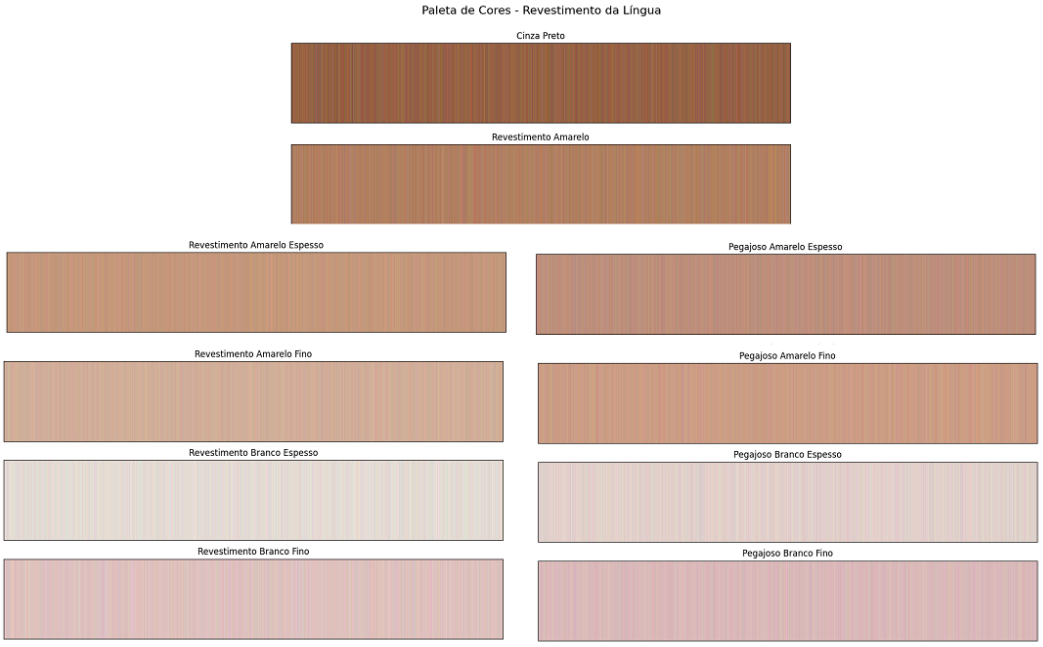
결과 및 토론



색상 팔레트 - 혀 본체

1. 창백한 흰색 혀
2. 창백한 빨간색 혀
3. 푸르스름한 보라색 혀
4. 창백한 보라색 혀
5. 진홍색 혀
6. 짙은 빨간색 혀
7. 보라색 빨간색 혀

Tongue Coating 혀 색상 팔레트 **-** 혀 코팅



- 상단 (두 항목):
  - 왼쪽: 회색 검은색
  - 오른쪽: 노란색 코팅
- 중간 (네 항목):
  - 왼쪽: 두꺼운 노란색 코팅
  - 오른쪽: 끈적한 두꺼운 노란색 코팅
  - 왼쪽: 얇은 노란색 코팅
  - 오른쪽: 끈적한 얇은 노란색 코팅
- 하단 (네 항목):
  - 왼쪽: 두꺼운 흰색 코팅
  - 오른쪽: 끈적한 두꺼운 흰색 코팅
  - 왼쪽: 얇은 흰색 코팅
  - 오른쪽: 끈적한 얇은 흰색 코팅

스크립트에서 생성된 총 데이터는 각 RGB 채널을 고려하여 5,100만 개의 데이터 포인트에 달합니다. 각 데이터 포인트는 환자에 해당할 수 있으며, 이는 콜롬비아나 대한민국과 같은 국가의 인구와 비교할 수 있어 생성된 데이터의 규모와 범위를 보여줍니다. 또한, 각 최종 팔레트는 30,000개의 데이터 포인트로 구성되며, 17개의 범주를 고려하면 총 510,000개의 데이터 포인트가 있습니다.

이미지를 주의 깊게 관찰하면 끈적한 코팅과 순수한 코팅의 차이가 미묘합니다. 이러한 차이는 큰 RGB 값 변동보다는 음영과 더 관련이 있을 수 있습니다. "끈적한" 질감은 색상의 시각적 인식에 영향을 미칠 수 있지만, 이러한 변동에 대한 보다 구체적인 처리가 없으면 RGB 코드 시각화만으로 구별하기 어려울 수 있습니다. 끈적한 코팅은 더 어두워 보일 수

있으며, 이는 같은 기본 색상의 더 어두운 음영으로 인식될 수 있습니다. 색상의 강도의 작은 변동은 기본 색상이 동일하더라도 다른 인식을 만들 수 있습니다.

