**Measuring Visual Sycophancy in Multimodal Models**

17 Agosto 2024

<https://www.arxiv.org/pdf/2408.09111>

**~~COME MIGLIORARE LO STATO DELL’ARTE~~**

### ****~~Integrazione di Knowledge Graphs (KGs) per la Correzione del Bias Visivo~~****

**~~Proposta:~~**

* ~~Integrare~~ **~~Knowledge Graphs (KGs)~~** ~~per migliorare la capacità dei modelli multimodali di riconoscere e correggere il bias visivo, utilizzando informazioni strutturate per evitare risposte influenzate da segnali visivi fuorvianti.~~

**~~Metodologia:~~**

* ~~Collegare un Knowledge Graph al modello multimodale per fornire un controllo aggiuntivo sulle risposte. Il Knowledge Graph può contenere informazioni su concetti chiave, relazioni tra oggetti e conoscenze logiche, aiutando il modello a verificare la coerenza delle risposte con la base di conoscenza.~~
* ~~Implementare un~~ **~~modulo di verifica logica~~** ~~che confronta la risposta proposta con le informazioni del KG per ridurre l'influenza dei segnali visivi quando sono in conflitto con la conoscenza esistente.~~

**~~Esempio Pratico:~~**

* ~~In una domanda basata su un'immagine che mostra un gatto e un cane, se l'immagine enfatizza visivamente il gatto ma la domanda testuale richiede informazioni sul cane, il Knowledge Graph può correggere il modello fornendo informazioni contestuali su entrambi gli animali, aiutando a rispondere correttamente.~~

**~~Risultato Atteso:~~**

* ~~Un sistema più robusto contro i bias visivi, in grado di utilizzare informazioni strutturate da KGs per garantire che le risposte multimodali siano logiche e corrette.~~

### ****~~Sviluppo di un Framework di Mitigazione per la Visual Sycophancy~~****

**~~Proposta:~~**

* ~~Progettare un framework che utilizza tecniche di~~ **~~calibrazione multimodale~~** ~~per bilanciare le informazioni visive e testuali. Questo framework potrebbe essere basato su un modello di attenzione che adatta dinamicamente il peso attribuito alle modalità, riducendo l'eccessiva fiducia nei segnali visivi.~~

**~~Metodologia:~~**

* ~~Sviluppare un modulo di attenzione cross-modale che attribuisca pesi dinamici alle informazioni visive e testuali, in base alla coerenza tra le due modalità.~~
* ~~Integrare meccanismi di interpretabilità come~~ **~~SHAP~~** ~~o~~ **~~LIME~~** ~~per visualizzare quali caratteristiche (testuali o visive) influenzano maggiormente le risposte del modello, e applicare tecniche di interpretabilità per ridurre la fiducia eccessiva nelle informazioni visive.~~

**~~Esempio Pratico:~~**

* ~~In un compito di Visual Question Answering (VQA), il framework potrebbe analizzare la coerenza tra la domanda testuale e l'immagine prima di generare la risposta. Se l'immagine mostra un oggetto fuorviante (ad esempio, una banana evidenziata visivamente in una domanda su una mela), il modulo di attenzione potrebbe ridurre il peso delle informazioni visive nel calcolo finale della risposta.~~

**~~Risultato Atteso:~~**

* ~~Un framework che mitiga in modo efficace la~~ **~~visual sycophancy~~**~~, migliorando l'accuratezza del modello in contesti critici dove il testo e l'immagine potrebbero non essere perfettamente allineati.~~

**~~Sviluppo di un Dataset per Testare la Robustezza contro la Visual Sycophancy~~**

**~~Proposta:~~**

* ~~Creare un nuovo~~ **~~dataset multimodale~~** ~~progettato per testare la robustezza dei modelli contro la visual sycophancy. Questo dataset potrebbe includere immagini con segnali visivi ambigui o fuorvianti, accanto a domande testuali che richiedono risposte basate su conoscenze logiche o testuali piuttosto che sull’immagine.~~

**~~Metodologia:~~**

* ~~Progettare un dataset con immagini manipolate visivamente (come dimensioni dei testi, colori, evidenziazione di elementi visivi errati) e domande a risposta multipla che richiedono una fusione coerente di informazioni visive e testuali.~~
* ~~Valutare le performance di modelli multimodali utilizzando metriche come~~ **~~accuracy~~** ~~e~~ **~~log-likelihood~~** ~~delle risposte in scenari neutrali e con bias visivo.~~

**~~Esempio Pratico:~~**

* ~~Un’immagine con un oggetto evidenziato in modo fuorviante (ad esempio, un libro colorato di rosso in una domanda sulla natura) accanto a una domanda che richiede di ignorare l’oggetto e concentrarsi su informazioni testuali per rispondere correttamente.~~

**~~Risultato Atteso:~~**

* ~~Un dataset che possa essere usato per benchmark di modelli multimodali, migliorando la capacità della comunità di ricerca di testare la robustezza dei modelli contro la visual sycophancy.~~

**~~PERCHE’ MIGLIORA LO STATO DELL’ARTE~~**

**~~Integrazione di Knowledge Graphs (KGs) per la Correzione del Bias Visivo~~**

**~~Perché Migliora lo Stato dell'Arte:~~**

* **~~Maggiore Coerenza e Precisione~~**~~: L'integrazione dei KGs consente di arricchire le capacità dei modelli multimodali aggiungendo un controllo basato su conoscenze strutturate, che può verificare e correggere le risposte influenzate da segnali visivi fuorvianti. Questo è un approccio innovativo rispetto ai modelli attuali, che spesso si affidano esclusivamente ai segnali visivi o testuali senza una valida rete di supporto logico.~~
* **~~Riduzione della Visual Sycophancy~~**~~: I modelli multimodali tendono a enfatizzare eccessivamente gli input visivi, portando a errori quando l'immagine è ambigua o ingannevole. L'uso dei KGs aiuta a mitigare questo problema, consentendo al modello di basare le risposte su una fonte di conoscenza logica e coerente, invece di affidarsi esclusivamente all'input visivo.~~
* **~~Applicazione Cross-Domain~~**~~: I KGs permettono di generalizzare meglio su domini complessi (come medicina, finanza, e altre aree ad alto rischio) dove la precisione è fondamentale, migliorando la capacità dei modelli di evitare bias in contesti specifici.~~

**~~Limiti:~~**

* **~~Dipendenza dai KGs~~**~~: L'efficacia di questo approccio dipende dalla qualità e dalla copertura dei KGs utilizzati. Se i KGs sono incompleti o imprecisi, potrebbero fornire informazioni errate, riducendo l'efficacia del modello.~~
* **~~Costi Computazionali~~**~~: L'integrazione di KGs aggiunge una complessità computazionale, soprattutto quando i modelli devono consultare frequentemente una vasta rete di conoscenza.~~

