國立臺灣師範大學資訊工程學系

國科會大專學生研究計劃書

從音樂生成歌曲封面

Generating Album Covers from Music

指導教授 葉梅珍 教授

學生 陳炫佑 撰

中華民國 114 年 2 月

Abstract

隨著數位音樂市場的發展，獨立音樂創作者面臨封面設計成本高昂的挑戰，而生成式 AI 提供了一種低成本且高效的視覺設計解決方案。然而，目前 AI 生成的圖像品質仍受到輸入提示詞的影響，如何有效設計提示詞 (prompt) 以提升生成結果仍是一個值得探討的問題。雖然音樂與文字皆爲 sequential data，但音樂的語意表達並非跟語言一樣精確，因此本研究將探討 '音樂提示詞'(music prompt) 的清晰度對於圖像生成模型（特別是擴散模型）的影響，並分析如何透過最佳化提示詞設計來提高生成結果的品質與一致性。透過實驗與討論，本研究的成果希望爲音樂視覺化的 AI 生成技術提供更具體的指導方針，並促進音樂創作者與 AI 技術之間的互動與創新。

一、研究動機與研究問題

應用方面

隨著數位音樂平台的興起，音樂創作已不再局限於大型唱片公司或專業錄音室，許多具有創意與才華的創作者，特別是學生或初學者，皆能夠透過這些平台輕鬆發布自己的音樂作品。然而這些獨立創作者往往面臨資源有限的困境，使得他們無法負擔專業的視覺設計成本，特別是在音樂專輯或單曲封面設計方面。

封面設計不僅僅是音樂作品的視覺呈現之一，更是潛在聽眾對該作品的第一印象。在當今的數位時代，視覺吸引力對於內容的傳播與影響力至關重要。一個吸引人的封面能夠提高點擊率與曝光度，而缺乏視覺吸引力的封面可能導致優質音樂作品被忽略。因此如何讓資源有限的創作者獲得高品質且具創意的封面設計，是值得關注的問題。

然而對於許多獨立創作者而言，聘請專業設計師往往需要一筆不小的開支，這使得他們在視覺設計上處於劣勢。即便是一些富有創意的學生，由於缺乏設計經驗或工具，也難以製作出能夠吸引目光的封面作品，進而影響作品的曝光度與市場競爭力。在數位音樂市場競爭日益激烈的背景下，如何突破這樣的困境，成爲許多創作者關心的議題。

近年來生成式人工智慧（Generative AI）的興起爲這類問題提供了一個創新的解決方案。生成式 AI 有潛力能夠依照簡單的文字描述或基本設計要求，自動生成高質量的封面設計，甚至能夠爲整張專輯提供一致性的視覺風格。這些 AI 設計工具操作簡單，使用者無需具備專業設計能力，只需提供關鍵字或概念，系統即可自動生成具視覺衝擊力的圖像，爲創作者提供更具彈性的設計選擇。

舉例來說一位學生創作者可以透過生成式 AI，在數分鐘內產生多種風格的封面設計，並從中挑選最符合其音樂風格的版本，無需仰賴專業設計師，也不必花費過多時間與金錢。這不僅能夠降低設計成本，也能夠爲創作者帶來更多元的視覺表現方式，提高音樂作品的市場競爭力與曝光率。

然而，大多數的影像生成工具需要使用者提供文字提示，而將音樂轉換成文字 prompt 並不一定容易。因此本計畫的研究目標爲輸入一個歌曲 (可能包含歌詞以及音檔)，能自動生成適當的封面圖像，如圖1所示。

技術方面

在技術層面上，根據一系列試驗與研究，發現了幾個關鍵問題與技術挑戰，這些因素可能影響系統的效能、可行性以及適用性，進而影響生成式 AI 在音樂封面設計領域的實際應用。