시각적 분석 끈적한 코팅은 약간 더 어두워 보이며 음영 차이를 나타냅니다. 얇은 흰색 코팅과 끈적한 얇은 흰색 코팅의 차이는 더 미묘하지만, 끈적한 코팅은 약간 더 어두워 보일 수 있으며 주요 차이는 색상의 음영에 있습니다.

데이터 증강 RGB만으로 구성된 데이터는 실제 이미지가 아니므로 회전, 반전, 자르기와 같은 많은 전통적인 데이터 증강 기술이 직접적으로 적용되지 않습니다. 따라서 제어된 노이즈 추가 기술이 이 특정 경우에 가장 적합한 기술로 두드러집니다.

제어된 노이즈 추가 선택 이유

1. 색상 보존: RGB 값에 작은 노이즈를 추가함으로써 해당 클래스의 정의된 색상 범위 내에서 데이터를 유지하면서 조명이나 카메라 품질의 차이로 인한 자연 변동을 시뮬레이션할 수 있습니다.
2. 모델 강건성: 이러한 작은 변동을 도입하면 모델이 입력의 미세한 변화를 더욱 강건하게 처리할 수 있어, 실제 응용 프로그램에서 이러한 변동이 흔한 경우 모델을 더욱 잘 준비시킬 수 있습니다.

고려할 기타 기술 노이즈 추가는 주요 기술이며 매우 유용하지만 특정 필요에 따라 다른 접근 방식을 고려할 수 있습니다:

- 밝기 조정: 사각형의 밝기를 변경하여 다른 조명 조건을 시뮬레이션할 수 있습니다. 이것은 원래 색상을 여전히 나타내는 한도 내에서 RGB 값을 균일하게 조정하여 수행할 수 있습니다.

합성 데이터 생성의 대안 및 표준 우리의 접근 방식은 RGB 코드 생성에 중점을 둔 연구 목표에 특히 적합합니다. 이 선택은 간단하고 효율적이라는 점에서 강화됩니다. 제어된 변동성을 통해 합성 RGB 코드를 생성하는 것은 빠르고 정확하게 대량의 데이터를 생성할 수 있는 견고하고 잘 정립된 방법입니다. 합성 이미지 생성에 있어서는 생성적 적대 신경망(GAN)과 같은 다른 기술이 장점을 제공할 수 있지만, 더 큰 복잡성과 컴퓨팅 요구 사항도 도입됩니다. 우리의 연구 본질을 고려할 때, 선택한 접근 방식은 적절할 뿐만 아니라 연구의 특정 요구를 충족시키는 데 있어 효과적입니다.

의학에서 합성 데이터 생성의 중요성 의학에서 합성 데이터 생성은 여러 가지 이유로 중요합니다:

- 프라이버시 및 보안: 합성 데이터는 연구자가 환자의 프라이버시를 침해하지 않고 건강 정보를 사용할 수 있게 합니다. 이는 민감한 데이터 보호가 필수적인 분야에서 중요합니다.
- 샘플 확장: 합성 데이터는 실제 데이터가 제한된 상황에서 데이터 세트를 확장할 수 있게 합니다. 이는 기계 학습 모델의 강건성을 향상시키고 더 정확한 통계 분석을 수행하는 데 필수적입니다.
- 시뮬레이션 및 예측: 합성 데이터는 새로운 치료법 개발 및 건강 정책 평가에 있어 고가의 시간 소모적 임상 시험 없이 통제된 환경에서 시뮬레이션 및 예측을 수행하는 데 유용합니다.

- **AI** 모델 훈련: 합성 데이터 사용은 인공지능 모델 훈련에 필수적이며, 이러한 모델이 실제 상황에서 잘 일반화하고 작은 데이터 세트에서 볼 수 없는 패턴을 감지할 수 있도록 보장합니다.

AI를 사용한 초기 패턴 감지의 중요성 RGB 값이 그렇게 지배적이지 않지만 존재하는 색상과 같은 미묘한 차이를 포착하는 것은 눈에 보이는 증상이 나타나기 전에도 혀 분석에서 초기 패턴을 감지하는 데 중요할 수 있습니다. 이 개념은 중국 의학의 기본 원칙인 **"治未病"** (Zhì wèi bìng), 즉 "병이 생기기 전에 치료하다"와 일치합니다. 동일한 접근 방식은 혀의 특정 영역, 혀 전체 또는 코팅 패턴의 색상이 정상 표준으로 돌아올 때 예후에도 적용됩니다. 이 맥락에서 환자의 이미지 기록은 프로세스를 용이하게 하여 AI가 이미지를 분석하고 패턴을 분류하는 데 있어 신뢰할 수 있는 도우미로서의 기능을 향상시킬 수 있습니다.