**~~Sviluppi Futuri:~~**

* **~~Espansione dei Knowledge Graphs~~**~~: I KGs dovrebbero essere costantemente aggiornati e ampliati per coprire una gamma più ampia di domini e contesti, assicurando che la loro integrazione sia utile in una varietà di applicazioni.~~
* **~~Integrazione di KGs Dinamici~~**~~: Lo sviluppo di KGs dinamici, che si aggiornano in tempo reale, potrebbe ulteriormente migliorare la capacità del modello di rispondere correttamente a scenari in evoluzione.~~

**~~Sviluppo di un Framework di Mitigazione per la Visual Sycophancy~~**

**~~Perché Migliora lo Stato dell'Arte:~~**

* **~~Bilanciamento Dinamico delle Modalità~~**~~: La proposta di un~~ **~~modulo di attenzione cross-modale~~** ~~che adatta dinamicamente il peso delle informazioni visive e testuali è un progresso rispetto ai modelli attuali. Questi ultimi tendono a dare troppo peso ai segnali visivi, anche quando sono meno rilevanti. Questo nuovo approccio aumenta la capacità dei modelli di prendere decisioni più informate.~~
* **~~Interpretabilità e Trasparenza~~**~~: L'integrazione di tecniche di interpretabilità come~~ **~~SHAP~~** ~~o~~ **~~LIME~~** ~~consente agli sviluppatori e agli utenti di comprendere meglio le decisioni prese dai modelli. Questo aumenta la trasparenza e la fiducia nei modelli multimodali, riducendo al contempo il rischio di errori derivanti da segnali visivi fuorvianti.~~
* **~~Applicabilità in Scenari Critici~~**~~: Questo framework migliora l'accuratezza dei modelli in contesti critici (es. medicina, sicurezza), dove l'allineamento tra testo e immagine è essenziale per evitare errori costosi.~~

**~~Limiti:~~**

* **~~Calibrazione Complessa~~**~~: Bilanciare correttamente le informazioni visive e testuali può essere complesso, soprattutto quando il peso attribuito a ciascuna modalità deve essere adattato dinamicamente in base a diverse situazioni. La calibrazione del modello richiede quindi iterazioni e test significativi.~~
* **~~Problemi di Interpretabilità Limitata~~**~~: Anche con tecniche come SHAP e LIME, l'interpretazione delle decisioni cross-modali potrebbe rimanere difficile in scenari complessi, in cui la relazione tra testo e immagine è ambigua o non ben definita.~~

**~~Sviluppi Futuri:~~**

* **~~Ottimizzazione del Modulo di Attenzione~~**~~: Future ricerche potrebbero focalizzarsi su un modulo di attenzione ancora più raffinato, che adatti il peso delle modalità in modo predittivo, anticipando le incoerenze tra testo e immagine.~~
* **~~Ampliamento delle Tecniche di Interpretabilità~~**~~: L'integrazione di tecniche di interpretabilità più sofisticate e specifiche per modelli multimodali potrebbe migliorare ulteriormente la trasparenza del sistema.~~

**~~Sviluppo di un Dataset per Testare la Robustezza contro la Visual Sycophancy~~**

**~~Perché Migliora lo Stato dell'Arte:~~**

* **~~Benchmarking Specializzato~~**~~: Lo sviluppo di un dataset dedicato alla~~ **~~visual sycophancy~~** ~~fornisce un nuovo strumento per la valutazione dei modelli multimodali. Attualmente, i dataset standard non sono progettati specificamente per valutare la resistenza dei modelli a segnali visivi ingannevoli o ambigui. Questo nuovo dataset colmerebbe questa lacuna, consentendo una valutazione più rigorosa e mirata.~~
* **~~Fusione di Informazioni Visive e Testuali~~**~~: Creare un dataset che costringa i modelli a bilanciare correttamente informazioni testuali e visive migliora la loro robustezza. Aumenta anche la capacità della comunità di ricerca di progettare modelli che riducano l'influenza sproporzionata dei segnali visivi.~~
* **~~Validazione e Test in Contesti Critici~~**~~: Questo dataset può essere utilizzato per testare i modelli in situazioni reali in cui l'accuratezza è cruciale. La costruzione di immagini volutamente fuorvianti o ambigue consente di vedere come i modelli multimodali reagiscono sotto stress, un aspetto che finora non è stato sufficientemente esplorato.~~

**~~Limiti:~~**

* **~~Sviluppo del Dataset~~**~~: Creare un dataset che rappresenti accuratamente una vasta gamma di scenari fuorvianti richiede tempo e risorse. Potrebbe essere difficile ottenere immagini e domande sufficientemente diversificate e complesse per testare tutti i tipi di bias visivi.~~
* **~~Overfitting ai Test~~**~~: C'è il rischio che i modelli si adattino eccessivamente al dataset, migliorando le prestazioni solo sui dati di test e non generalizzando bene in scenari reali.~~

**~~Sviluppi Futuri:~~**

* **~~Espansione del Dataset~~**~~: Aggiungere continuamente nuove immagini, domande e scenari per aumentare la difficoltà del dataset, espandendo la gamma di situazioni in cui i modelli multimodali vengono testati.~~
* **~~Collaborazione con Comunità di Ricerca~~**~~: Coinvolgere la comunità per testare modelli su questo dataset e contribuire con nuove immagini e domande potrebbe accelerare la creazione di un benchmark più solido e universale.~~

**Exposing and Mitigating Spurious Correlations for Cross-Modal Retrieval**

6 Aprile 2023

https://arxiv.org/pdf/2304.03391

**COME MIGLIORARE LO STATO DELL’ARTE**

**Analisi delle Interazioni tra Bias Visivi e Correlazioni Spurie nei Modelli Multimodali**

### Proposta:

Indagare **come i bias visivi, come la visual sycophancy**, interagiscono con le **correlazioni spurie** nei modelli multimodali. L'obiettivo è identificare se e come le debolezze del modello si amplificano quando le informazioni visive e testuali sono in conflitto, valutando la capacità del modello di ignorare le correlazioni spurie o correggere il proprio comportamento in presenza di segnali visivi fuorvianti.

### Metodologia:

1. **Definire scenari conflittuali**: Creare un set di dati con esempi in cui immagini e testi presentano **contraddizioni esplicite** o informazioni ambigue. Per esempio, un'immagine potrebbe contenere un cane senza frisbee, ma la descrizione testuale potrebbe fare riferimento a un frisbee (correlazione spuria).
2. **Manipolazioni visive intenzionali**: Utilizzare tecniche di data augmentation visiva, come il **highlighting o l'enfasi visiva**, su regioni che contengono o dovrebbero contenere oggetti correlati in modo spurio. Questo permette di studiare se il modello cambia la propria risposta a causa di un bias visivo o di una correlazione spuria.
3. **Test multimodale incrociato**: Verificare la reazione del modello a queste contraddizioni, misurando se si affida più alle informazioni visive (bias visivo) o a correlazioni testuali imparate durante l'addestramento (correlazione spuria).
4. **Metriche di valutazione**: Utilizzare metriche personalizzate che misurano l'affidabilità del modello, come la frequenza con cui il modello ignora correttamente o sbaglia in presenza di segnali spurii (sia visivi che testuali). Proporre una nuova metrica che combini **sycophancy visiva** e correlazioni spurie.