1. **WAV** 檔案過大，資料集佔用大量儲存空間。
   由於 WAV 格式爲未壓縮的音訊檔案，其容量遠大於其他常見的壓縮格式（如 MP3 或 FLAC）。當處理大量音訊數據時，這不僅會對儲存空間造成壓力，也可能影響數據讀取與處理的效率。在進行大規模訓練或推論時，必須考量更有效的數據管理策略，例

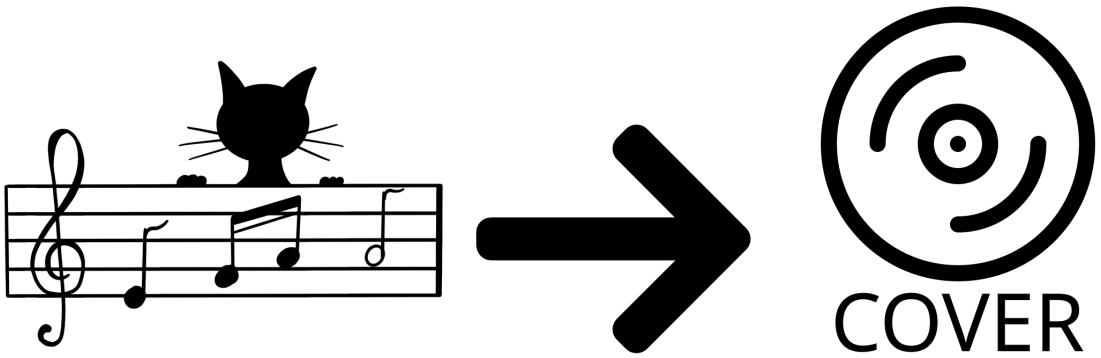

Figure 1: 本研究計畫的目標：從歌曲生成封面。

如使用更高效的存儲格式、採用動態壓縮技術，或是透過資料流（streaming）方式來讀取數據，以減少記憶體與儲存空間的負擔。

2. **QWen**、**LLaMA** 等大型語言模型（**LLM**）過於龐大，無法與其他模型一同放入 **VRAM** 中。
   目前許多先進的 LLM，例如 QWen [14] 與 LLaMA [13]，參數規模動輒數十億至數千億，推論時的 VRAM 需求極高。如果需要將 LLM 與其他模型（如擴散模型（Diffusion Model）或語音合成模型）結合使用，可能會因顯存不足而無法同時運行，導致系統無法發揮最佳效能。可能的解決方案包括模型裁剪（pruning）、量化（quantization），或透過混合精度（mixed precision）來降低 VRAM 佔用。此外，也可考慮透過分散式計算與記憶體優化技術來提升計算效率。

3. 是否所有的 **sequential data** 都可以以文字的形式處理，並將其 **embed** 後輸入至 **diffusion model**？
   這是一個關鍵的研究問題。雖然許多序列數據（如語音、股價、時間序列等）可以轉換爲嵌入（embedding），但這是否適用於所有 sequential data 仍需進一步驗證。例如，語音數據的時序關聯性與語言文字的關聯性不同，可能會影響 diffusion model 對其進行建模的能力。因此，可能需要設計專門的 embedding 方法，使其能夠更有效地保留原始數據的特性。

4. **prompt** 的精確程度是否與 **diffusion model** 生成的結果有相關性？
   目前許多 diffusion model（如 Stable Diffusion）依賴文本提示（prompt）來引導生成過程。然而，prompt 的細節程度、語意準確性與模型訓練時的對齊程度，可能都會影響最終的輸出結果。例如，若 prompt 過於模糊，模型可能難以產生符合預期的內容；相反，若 prompt 過於精確，則可能會限制模型的創造力。因此，如何設計符合音樂描述且高效的 prompt，使其能夠既準確引導生成結果，又能保留一定的靈活性，是值得深入研究的議題。

綜上所述，這些技術問題與限制在實踐中可能會影響系統的表現，需要透過各種優化方法與技術創新來解決，以提高整體效能與可用性。

二、文獻回顧與探討

近年來，深度學習技術在音樂生成與理解領域取得了長足進展，推動了旋律創作、音樂風格轉換、歌詞生成與音樂視覺化等多種應用。特別是基於大型語言模型（LLMs）的音樂生成系統，如 MusicBERT 和 MusicAgent，在旋律補全、伴奏建議與風格分類等任務上展現了優異的表現。