초기 감지 및 그 혜택

1. 미묘한 이상 감지: 혀 색상의 미묘한 변동은 명백한 증상이 나타나지 않은 건강 문제를 나타낼 수 있습니다. 이러한 미묘한 차이를 식별하면 초기 개입이 가능하여 더 심각한 질병의 발병을 예방할 수 있습니다.
2. 정확한 분석: 인공지능(AI) 기술을 사용하면 대량의 데이터를 분석하고 인간의 눈에 보이지 않는 패턴을 식별할 수 있습니다. AI는 혀 색상의 미세한 변화를 감지하여 진단의 더 정확한 기반을 제공합니다.
3. 조기 개입: 초기 감지는 증상이 나타나기 전 초기 단계에서 개입이 가능하도록 하여 치료 결과를 크게 개선하고 합병증 위험을 줄일 수 있습니다.
4. 전통 원칙과의 정렬: **"治未病"** (Zhì wèi bìng)은 예방과 건강 유지에 중점을 둔 중국 전통 의학의 고급 원칙입니다. 이 원칙에 AI를 통합하면 오래된 입증된 개념에 대한 현대적이고 진보된 접근 방식을 가능하게 하여 예방 치료의 가능성을 확장할 수 있습니다.

미래 응용 및 프로젝트의 두 번째 단계 이 연구는 합성 데이터를 대규모로 사용하여 합성 신경망(CNN)을 사용하여 기계 학습 모델을 훈련하고 개선하는 길을 열어줍니다. 이 응용 프로그램은 프로젝트의 두 번째 실험 단계로, 혀 색상 분석에 기반한 진단의 정확도와 효율성을 향상시키기 위해 고급 AI를 통합할 수 있습니다.

중국 의학 연구를 위한 합성 데이터 프로젝트의 가속화에서 **LLM** 및 생성 **AI**의 역할 이 프로젝트에서는 사람이 수작업으로 작성한 코드가 없습니다. 모든 코드는 프롬프트 엔지니어링 기술을 통해 얻어졌습니다. 이 혁신적인 방법은 고급 AI 도구가 소프트웨어 개발을 어떻게 변형시킬 수 있는지를 보여주며, 이전에 광범위한 인간 개입이 필요했던 프로세스를 자동화합니다. 코드 처리 및 실행은 Kaggle 플랫폼에서 수행되었으며, Kaggle은 연구원 및 개발자가 높은 비용 없이 강력한 컴퓨팅 리소스를 활용할 수 있는 관대한 무료 플랜을 제공합니다. Kaggle은 대량의 데이터 및 복잡한 모델과 함께 작업할 수 있는 접근 가능하고 효율적인 환경을 제공하는 것으로 널리 인정받고 있습니다. 이 연구는 합성 데이터를 대규모로 사용하여 합성 신경망(CNN)을 사용하여 기계 학습 모델을 훈련하고 개선하는 길을 열어줍니다. 이 접근 방식은 프로젝트의 두 번째 실험 단계로 계획되어 있으며, 예측 모델의 정확도가 더욱 향상될 것으로 예상됩니다. 초보 기계 학습 전문가도 몇 시간 내에 아이디어를 테스트, 프로토타입 및 인큐베이션하고 기본 수준의 프로젝트를 완료할 수 있습니다. 이는 경험 부족으로 인해 몇 달이 걸리던 작업을 단축시킵니다. 생성 AI 도구와 접근 가능한 컴퓨팅 플랫폼은 다양한 과학 분야의 더 많은 개인이 프로그래밍

배경이 거의 없어도 AI 분야에서 중요한 기술적 및 과학적 발전에 더 직접적으로 기여할 수 있도록 민주화합니다.