### Esempio Pratico:

Immagina un dataset dove alcune immagini mostrano **cani senza frisbee**, ma durante l'addestramento, il modello ha imparato erroneamente che frisbee e cani sono spesso correlati. Se l'immagine contiene un cane senza frisbee e un **highlight visivo** segnala una zona che suggerisce un frisbee, il modello potrebbe erroneamente generare una descrizione come "Un cane che gioca con un frisbee", cadendo sia nel **bias visivo** che nella correlazione spuria.

Testando il modello in questi scenari, si può valutare se:

* Il modello **dà priorità all'informazione visiva** evidenziata, ignorando il fatto che il frisbee non sia presente.
* Il modello **applica correlazioni spurie** tratte dai dati di addestramento, producendo una descrizione errata, indipendentemente dai segnali visivi.

### Risultati Attesi:

L'analisi rivelerà se le correlazioni spurie e il bias visivo sono problemi interdipendenti o indipendenti nei modelli multimodali. Si può dimostrare che la presenza di segnali visivi fuorvianti tende a **rafforzare l'effetto delle correlazioni spurie**, portando a errori sistematici nelle descrizioni generate o nelle risposte a domande basate su immagini.

### Impatto:

Questo studio fornirebbe una comprensione più profonda delle debolezze dei modelli multimodali, suggerendo miglioramenti per la mitigazione con tecniche combinate, come:

* **Modelli di attenzione bilanciati** tra modalità visiva e testuale.
* **Rafforzamento di controlli incrociati** tra le informazioni visuali e testuali, rendendo il modello più robusto alle contraddizioni.

Questa analisi fornirebbe strumenti per migliorare l'accuratezza e l'affidabilità dei modelli in applicazioni critiche, come l'assistenza medica o le analisi legali, dove l'influenza di correlazioni spurie o segnali visivi ambigui può portare a errori gravi.

### ****Utilizzo di Knowledge Graphs (KGs) per la Verifica Semantica delle Descrizioni****

**Proposta:**

* Integrare **Knowledge Graphs (KGs)** per migliorare la precisione semantica delle descrizioni generate dai modelli di retrieval, verificando che gli oggetti e le relazioni menzionate nei testi siano coerenti con ciò che è presente nell'immagine. Questo ridurrebbe l'impatto delle correlazioni spurie utilizzando una base di conoscenza esplicita per garantire la coerenza concettuale.

**Metodologia:**

* Implementare un processo di verifica semantica dove le descrizioni generate dal modello (ad esempio, "Due cani giocano con un frisbee") vengono confrontate con un **KG multimodale**. Il KG fornisce conoscenze esplicite riguardanti oggetti e le loro possibili associazioni (ad esempio, frisbee e cani non devono apparire insieme se il frisbee è stato rimosso).
* Se la descrizione non è coerente con le informazioni presenti nel KG, viene generato un avviso, e il modello può correggere la descrizione, migliorando così la robustezza contro correlazioni spurie.

**Esempio Pratico:**

* Supponiamo che in un'immagine sintetica siano stati rimossi tutti gli oggetti correlati al frisbee. Il KG, con una struttura semantica, rileva che la descrizione proposta dal modello menziona ancora "frisbee". Il sistema corregge la descrizione eliminando tale riferimento.

**Risultato Atteso:**

* Un significativo miglioramento nella coerenza semantica delle descrizioni generate, con una forte riduzione delle correlazioni spurie, poiché il KG funge da filtro per assicurare che le relazioni descritte siano basate sugli oggetti presenti.

### ****Generazione Dinamica di Prompt Controfattuali per Ridurre il Bias Spurio****

**Proposta:**

* Progettare un sistema di generazione dinamica di **prompt controfattuali** che sfidino il modello durante l'inferenza, proponendo domande o scenari che escludono in modo esplicito certi oggetti, costringendo il modello a focalizzarsi sugli oggetti presenti.

**Metodologia:**

* Creare una pipeline di generazione di prompt controfattuali in cui, durante l'inferenza, viene richiesto al modello di generare descrizioni basate su informazioni ambigue o parzialmente incomplete. Ad esempio, si potrebbero generare prompt come "Descrivi l'immagine senza menzionare nessun giocattolo" in un'immagine di cani, per verificare che il modello non menzioni oggetti non presenti.

**Esempio Pratico:**

* In una query di retrieval immagine-testo, il modello potrebbe ricevere un prompt controfattuale come "Descrivi i cani senza menzionare accessori". Questo approccio costringerebbe il modello a concentrarsi sui cani e a evitare di menzionare correlazioni spurie come frisbee o giocattoli.

**Risultato Atteso:**

* Migliore capacità del modello di gestire scenari complessi, riducendo la dipendenza da correlazioni spurie. Inoltre, si forzerebbe il modello a fornire descrizioni più accurate e meno ridondanti.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**Analisi delle Interazioni tra Bias Visivi e Correlazioni Spurie nei Modelli Multimodali**

**Perché Migliora lo Stato dell'Arte:**

* **Esplorazione di Interazioni Tra Modalità**: La proposta di analizzare come le **correlazioni spurie** si intrecciano con i **bias visivi**, come la visual sycophancy, aggiunge una nuova dimensione all'analisi dei modelli multimodali. Attualmente, le ricerche tendono a trattare queste problematiche separatamente, mentre questa proposta evidenzia l'importanza di comprenderne le interazioni. La capacità di un modello di ignorare correlazioni spurie in presenza di segnali visivi fuorvianti è un aspetto poco esplorato, ma cruciale per garantire la robustezza in scenari reali.
* **Generazione di Scenari Conflittuali**: La creazione di dataset con scenari in cui testi e immagini presentano informazioni contrastanti fornisce una metodologia nuova per testare i modelli. Questo approccio permette di valutare come il modello bilancia le modalità visive e testuali in condizioni di conflitto, spingendo lo stato dell'arte verso lo sviluppo di modelli più resilienti.
* **Metriche Nuove e Combinazione di Bias**: Proporre nuove metriche che combinino la **visual sycophancy** e le **correlazioni spurie** permette una valutazione più dettagliata del comportamento dei modelli in scenari complessi, fornendo un nuovo strumento per confrontare l'affidabilità dei modelli multimodali.

**Limiti:**

* **Dataset Complessi da Creare**: La creazione di un dataset di scenari conflittuali visivo-testuali richiede una notevole quantità di risorse e può essere complicata da progettare in modo bilanciato, rischiando di coprire solo una gamma limitata di situazioni.
* **Difficoltà nella Generalizzazione**: Anche se l'analisi delle interazioni tra bias visivi e correlazioni spurie è innovativa, potrebbe essere difficile generalizzare i risultati a un'ampia gamma di scenari reali, specialmente in domini non inclusi nel dataset.