在多模態學習（Multimodal Learning）領域，CLIP（Contrastive Language-Image Pretraining）透過對比學習統一圖像與文本嵌入空間，爲音樂與視覺的跨模態結合提供了技術支撐。此外，擴散模型（Diffusion Models, DMs）作爲新興的生成方法，在影像合成方面展現了卓越的性能，特別是潛在擴散模型（Latent Diffusion Models, LDMs）進一步降低了計算需求，使得影像生成更加高效。

本節將回顧與本研究相關的生成模型、多模態學習及音樂理解技術，探討它們在不同應用場景中的發展與挑戰，並分析這些技術如何爲音樂視覺封面生成提供理論與技術基礎。

影像生成

高解析度圖像合成（High-Resolution Image Synthesis）是計算機視覺領域的一大挑戰。傳統方法如生成對抗網絡（GANs）雖能生成高品質圖像，但易遭遇模式崩潰（mode collapse）與訓練不穩定等問題。基於自回歸（autoregressive）方法的模型，如 Transformer-based 生成模型，雖具有良好的密度估計能力，但計算成本極高。

近年來，擴散模型（Diffusion Models, DMs）在影像生成領域取得突破，透過逐步去噪（denoising process）來學習數據分佈。然而，傳統 DMs 直接在像素空間運行，導致訓練和推理成本高昂。Latent Diffusion Models (LDMs)（[1]）提出將擴散過程轉移至預訓練的自編碼器（autoencoder）潛在空間，以降低計算需求，同時保持圖像細節。該方法利用交叉注意力（cross-attention）機制來控制圖像生成，支援文本到圖像生成（text-to-image generation）、圖像修補（inpainting）與超解析度（super-resolution）等任務。

實驗顯示，LDMs 在 ImageNet、CelebA-HQ、LSUN-Churches 等數據集上達到與最先進模型相媲美的生成品質，並顯著降低計算資源需求。特別是在文本到圖像合成（text-to-image synthesis）上，LDMs 能以較少參數量達到與 DALL·E、GLIDE 相近的性能，驗證其在多模態生成上的潛力。未來研究可探索 LDMs 在更大規模數據集上的適應性，或與 Transformer 結合以進一步提升生成質量與控制能力。擴散模型相較於傳統的生成對抗網絡（**GAN**）與自回歸模型（**Autoregressive Models**），在影像品質與多樣性上展現更卓越的性能。（見 [7]）

擴散模型的發展與主要 **T2I** 模型

擴散模型的發展歷程涵蓋了 DDPM（Denoising Diffusion Probabilistic Models）及其改進方法。代表性的 T2I 擴散模型包括：

- **GLIDE**（2021）：使用 *Classifier-Free Guidance* 提高影像品質。
- **Imagen**（2022）：採用大型語言模型作爲文本編碼器，提升文本與影像的一致性。
- **Stable Diffusion**（2022）：透過潛在空間（*Latent Space*）進行擴散，大幅減少計算資源需求。
- **DALL-E 2**（2022）：結合 CLIP 編碼器與擴散模型，提升語義理解能力。

技術改進與應用

論文探討了擴散模型的多項改進方法，包括：

- 模型優化：如 *DiT*（*Diffusion Transformer*）提升架構效能，*Free-U* 增強 U-Net 特徵學習能力。

- 靈活控制：如 *ControlNet* 允許用户透過邊緣檢測或姿態控制生成內容。
- 應用拓展：擴散模型已應用於文本到影片（Sora, Stable Video Diffusion）、文本到 *3D*（Magic3D）、圖像編輯（InstructPix2Pix）等領域。