기술 설명 프롬프트 엔지니어링 및 코드 설계 고급 아이디어와 코드 구조 설계는 저자가 구상했지만, 실질적인 구현은 GPT-4o(미국) 및 DeepSeek-Coder-V2(중국)라는 생성 인공지능을 사용하여 수행되었습니다. 대형 언어 모델(LLM)은 방대한 양의 텍스트로 학습된 AI 모델로, 인상적인 방식으로 자연어를 이해하고 생성할 수 있습니다. OpenAI에서 개발한 GPT-4o는 복잡한 코드를 상세한 설명에서 생성하는 능력으로 잘 알려져 있습니다. 중국 AI인 DeepSeek-Coder-V2는 기술 문제를 해결하고 효율적인 프로그래밍 솔루션을 제공하는 데 중점을 둡니다. 이 프로젝트에서는 이 두 LLM을 모두 사용하여 필요한 모든 코드를 생성했으며, 이들의 효율성을 입증하고 더 빠르고 정확한 구현에 기여했습니다. LLM은 상세한 설명과 정의된 목표에서 코드를 생성하는 데 도움을 주어 시간 절약 및 인간 오류를 줄여줍니다. 이들은 기술 문제를 신속하게 해결하고 코드 작성 프로세스의 일부를 자동화하여 개발자가 고급 논리와 프로젝트 혁신에 더 집중할 수 있도록 하여 생산성을 향상시킵니다. 이 기사에서 생성된 모든 코드는 GPT-4o 및 DeepSeek-Coder-V2를 사용하여 생성되었으며, 이러한 기술이 중국 의학 연구를 위한 합성 데이터 프로젝트를 얼마나 크게 가속화할 수 있는지 보여줍니다.

인공지능을 사용한 설계 및 구현 **-** 인공지능 지원 개발 방법론 개요 초기에는 저자가 프로젝트 목표를 정의하고 이를 달성하기 위해 필요한 논리를 개요합니다. 이 과정에는 AI가 필요한 코드를 생성하는 데 도움이 되는 상세한 프롬프트를 작성하는 것이 포함됩니다.

1. 목표 및 구조 정의: 저자는 합성 데이터 생성, RGB 값 변동성 및 색상 팔레트 시각화를 포함하여 코드의 목표를 명확하게 설정합니다. 이 단계는 모든 기능 및 비기능 요구 사항이 잘 이해되고 문서화되었음을 보장하여 소프트웨어 엔지니어링 관행을 준수합니다.
2. 특정 프롬프트 생성: 프롬프트 엔지니어링 기술을 사용하여 저자는 코드의 각 단계를 구체화하는 상세한 지침을 작성합니다. 그런 다음 이러한 프롬프트는 GPT-4o 및 DeepSeek-Coder-V2 AI에 의해 처리되어 해당 코드가 생성됩니다. 이 접근 방식은 프로그래밍 언어와 코딩에 대한 기본 지식만으로도 코드 개발을 가능하게 합니다.
3. 코드 생성 및 정제: AI는 제공된 프롬프트를 기반으로 코드를 생성합니다. 저자는 생성된 코드를 검토하고 필요한 조정을 수행하며 프롬프트를 재정의하여 코드의 정확성과 기능을 개선합니다. 이 검토 및 정제 주기는 코드 품질을 보장하기 위해 필수적이며, 지속적 통합(CI) 및 지속적 개선(CI) 관행에 부합합니다.
4. 실행 및 검증: 생성된 코드는 Kaggle 플랫폼에서 실행됩니다. 정의된 목표를 충족하고 기대한 대로 작동하는지 확인하기 위해 검증이 수행됩니다. 이 프로세스에는 자동화된 테스트 및 수동 테스트를 포함하여 요구 사항 준수를 확인하는 검증 및 검증(V&V) 관행을 따릅니다.
5. 반복 및 개선: 피드백 및 실행 결과를 기반으로, 이상적인 솔루션을 달성하기 위해 프롬프트를 정제하고 코드를 조정하는 과정을 여러 번 반복할 수 있습니다. 이 반복적 접근 방식은 지속적인 피드백이 최종 제품을 개선하는 데 사용되는 민첩한 개발 주기와 일치합니다.

이 방법론은 프로젝트 개발을 가속화할 뿐만 아니라 프로그래밍 언어와 코딩에 대한 기본 지식을 가진 개발자도 높은 품질의 프로젝트를 수행할 수 있도록 합니다. 채택된 관행은

생성된 코드가 반복적 설계, 지속적 통합 및 지속적 개선을 포함한 최고의 소프트웨어 개발 및 기계 학습 관행과 일치하도록 보장합니다.