**Sviluppi Futuri:**

* **Espansione dei Dataset**: La creazione di dataset più ampi e diversificati per testare queste interazioni in diversi contesti (come medicina, assistenza legale, ecc.) potrebbe migliorare l'applicabilità dei risultati.
* **Metodi di Mitigazione Automatizzati**: Sviluppare metodi che automatizzino la mitigazione dei bias visivi e delle correlazioni spurie, applicando correzioni in tempo reale durante l'inferenza.

**Utilizzo di Knowledge Graphs (KGs) per la Verifica Semantica delle Descrizioni**

**Perché Migliora lo Stato dell'Arte:**

* **Verifica Semantica Aggiuntiva**: L'integrazione dei **Knowledge Graphs (KGs)** aggiunge un ulteriore livello di controllo alla coerenza delle descrizioni generate dai modelli multimodali. Questo aiuta a contrastare le correlazioni spurie utilizzando informazioni strutturate che garantiscono che gli oggetti menzionati nel testo siano presenti o coerenti con l'immagine.
* **Correzione Basata sulla Conoscenza**: Attualmente, i modelli multimodali basano le loro decisioni solo sui segnali appresi durante l'addestramento, spesso cadendo in correlazioni spurie. I KGs forniscono una base di conoscenza esterna che può correggere le descrizioni, migliorando la coerenza logica e riducendo errori causati da correlazioni errate.
* **Applicazione nei Contesti Critici**: In ambiti come la medicina o la sorveglianza, dove le descrizioni devono essere estremamente accurate, i KGs possono fungere da meccanismo di sicurezza per evitare errori causati da segnali fuorvianti.

**Limiti:**

* **Qualità del Knowledge Graph**: L'efficacia dipende fortemente dalla qualità e dalla completezza del KG. Se il KG contiene informazioni incomplete o errate, potrebbe introdurre nuovi errori o non essere in grado di correggere le correlazioni spurie.
* **Costi Computazionali**: Integrare un KG richiede risorse computazionali aggiuntive, il che potrebbe rallentare l'inferenza in contesti ad alte prestazioni o su larga scala.

**Sviluppi Futuri:**

* **Creazione di Knowledge Graphs Dinamici**: Lo sviluppo di KGs che si aggiornano dinamicamente o che apprendono durante l'esecuzione del modello potrebbe migliorare ulteriormente la capacità di correzione semantica.
* **KGs Specializzati per Domini**: KGs specifici per determinati domini (medico, legale, ecc.) potrebbero essere integrati per garantire che le descrizioni siano coerenti con le conoscenze specialistiche del settore.

**Generazione Dinamica di Prompt Controfattuali per Ridurre il Bias Spurio**

**Perché Migliora lo Stato dell'Arte:**

* **Forzare la Riduzione delle Correlazioni Spurie**: La generazione dinamica di **prompt controfattuali** costringe il modello a non affidarsi a correlazioni spurie, obbligandolo a focalizzarsi su informazioni pertinenti. Questo è un approccio nuovo ed efficace per ridurre la dipendenza del modello da associazioni sbagliate, migliorando la capacità di gestire scenari complessi.
* **Miglioramento dell'Accuratezza**: I prompt controfattuali, che esplicitamente escludono oggetti o categorie specifiche, aiutano a migliorare la precisione delle descrizioni, costringendo il modello a concentrarsi su ciò che è realmente presente nell'immagine.
* **Robustezza nei Sistemi di Retrieval**: L'utilizzo di prompt controfattuali può aiutare a migliorare la robustezza dei sistemi di retrieval immagine-testo, specialmente in contesti in cui è cruciale evitare errori derivanti da correlazioni spurie (ad esempio, nella sorveglianza o nel rilevamento di anomalie).

**Limiti:**

* **Costruzione Complessa di Prompt**: Progettare prompt controfattuali efficaci può essere complesso e richiede una profonda comprensione dei potenziali bias presenti nei modelli. Inoltre, potrebbe essere difficile coprire tutti i tipi di correlazioni spurie in modo esaustivo.
* **Applicabilità Limitata**: Questo approccio potrebbe non funzionare in tutte le applicazioni multimodali, specialmente in scenari in cui è difficile creare prompt controfattuali che siano significativi e utili per ogni immagine.

**Sviluppi Futuri:**

* **Ottimizzazione Automatica dei Prompt**: L'uso di tecniche di machine learning per generare automaticamente prompt controfattuali ottimizzati potrebbe rendere questo approccio più scalabile ed efficiente.
* **Espansione ad Altri Domini**: Applicare questo approccio a modelli multimodali in altri domini (come il riconoscimento facciale o la sorveglianza video) potrebbe migliorare la capacità di identificare errori nelle descrizioni generate in tempo reale.

**PROPOSTA INTERESSANTE**

**Titolo della Proposta di Ricerca:**

**Analisi delle Interazioni tra Bias Visivi e Correlazioni Spurie nei Modelli Multimodali con Verifica Semantica tramite Knowledge Graphs e Generazione di Prompt Controfattuali**

**Obiettivo della Ricerca:**

L'obiettivo della ricerca è studiare l'interazione tra bias visivi e correlazioni spurie nei modelli multimodali, in particolare in scenari conflittuali tra input visivi e testuali. L'approccio proposto mira a sviluppare un framework che combini la **verifica semantica tramite Knowledge Graphs (KGs)** e la **generazione di prompt controfattuali** per ridurre le dipendenze da correlazioni spurie e migliorare la robustezza del modello multimodale.

**Background:**

I modelli multimodali che combinano input visivi e testuali sono spesso soggetti a:

* **Bias Visivo (Visual Sycophancy)**: L'eccessiva dipendenza dai segnali visivi che può portare a errori quando le informazioni visive sono fuorvianti.
* **Correlazioni Spurie**: Associazioni errate apprese durante l'addestramento che portano il modello a fare inferenze inappropriate, come associare in modo errato un cane a un frisbee anche quando quest'ultimo non è presente.

Questi problemi possono portare a errori sistematici nei modelli multimodali, soprattutto in contesti critici come la diagnostica medica e le analisi legali.

**Proposta di Ricerca:**

L'approccio proposto combina tre elementi principali:

1. **Analisi delle Interazioni tra Bias Visivi e Correlazioni Spurie**: Studiare come i bias visivi e le correlazioni spurie interagiscono, amplificando gli errori nei modelli multimodali.
2. **Verifica Semantica tramite Knowledge Graphs (KGs)**: Utilizzare KGs per verificare la coerenza semantica tra gli oggetti rilevati nelle immagini e le descrizioni generate, aiutando a filtrare e correggere le correlazioni spurie.
3. **Generazione di Prompt Controfattuali**: Generare prompt controfattuali per sfidare il modello e ridurre la dipendenza da correlazioni spurie, forzando il modello a concentrarsi sugli oggetti presenti nell'immagine.