挑戰與未來發展

論文探討了擴散模型的倫理與安全挑戰，如數據偏差可能導致刻板印象，生成內容可能被濫用（如僞造影像）。未來的發展方向包括：

- 設計更安全與公平的模型，減少倫理風險。
- 降低計算成本，讓擴散模型更易於部署。
- 多模態學習，結合影像與語言模型，建立統一的 AI 框架。

MusicAgent: An AI Agent for Music Understanding and Generation with Large Language Models 是由 D. Yu 等人於 2023 年發表的研究，提出了一個創新的基於大型語言模型（LLM）的 AI 系統——MusicAgent，旨在推動音樂理解與生成的研究。這個系統不僅能夠理解和生成音樂，還能夠在多種音樂任務中提供高效的支持，從旋律創作到音樂風格轉換，再到歌詞生成等。MusicAgent 結合了語言模型的強大推理能力和音樂領域的專業知識，實現了多模態音樂生成，進一步拓展了 LLM 在音樂領域的應用（見 [5]）。

此外最新的研究也開始探索生成式技術在音樂視覺化上的應用，例如文本到影片生成的嘗試 [6]。

多模態模型

CLIP（Contrastive Language-Image Pretraining）由 OpenAI 提出，透過對比學習訓練圖像與文本編碼器，在 4 億組（image, text）配對數據上學習統一的多模態嵌入空間。CLIP 能透過自然語言進行零樣本學習（zero-shot learning），在多種視覺任務中展現強大泛化能力，甚至可匹敵 ResNet-50 在 ImageNet 上的表現。相比傳統監督學習，CLIP 更具遷移能力與魯棒性，但在細粒度分類等任務上仍有限制。未來可探索更高效的訓練方法與多模態結合（見 [3]）。

MusicBERT（[2]）針對符號音樂理解提出大規模預訓練模型。該模型採用 OctupleMIDI 編碼，將音符表示爲八元組（包含時間簽名、節奏、樂器、音高等），使音樂序列比傳統 REMI 編碼（Huang & Yang, 2020）更短、更具通用性。作者設計小節級遮罩（bar-level masking）策略，在預訓練過程中避免資訊洩漏，提升模型的音樂表示能力。

爲了有效訓練 MusicBERT，研究者建構了 Million MIDI Dataset (MMD)，包含超過 150 萬首音樂作品，遠超過先前的 Lakh MIDI Dataset（Raffel, 2016）。實驗結果顯示，MusicBERT 在旋律補全、伴奏建議、風格分類等音樂理解任務上均超越傳統方法，如 PiRhDy（Liang et al., 2020）與 Melody2Vec（Hirai & Sawada, 2019）。進一步的消融研究證實，OctupleMIDI 編碼與小節級遮罩策略顯著提升了模型表現。此外，與未經預訓練的模型相比，MusicBERT 顯示出強大的遷移學習能力，證明大規模符號音樂預訓練的價值。

相關工具

fairseq: A Fast, Extensible Toolkit for Sequence Modeling 是由 M. Ott 等人於 2019 年在 NAACL-HLT 會議上發表的文章，介紹了一個開源且高效的序列建模工具——fairseq。這個工具庫是由 Facebook AI Research（FAIR）開發，旨在爲自然語言處理（NLP）領域的各類任務提供靈活且高效的支持，特別是機器翻譯、語言建模、文本生成等。fairseq 基於 PyTorch 框架，並充分利用了該框架的優勢，提供了易於擴展和高效訓練的功能。