사용된 코드가 포함된 노트북 링크 노트북 링크

참고 문헌

1. Kennedy, S., 2024. Weighing the pros and cons of synthetic healthcare data use. Health Tech Analytics. Available at: https://www.techtarget.com/healthtechanalytics/feature/Weighing-the-pros-and-cons-of-synthetic-healthcare-data-use [Accessed 23 March 2004].
2. Murtaza, H., Ahmed, M., Khan, N.F., Murtaza, G., Zafar, S. and Bano, A., 2023. Synthetic data generation: State of the art in the healthcare domain. Computer Science Review, 48, p.100546. Available at: https://www.sciencedirect.com/science/article/pii/S1574013723000138 [Accessed 7 April 2004].
3. Chen, R.J., Lu, M.Y., Chen, T.Y. et al., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5, pp.493-497. Available at: https://doi.org/10.1038/s41551-021-00751-8 [Accessed 15 February 2004].
4. Arora, A., 2023. Synthetic data: the future of open-access health-care datasets? The Lancet, 401(10381), p.997. Available at: https://doi.org/10.1016/S0140-6736(23)00324-0 [Accessed 12 May 2004].
5. Kamińska, J., 2022. How healthcare enterprises can benefit from synthetic health data, including 3 practical use-cases. Statice. Available at: https://www.statice.ai/post/healthcare-synthetic-health-data-3-use-cases [Accessed 19 January 2004].
6. Shaip, 2024. Synthetic data in healthcare: Definition, Benefits, and Challenges. Shaip. Available at: https://www.shaip.com/blog/synthetic-data-in-healthcare/ [Accessed 28 June 2004].
7. Agarwal, S., 2024. Navigating The Potential And Perils Of Synthetic Data In Healthcare. Forbes. Available at: https://www.forbes.com/sites/shashankagarwal/2024/01/27/navigating-the-potential-and-perils-of-synthetic-data-in-healthcare/ [Accessed 5 February 2004].
8. Madden, B., 2023. Synthetic Data in Healthcare: the Great Data Unlock. Hospitalogy. Available at: https://hospitalogy.com/articles/2023-11-02/synthetic-data-in-healthcare-great-data-unlock/ [Accessed 22 March 2004].
9. Giuffrè, M. and Shung, D.L., 2023. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digital Medicine, 6, p.186. Available at: https://doi.org/10.1038/s41746-023-00927-3 [Accessed 4 June 2004].
10. Tang, Z., Li, P., Zeng, Y., Zhang, Y., Zheng, Y., Li, H., Wei, Q. and Xiao, Z., 2021. Progress in Computer-Aided Tongue Image Analysis Diagnosis [Original title in Chinese: 计算机辅助舌象分析诊断的研究进展]. biomedRxiv. Available at: http://www.biomedrxiv.org.cn/article/doi/bmr.202106.00006 [Accessed 29 April 2004].
11. Chen, J.J., Yeh, S.Y. and Chiang, J.Y., Feature Extraction For The Automatic Tongue Diagnosis [Original title in Chinese: 中醫舌診電腦化之特徵擷取方法]. China Medical College Hospital and National Sun Yat-Sen University. Available at: https://image.cse.nsysu.edu.tw/ServerPaper/24/%A4%A4%C2%E5%A6%DE%B6E

%B9q%B8%A3%A4%C6%A4%A7%AFS%BCx%C2%5E%A8%FA%A4%E8%AAk.pdf [Accessed 13 May 2004].
12. Hui, S.C., He, Y. and Thach, D.T.C., 2008. Machine learning for tongue diagnosis. In: Information, Communications & Signal Processing, 2007 6th International Conference on. IEEE Xplore. Available at: https://doi.org/10.1109/ICICS.2007.4449631 [Accessed 8 March 2004].
13. Xue, W. and Huang, S., 2001. Objective Study on Traditional Chinese Medicine Tongue Diagnosis [Original title in Chinese: 中医舌诊客观化研究]. Engineering Science, 3(1). Available at: https://www.engineering.org.cn/ch/article/21120/detail [Accessed 17 February 2004].
14. Weng, W. and Cao, Y., 1995. Development and Application of a True Color Image System for Traditional Chinese Medicine Tongue Image [Original title in Chinese: 中医舌象真彩色图像系统的研制与应用]. Journal of Practical Traditional Chinese and Western Medicine, (4), p.165. [Accessed 22 January 2004].

Ephraim Ferreira Medeiros Ephraim Ferreira Medeiros는 침술사이자 중국 의학 연구원이며, Google 인증 UX 디자이너이자 Michigan University의 Interaction Designer입니다. 또한 Deeplearning.ai에서 주니어 레벨 기계 학습 전문가 인증을 받았으며, Acupunturabrasil.org 사이트의 웹마스터이자 창립자입니다. Ephraim은 침술 및 대체 요법 연구 센터 (CEATA)의 소속 연구원입니다.