**Struttura dei Moduli:**

**1. Modulo di Rilevazione degli Oggetti (Modello Visivo)**

* **Funzione**: Identificare gli oggetti presenti nell'immagine.
* **Architettura**:
  + Un modello di **object detection** (ad esempio, **Faster R-CNN** o **YOLO**), che rileva e classifica gli oggetti nelle immagini.
  + **Output**: Una lista di oggetti presenti nell'immagine, con metadati come coordinate di bounding box e categorie di oggetti.

**2. Modulo Multimodale per la Generazione di Descrizioni**

* **Funzione**: Generare una descrizione testuale basata su input visivi e testuali.
* **Architettura**:
  + Un modello multimodale come **CLIP** (Contrastive Language-Image Pretraining) o un modello **VQA** (Visual Question Answering) per generare risposte o descrizioni combinate.
  + **Output**: Una descrizione testuale dell'immagine, che può includere correlazioni spurie (ad esempio, "Un cane gioca con un frisbee").

**3. Knowledge Graph (KG) per la Verifica Semantica**

* **Funzione**: Verificare la coerenza tra la descrizione generata dal modello multimodale e gli oggetti rilevati nell'immagine.
* **Costruzione**:
  + **Entità**: Oggetti e concetti rilevati nell'immagine (ad esempio, "cane", "frisbee", "giardino").
  + **Relazioni**: Associazioni tra oggetti e le loro relazioni semantiche (ad esempio, "Un cane può giocare con un frisbee solo se il frisbee è presente").
  + Il KG viene popolato utilizzando fonti strutturate come **DBpedia**, **Wikidata**, o ontologie specifiche per il dominio (ad esempio, ontologie mediche).
* **Verifica**:
  + Confrontare l'elenco degli oggetti rilevati nell'immagine con la descrizione generata.
  + **Correzione**: Se la descrizione contiene oggetti non presenti nell'immagine (ad esempio, menzione di un frisbee inesistente), il KG segnala l'incongruenza e può correggere l'output.

**4. Modulo di Generazione di Prompt Controfattuali**

* **Funzione**: Generare prompt controfattuali che costringano il modello a evitare correlazioni spurie.
* **Architettura**:
  + Un modulo NLP che genera prompt testuali durante l'inferenza (ad esempio, "Descrivi l'immagine senza menzionare certi oggetti").
  + Questi prompt sfidano il modello a focalizzarsi su informazioni testuali più specifiche e accurate.
* **Esempio Pratico**:
  + Se l'immagine contiene solo un cane, il prompt potrebbe essere "Descrivi il cane senza menzionare giocattoli". Questo forza il modello a concentrarsi su ciò che è visibile nell'immagine, riducendo la probabilità di menzionare oggetti inesistenti come frisbee.

**5. Modulo di Valutazione delle Correlazioni Spurie**

* **Funzione**: Valutare quanto il modello multimodale dipende da correlazioni spurie e bias visivi.
* **Architettura**: Metriche personalizzate che combinano la misurazione di:
  + **Visual Sycophancy**: Frequenza con cui il modello si affida a segnali visivi fuorvianti.
  + **Correlazioni Spurie**: Frequenza con cui il modello genera descrizioni basate su associazioni errate (ad esempio, menzionare un frisbee inesistente).
* **Output**: Report che quantifica la robustezza del modello in presenza di bias visivi e correlazioni spurie.

**Metodologia:**

**1. Creazione di Scenari Conflittuali**

* **Dataset**: Creare un set di dati multimodale con immagini e testi contraddittori, progettati per testare la capacità del modello di gestire bias visivi e correlazioni spurie.
* **Tecniche di Manipolazione Visiva**: Utilizzare tecniche di **data augmentation** per enfatizzare visivamente oggetti inesistenti (ad esempio, highlighting di una zona che suggerisce un frisbee, anche se non presente).

**2. Test Multimodale Incrociato**

* **Esecuzione degli Esperimenti**: Testare il modello in presenza di scenari conflittuali, monitorando come reagisce agli input visivi rispetto alle correlazioni spurie apprese durante l'addestramento.
* **Metriche di Valutazione**: Implementare metriche per misurare:
  + **Sycophancy visiva**: Frequenza con cui il modello si affida a segnali visivi fuorvianti.
  + **Dipendenza da correlazioni spurie**: Frequenza con cui il modello genera descrizioni errate basate su associazioni spurie.

**3. Integrazione di Knowledge Graphs (KGs) per la Verifica Semantica**

* **Verifica Semantica**: Utilizzare il KG per confrontare gli oggetti rilevati nell'immagine con la descrizione generata dal modello multimodale.
* **Correzione dell'Output**: Se viene rilevata un'incongruenza, il sistema corregge l'output eliminando correlazioni spurie (ad esempio, eliminando la menzione del frisbee se non è presente nell'immagine).

**4. Generazione di Prompt Controfattuali**

* **Creazione Dinamica di Prompt**: Durante l'inferenza, vengono generati prompt controfattuali che sfidano il modello a descrivere l'immagine senza menzionare oggetti che non sono visibili.

**Esempio Pratico:**

Un'immagine mostra un cane senza frisbee. Il modello multimodale, basandosi su correlazioni spurie, genera la descrizione "Un cane che gioca con un frisbee".

* Il **modulo di rilevamento visivo** non ha rilevato alcun frisbee.
* Il **KG** rileva l'incongruenza tra la descrizione e gli oggetti effettivamente rilevati e corregge l'output.
* Il **prompt controfattuale** può inoltre forzare il modello a concentrarsi su altri dettagli pertinenti, come "Descrivi il cane senza menzionare giocattoli".

**Risultati Attesi:**

1. **Comprensione delle Interazioni**: La ricerca chiarirà l'interazione tra bias visivi e correlazioni spurie, fornendo nuovi strumenti per affrontare queste problematiche nei modelli multimodali.
2. **Verifica Semantica Migliorata**: L'integrazione del KG migliorerà la precisione semantica delle descrizioni, riducendo drasticamente le correlazioni spurie.
3. **Riduzione della Sycophancy e delle Correlazioni Spurie**: La generazione di prompt controfattuali forzerà il modello a rispondere solo in base a ciò che è effettivamente presente nell'immagine.

**Impatto:**

* **Miglioramento dell'affidabilità dei modelli multimodali** in contesti critici, come la diagnostica medica o le analisi legali.
* **Sviluppo di nuove tecniche combinate** (Knowledge Graphs e prompt controfattuali) per migliorare la robustezza dei modelli.
* **Nuovi dataset e benchmark** per valutare i modelli multimodali, migliorando lo stato dell'arte nella gestione dei bias visivi e delle correlazioni spurie.