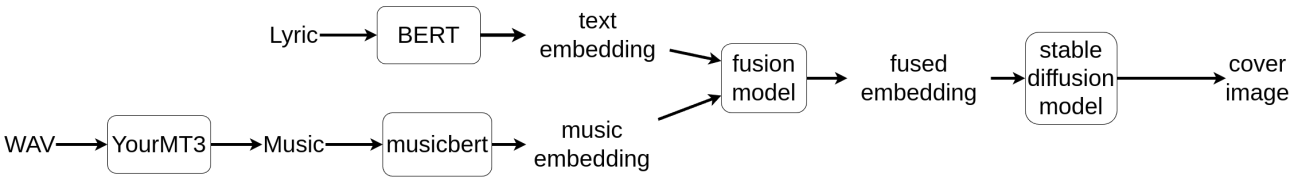

Figure 2: 模型架構：包括從音樂與歌詞計算 embeddings 與生成圖像兩個部分。

fairseq 的設計特點包括高度模組化的架構，能夠輕鬆實現各種新模型或算法的整合，支援多種深度學習架構，譬如 Transformer、LSTM、CNN 等，這些架構廣泛應用於序列建模任務。該工具庫特別注重計算效率，支援混合精度運算（Mixed Precision Training）和動態批量處理（Dynamic Batching），能夠加速模型的訓練過程。此外，fairseq 支援多 GPU 和分散式訓練，使其能夠處理大規模數據集並加快模型訓練速度（參見 [4]）。

綜合上述文獻，可以看出目前生成模型、擴散模型、多模態學習及音樂理解與生成的前沿技術，均爲本研究提供了堅實的理論與技術基礎。本研究旨在利用這些先進技術，將歌詞與音樂特徵融合，生成能夠準確體現音樂情感與意境的視覺封面。透過借鑒 LDMs 的生成策略、CLIP 的多模態對齊、以及 MusicBERT 和 MusicAgent 在音樂理解與生成上的成功經驗，本研究期望構建一個低成本且高效率的封面生成系統，從而彌補資源有限的音樂創作者在視覺呈現上的不足。

四、研究方法及步驟

本研究分爲多個階段，首先聚焦於歌詞的數據分析與音樂封面生成，以驗證研究方法的可行性。隨後，研究範圍擴展至音樂元素，包括旋律、節奏等，以探討音樂提示詞與生成封面品質的關聯性。

研究步驟

步驟一：歌詞數據蒐集與預處理

數據蒐集：使用網路爬蟲技術蒐集大量各類型的歌詞，涵蓋不同語言、流派與時代，以確保資料的多樣性。

數據預處理：

文本標準化：轉換所有文字爲小寫，移除標點符號與特殊字符，以減少噪音數據的影響。

詞向量化：使用詞嵌入技術（如 Word2Vec、BERT 或 FastText）將歌詞轉換爲詞向量，爲後續分析與模型設計提供數值表示。

步驟二：歌詞意境分析與特徵提取

意境分析：利用 Bert 模型分析歌詞表達的情感，如樂觀、悲傷、思念等，並標註其情感類別（參見 [15]）。

主題提取：基於自然語言處理技術，使用 LDA（Latent Dirichlet Allocation）或 Transformer 模型對歌詞進行主題分類，例如愛情、自然、旅行等。

向量轉換：將上述特徵轉換爲詞向量，以便用於後續模型的訓練與設計。

步驟三：**Fusion** 模型建構

Token Fusion 設計：將歌詞的情感特徵與主題特徵融合爲 Token Vector，並轉換爲適用於 Diffusion Model 的 prompt，確保生成的封面與歌詞的意境匹配。

神經網絡模型訓練：構建深度學習模型，訓練 Fusion 模型，使其能有效關聯 Token Vector 與歌詞意境，並生成高質量的 prompt。

步驟四：使用 **Diffusion Model** 生成封面

利用步驟三中生成的 prompt，搭配其他控制超參數（Hyperparameters），將數據輸入 Diffusion Model，生成音樂封面，並透過調整模型參數優化生成結果。

步驟五：拓展模型至音樂範圍

資料蒐集：使用 ffmpeg 與爬蟲技術下載對應歌曲的 WAV 檔案，並確保音訊資料的多樣性。

資料前處理：提取 WAV 檔案的音樂特徵，如音符、旋律、節奏速度、和弦進行等，轉換爲結構化數據。

Fusion 模型擴展：將歌詞與音樂的嵌入向量（embedding）進行融合，並調整 Fusion 模型，使其能同時考慮歌詞與音樂的特徵，生成更具表現力的 prompt。