### Perché questa ricerca migliorerebbe lo stato dell'arte?

1. **Riduzione delle Correlazioni Spurie**:
   * **Problema attuale**: I modelli multimodali spesso apprendono correlazioni spurie durante la fase di addestramento, basandosi su associazioni frequenti ma non sempre corrette (ad esempio, "cani giocano con frisbee"). Questi errori si verificano quando il modello fa affidamento su correlazioni statistiche piuttosto che su una comprensione profonda dei dati.
   * **Miglioramento proposto**: Integrando un **Knowledge Graph (KG)** nel processo, il modello può verificare le descrizioni generate confrontandole con una base di conoscenza strutturata. Questo permette di eliminare correlazioni errate o non valide, migliorando la precisione delle risposte generate. Il KG funge da **filtro semantico** che permette al modello di correggere gli errori in tempo reale.
2. **Mitigazione del Bias Visivo (Visual Sycophancy)**:
   * **Problema attuale**: I modelli multimodali tendono a dare troppa importanza ai segnali visivi anche quando questi sono fuorvianti. Ciò può portare a risposte o descrizioni basate su informazioni visive che non sono rilevanti o accurate rispetto alla domanda o al contesto.
   * **Miglioramento proposto**: L'uso di **prompt controfattuali** aiuta a sfidare il modello durante l'inferenza, costringendolo a ignorare oggetti visivi che non sono pertinenti. Questo rende il modello meno propenso a cadere nel bias visivo, migliorando la sua capacità di prendere decisioni basate su più fonti informative (visive e testuali).
3. **Framework Dinamico e Interattivo**:
   * **Problema attuale**: Molti approcci multimodali si affidano a pipeline statiche, dove il modello genera una risposta senza meccanismi di verifica o revisione dell'output. Questo significa che una volta generata una descrizione errata, non ci sono processi che la correggono.
   * **Miglioramento proposto**: Il framework proposto combina diversi moduli (rilevazione visiva, verifica con KG, generazione di prompt controfattuali), creando un sistema **dinamico** che può correggere e migliorare l'output in tempo reale. Questo processo di feedback continuo migliora la robustezza del sistema, aumentando l'affidabilità delle descrizioni.
4. **Applicazioni in Contesti Critici**:
   * **Problema attuale**: In contesti ad alta criticità come la diagnostica medica o le analisi legali, l'affidabilità dei modelli multimodali è essenziale. Gli errori dovuti a bias visivi o correlazioni spurie possono portare a conclusioni sbagliate con conseguenze significative.
   * **Miglioramento proposto**: Il framework proposto non solo aumenta l'accuratezza e la coerenza delle descrizioni, ma fornisce anche un sistema trasparente che permette di visualizzare e correggere gli errori. Questo migliora l'affidabilità in settori dove la precisione è cruciale.

### Limiti della Ricerca:

1. **Dipendenza dalla Qualità del Knowledge Graph**:
   * **Sfida**: Il **KG** è efficace solo nella misura in cui è completo e accurato. Se il KG non contiene conoscenze rilevanti o è carente di informazioni aggiornate, la sua capacità di correggere l'output sarà limitata.
   * **Impatto**: La qualità e la granularità del KG influenzeranno direttamente la capacità del sistema di rilevare e correggere correlazioni spurie. La costruzione e il mantenimento di un KG ricco e pertinente richiede tempo e risorse, specialmente se si lavora in domini altamente specifici come la medicina o il diritto.
2. **Scalabilità**:
   * **Sfida**: Il processo di verifica e correzione con KG e la generazione dinamica di prompt controfattuali richiedono risorse computazionali significative, specialmente quando si tratta di grandi dataset o di applicazioni in tempo reale.
   * **Impatto**: La scalabilità del sistema potrebbe essere un problema per applicazioni che richiedono un'elaborazione rapida e in tempo reale, come sistemi di raccomandazione o assistenti virtuali. Ottimizzare il sistema per lavorare in modo efficiente su larga scala richiederà ulteriori sviluppi.
3. **Compatibilità con Modelli Pre-addestrati**:
   * **Sfida**: Il framework proposto si basa sull'output di modelli pre-addestrati, che possono avere già appreso correlazioni spurie o bias visivi. In alcuni casi, la capacità di correggere questi errori potrebbe essere limitata se il modello multimodale di base non supporta meccanismi di integrazione con moduli di verifica esterni.
   * **Impatto**: L'integrazione del framework con modelli pre-addestrati potrebbe non essere immediata e richiedere modifiche alle pipeline esistenti. Inoltre, per modelli black-box molto complessi, potrebbero esserci limitazioni nell'accesso alle informazioni intermedie necessarie per la verifica.
4. **Difficoltà nell'Addestramento e nella Generazione di Prompt Controfattuali**:
   * **Sfida**: Generare prompt controfattuali efficaci richiede una comprensione approfondita del dominio e degli scenari specifici. Inoltre, il design di prompt che non siano troppo restrittivi o generici rappresenta una sfida tecnica.
   * **Impatto**: Il sistema di prompt controfattuali potrebbe non essere sempre efficace nel ridurre il bias visivo o le correlazioni spurie, soprattutto se i prompt non sono ben progettati. Potrebbe essere necessario un lavoro significativo di **fine-tuning** per garantire che i prompt controfattuali siano appropriati per ogni contesto.
5. **Casi di Complessità Elevata**:
   * **Sfida**: In scenari molto complessi, in cui gli oggetti visivi e le relazioni tra essi sono ambigui, il framework potrebbe non essere in grado di gestire correttamente tutte le situazioni. Ad esempio, se l'immagine mostra un cane vicino a un frisbee nascosto parzialmente o mal rilevato, il sistema potrebbe fallire nel correggere l'output.
   * **Impatto**: In questi casi, la capacità di interpretazione del modello potrebbe ancora essere soggetta a errori, limitando l'efficacia del framework in scenari non standard o altamente complessi.

### Conclusione:

Il framework proposto rappresenta un significativo passo avanti nello stato dell'arte per quanto riguarda la riduzione delle correlazioni spurie e del bias visivo nei modelli multimodali. Tuttavia, la sua efficacia dipende fortemente dalla qualità del Knowledge Graph, dalla scalabilità del sistema e dalla capacità di progettare prompt controfattuali efficaci. Nonostante i limiti, la ricerca ha il potenziale per migliorare l'accuratezza e la robustezza dei modelli in contesti critici, con un impatto positivo su diverse applicazioni reali, soprattutto in ambiti dove l'affidabilità dei modelli è essenziale.

**Semantic Alignment for Multimodal Large Language Models**

23 Agosto 2024

<https://arxiv.org/pdf/2408.12867>

**COME MIGLIORARE LO STATO DELL’ARTE**

**Analisi del ruolo delle semantiche contestuali nella mitigazione della sycophancy**

**Descrizione**: Studiare come le **semantiche contestuali** (guidate dalle altre immagini nel set, come proposto da SAM) possano aiutare a ridurre la sycophancy visiva. Il focus sarebbe su come il contesto visivo coerente possa mitigare l'influenza delle domande fuorvianti, rendendo il modello meno incline a rispondere in modo errato.

**Esperimento**:

* Utilizzare immagini con contesti fortemente correlati e debolmente correlati.
* Testare il modello con domande tendenziose su entrambi i gruppi.
* Misurare la differenza nella resistenza alla sycophancy tra immagini con e senza forte correlazione semantica.

**Risultati attesi**: Si prevede che le immagini con una correlazione semantica più forte abbiano una maggiore capacità di mitigare la sycophancy, poiché il modello sarà meno influenzato dalle informazioni ingannevoli delle query.