生成封面：按照步驟四的方法，使用擴展後的模型生成音樂封面，確保封面設計能夠體現歌詞與音樂的整體風格。並且將會設計一種對比實驗可以比較清晰與模糊的 prompt 之間是否存在顯著的差異以及如何量化 prompt 的清晰程度

步驟六：模型比較與評估

相似度評估：使用 CLIP 模型重新分析生成的封面，並計算其與原始歌詞及音樂特徵之間的相似度。([3])

用户評測：邀請專業設計師與音樂愛好者對生成的封面進行主觀評價，收集用户回饋以增加模型訓練的資料準確性。

模型優化：根據評估結果調整 Fusion 模型與 Diffusion Model 的權重與超參數，以提升封面生成的準確性與美感。

五、預期結果

本研究預期將證明基於歌詞特徵的自動封面生成技術在視覺呈現上的有效性，能夠準確捕捉並呈現音樂的情感與意境。同時，將建立一個有效連結音樂特徵與視覺圖像的量化模型爲未來生成技術提供理論基礎。在此基礎上研究將進一步擴展至旋律、節奏等音樂元素的應用，提升封面生成的表現力。最終這項技術將爲資源有限的音樂創作者提供高效、低成本的封面生成工具，減輕視覺設計負擔，提升創作效率。

六、需要指導教授指導內容

指導教授將在研究的各個階段提供專業指導，特別是在模型選擇、實驗設計及數據分析方面，確保研究方法的科學性與可行性。在歌詞特徵提取與情感分析階段，教授將協助優化 NLP 技術應用，提升模型的準確性。在 Fusion 模型構建與 Diffusion Model 生成過程中，指

導教授將提供技術支持，協助調整參數與優化生成結果。最後，教授將指導評估方法，確保結果具備學術價值，並提供修改建議以完善研究成果。

七、參考文獻

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[2] Q. Zeng, J. Tan, R. Zhang, Y. Wang, S. Zhang, and J. Xiao. MusicBERT: Symbolic music understanding with large-scale pre-training. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2021.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[4] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[5] D. Yu, K. Song, P. Lu, T. He, X. Tan, W. Ye, S. Zhang, and J. Bian. MusicAgent: An AI Agent for Music Understanding and Generation with Large Language Models. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[6] L. Song, J. Liu, X. Wang, and Z. Chen. Generative Disco: Text-to-Video Generation for Music Visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[7] C. Zhang, C. Zhang, M. Zhang, I. S. Kweon, and J. Kim. Text-to-Image Diffusion Models in Generative AI: A Survey. *arXiv preprint arXiv:2303.07909*, 2024.

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-Shot Text-to-Image Generation. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2021.

[9] Wang, Li, Gao, Boyan, Li, Yanran, Wang, Zhao, Yang, Xiaosong, Clifton, David A., and Xiao, Jun. Exploring the latent space of diffusion models directly through singular value decomposition. *arXiv preprint arXiv:2502.02225*, 2025.

[10] Cheng Xinle, Chen Zhuoming, and Jia Zhihao. CAT Pruning: Cluster-Aware Token Pruning For Text-to-Image Diffusion Models. *arXiv preprint arXiv:2502.00433*, 2025.

[11] Wiszenko, Michał, Stefański, Kacper, Malesa, Piotr, Modrzejewski, Mateusz, and Pokorzyński, Łukasz. Miditok Visualizer: a Tool for Visualization and Analysis of Tokenized Midi Symbolic Music. In *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

[12] Xu Junjie, Wu Xingjiao, Yao Tanren, Zhang Zihao, Bei Jiayang, Wen Wu, and He Liang. Aesthetic Matters in Music Perception for Image Stylization: an Emotion-Driven Music-to-Visual Manipulation. *arXiv preprint arXiv:2501.01700*, 2025.

[13] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023.

[14] Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., et al. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*, 2023.

[15] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.