**Studio del legame tra sycophancy e storytelling visivo**

**Descrizione**: Esplorare come la sycophancy possa influenzare negativamente il task di **visual storytelling** in cui il modello deve generare una narrazione coerente a partire da immagini diverse. In questo caso, si potrebbero inserire elementi fuorvianti nelle istruzioni o nelle query narrative per valutare se il modello si lascia ingannare dalle leading queries e genera una storia incoerente.

**Esperimento**:

* Creare un dataset di visual storytelling con domande fuorvianti che implicano connessioni errate tra le immagini.
* Valutare se il modello è in grado di ignorare le domande tendenziose e generare una storia coerente.

**Risultati attesi**: I modelli meno robusti alla sycophancy genereranno narrazioni che seguono suggerimenti fuorvianti, mentre i modelli più resistenti manterranno coerenza nella storia.

**Ablazione delle tecniche di contrastive decoding in presenza di disallineamento semantico**

**Descrizione**: Condurre uno studio di **ablazione** per verificare l'efficacia del **contrastive decoding** utilizzato per mitigare la sycophancy visiva in modelli che gestiscono immagini semanticamente disallineate. L'obiettivo è capire se tecniche come il Leading Query Contrastive Decoding (LQCD) proposto per la sycophancy possono essere applicate con successo anche in contesti multi-immagine.

**Esperimento**:

* Rimuovere selettivamente il contrastive decoding in test su immagini con disallineamento semantico e leading queries.
* Valutare l'effetto sull'accuratezza delle risposte.

**Risultati attesi**: Se il contrastive decoding è efficace anche in presenza di disallineamento semantico, questo dimostrerebbe la sua applicabilità per mitigare la sycophancy anche in contesti multimodali.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**Analisi del ruolo delle semantiche contestuali nella mitigazione della sycophancy**

**Perché Migliora lo Stato dell'Arte:**

* **Uso del contesto semantico per ridurre il bias**: La proposta di esplorare come il contesto visivo coerente, utilizzato in SAM (Semantic Alignment Models), possa ridurre la sycophancy visiva rappresenta un avanzamento importante. Attualmente, la sycophancy si verifica quando un modello è influenzato da domande fuorvianti che potrebbero farlo rispondere erroneamente. Il fatto di considerare il contesto globale delle immagini (anziché una singola immagine) può migliorare la capacità del modello di fare inferenze più robuste e basate su una visione più ampia del task.
* **Valutazione delle immagini correlate**: Creare set di immagini con correlazioni semantiche forti e deboli permette di testare come i modelli rispondano in contesti diversi. Questo approccio fornisce un nuovo modo di analizzare la robustezza contro le domande tendenziose, migliorando così la comprensione dei limiti attuali dei modelli multimodali.

**Limiti:**

* **Complessità della correlazione semantica**: Definire e quantificare in modo rigoroso la forza delle correlazioni semantiche tra le immagini può essere soggettivo e complesso. Potrebbe essere difficile progettare dataset che coprano adeguatamente un'ampia gamma di contesti visivi.
* **Generalizzazione del contesto**: Sebbene il contesto semantico possa ridurre il bias in scenari fortemente correlati, potrebbe non funzionare altrettanto bene in domini con ambiguità o immagini poco correlate.

**Sviluppi Futuri:**

* **Espansione dei dataset multimodali**: Creare dataset più ampi e variegati che includano contesti visivi sia forti che deboli, applicabili a diversi domini come medicina, sorveglianza e robotica.
* **Integrazione di modelli semantici avanzati**: Combinare il contesto semantico con reti di conoscenza esterne (ad esempio, Knowledge Graphs) per rafforzare la capacità del modello di resistere alle query fuorvianti.

**Studio del legame tra sycophancy e visual storytelling**

**Perché Migliora lo Stato dell'Arte:**

* **Indagine sulla coerenza narrativa**: L'idea di studiare come la sycophancy influisce sul task di **visual storytelling** è un'area relativamente poco esplorata. La capacità di un modello di generare una narrazione coerente da una serie di immagini è un compito complesso, e questo esperimento permette di valutare la resistenza del modello alle query fuorvianti che potrebbero interrompere la coerenza della storia.
* **Valutazione della robustezza narrativa**: Analizzare la capacità del modello di mantenere coerenza, nonostante domande che suggeriscono falsi collegamenti tra immagini, aiuta a comprendere meglio come la sycophancy influenzi non solo risposte singole, ma narrazioni visive più articolate.

**Limiti:**

* **Dataset personalizzati**: Creare dataset con domande e suggerimenti tendenziosi per il visual storytelling richiede una progettazione meticolosa e potrebbe risultare limitante per modelli generali o per compiti non narrativi.
* **Difficoltà nella valutazione della coerenza narrativa**: La coerenza narrativa è difficile da misurare in modo automatico. Potrebbe essere necessaria un'analisi umana per valutare adeguatamente quanto la sycophancy influenzi la qualità delle storie generate.

**Sviluppi Futuri:**

* **Metriche di coerenza narrativa automatica**: Sviluppare metriche automatiche per misurare la coerenza narrativa in contesti di storytelling visivo. Questo potrebbe includere valutazioni della continuità semantica e della logica temporale delle storie generate.
* **Applicazione a contesti reali**: Utilizzare questo approccio in applicazioni che richiedono narrazioni accurate, come la generazione di rapporti da immagini mediche o la documentazione visiva in contesti legali.

**Ablazione delle tecniche di contrastive decoding in presenza di disallineamento semantico**

**Perché Migliora lo Stato dell'Arte:**

* **Studio sistematico del contrastive decoding**: La proposta di eseguire uno studio di ablazione per valutare l'efficacia del **contrastive decoding** in presenza di disallineamento semantico rappresenta un contributo significativo. Molti modelli multimodali utilizzano tecniche di decoding per bilanciare le informazioni visive e testuali, ma è poco chiaro quanto siano efficaci quando il contesto visivo è incoerente o disallineato. Questo esperimento permetterà di valutare fino a che punto il contrastive decoding possa mitigare la sycophancy anche in scenari complessi.
* **Applicabilità a contesti multimodali più ampi**: Il contrastive decoding, se dimostrato efficace in contesti multimodali complessi, potrebbe essere applicato per migliorare la robustezza dei modelli in scenari che richiedono l'integrazione di molteplici fonti visive o testuali.

**Limiti:**

* **Effetto del disallineamento semantico**: Il disallineamento semantico tra immagini potrebbe essere difficile da controllare e definire. Creare test ben bilanciati che catturino diversi livelli di disallineamento semantico richiede attenzione alla progettazione del dataset.
* **Possibile perdita di precisione**: Rimuovere il contrastive decoding in modelli complessi potrebbe portare a una perdita di accuratezza nelle risposte, riducendo la capacità di gestione delle informazioni multimodali.

**Sviluppi Futuri:**

* **Miglioramento del contrastive decoding**: Sviluppare versioni più avanzate di contrastive decoding che possano adattarsi dinamicamente al grado di disallineamento semantico tra le immagini.
* **Applicazione a modelli multimodali complessi**: Esplorare l'applicabilità del contrastive decoding in modelli che gestiscono più fonti multimodali, come testo, video e audio, per mitigare la sycophancy in contesti multimediali più complessi.

**PROPOSTA INTERESSANTE**

**Titolo Proposto:**

**Mitigazione della Sycophancy nei Modelli Multimodali: Analisi delle Semantiche Contestuali, Visual Storytelling e Tecniche di Decoding Contrastivo**

**Obiettivi Generali:**

1. **Comprendere** come i modelli multimodali possono essere influenzati dalla sycophancy e come la semantica contestuale delle immagini può aiutare a ridurre questa tendenza.
2. **Esplorare** il ruolo della sycophancy nel task di visual storytelling e come le domande tendenziose influenzano la coerenza narrativa.
3. **Valutare** l'efficacia delle tecniche di contrastive decoding per migliorare la resistenza alla sycophancy, anche in contesti complessi come il disallineamento semantico tra immagini.

**Capitoli Principali:**

**1. Introduzione alla Sycophancy nei Modelli Multimodali**

* Descrivere cos'è la sycophancy nei modelli di intelligenza artificiale.
* Presentare il problema nei modelli multimodali (testo + immagini) e la sua rilevanza nelle applicazioni pratiche.
* Introdurre i concetti di semantica contestuale, visual storytelling e contrastive decoding.

**2. Analisi del Ruolo delle Semantiche Contestuali nella Mitigazione della Sycophancy**

* **Teoria**: Approfondire l'importanza del contesto visivo nelle risposte dei modelli.
* **Esperimento Proposto**: Analizzare come immagini con contesti forti e deboli influenzano la capacità del modello di evitare risposte errate a domande tendenziose.
* **Confronto**: Differenze tra modelli che usano o non usano un contesto semantico coerente.

**3. Studio del Legame tra Sycophancy e Visual Storytelling**

* **Teoria**: Esaminare come la sycophancy può portare a narrazioni incoerenti nel task di visual storytelling.
* **Esperimento Proposto**: Creare un dataset con immagini e query fuorvianti, valutando se il modello riesce a produrre una narrazione coerente o se segue gli indizi errati.
* **Implicazioni**: Impatto della sycophancy sulla generazione automatica di contenuti narrativi.

**4. Ablazione delle Tecniche di Contrastive Decoding in Presenza di Disallineamento Semantico**

* **Teoria**: Introduzione al contrastive decoding e alle sue varianti (ad es. Leading Query Contrastive Decoding).
* **Esperimento Proposto**: Testare se queste tecniche possono mitigare la sycophancy quando le immagini sono semanticamente disallineate.
* **Conseguenze**: Dimostrare l'efficacia delle tecniche contrastive non solo nel testo ma anche in contesti multimodali complessi.

**5. Sintesi e Discussione dei Risultati**

* Raccogliere e analizzare i risultati dei tre esperimenti.
* Confrontare i diversi approcci alla mitigazione della sycophancy (semantica contestuale, storytelling e decoding).
* Discussione dei limiti e delle opportunità per ulteriori ricerche.

**6. Conclusioni e Prospettive Future**

* Riassumere i contributi della tesi alla mitigazione della sycophancy nei modelli multimodali.
* Suggerire direzioni future per migliorare la robustezza dei modelli IA, con focus su contesti multimodali complessi.

**Aggiunte Potenziali:**

Per dare più coerenza e forza alla proposta di tesi, potresti considerare di aggiungere:

* **Un capitolo teorico sull'origine e l'impatto della sycophancy** nei modelli linguistici, per avere una base più solida da cui partire prima di passare agli esperimenti.
* **Un esperimento supplementare** che valuti come la dimensione del dataset o il tipo di query (ad esempio, ambigue vs. chiare) influenzi la sycophancy, permettendo di esplorare ulteriori variabili.
* **Tecniche di pre-training o fine-tuning** che migliorano la consapevolezza semantica nei modelli, oltre al contrastive decoding.

Queste integrazioni non solo darebbero un filo conduttore chiaro al lavoro, ma permetterebbero anche una trattazione approfondita e ben strutturata dei vari aspetti della sycophancy nei modelli multimodali, garantendo che ogni capitolo segua in modo logico e complementare gli altri.

### Come Migliorerebbe l'Attuale Stato dell'Arte

1. **Affinamento della comprensione della sycophancy nei modelli multimodali**:
   * La maggior parte della ricerca sulla **sycophancy** si concentra principalmente sui modelli linguistici, ossia su come questi modelli rispondono a domande tendenziose o errate senza una verifica indipendente delle informazioni. Estendere questo concetto ai **modelli multimodali** (che gestiscono immagini oltre che testo) è un'evoluzione cruciale, poiché i contesti visivi aggiungono una dimensione di complessità nella comprensione e nell'inferenza dei modelli.
   * Il focus sul **ruolo della semantica contestuale** come strumento per mitigare la sycophancy è una novità, soprattutto nell'ambito dell'elaborazione di immagini. Questo approccio potrebbe fornire una via per rendere i modelli meno vulnerabili a domande fuorvianti o a connessioni errate, migliorando la robustezza e l'accuratezza delle loro risposte.
2. **Applicazioni in visual storytelling e coerenza narrativa**:
   * Lo studio del **visual storytelling** introduce un aspetto innovativo rispetto alla sycophancy. Sebbene la generazione di narrazioni visive sia un campo in crescita, non esistono molti studi che analizzino come i modelli rispondano a query tendenziose che implicano connessioni sbagliate tra le immagini. Questo esperimento offre un nuovo terreno per capire come migliorare la **coerenza narrativa** prodotta da modelli AI, portando un contributo significativo nell'ambito delle applicazioni creative, dei media e della comunicazione visiva.
   * Inoltre, i modelli che resistono meglio alla sycophancy avrebbero implicazioni non solo in settori creativi, ma anche in ambiti come l'**intelligenza artificiale assistiva** e il **reporting automatico di eventi**, dove la narrazione accurata di ciò che è presente in una serie di immagini è essenziale.
3. **Approfondimento delle tecniche di contrastive decoding in contesti multimodali**:
   * Il **contrastive decoding**, utilizzato per mitigare la sycophancy nei modelli linguistici, è ancora in fase di esplorazione nei contesti multimodali. Il tuo lavoro approfondirebbe se e come queste tecniche possano essere applicate efficacemente quando le immagini sono semanticamente disallineate, fornendo un contributo fondamentale per migliorare la **robustezza** dei modelli multimodali.
   * Se l'ablazione di queste tecniche dimostra che il contrastive decoding è utile anche in scenari visivi complessi, ciò potrebbe aprire la strada a una **generalizzazione delle tecniche di mitigazione** in diversi contesti, rendendo i modelli più resistenti a influenze esterne ingannevoli.

### Limiti e Potenziali Sfide

1. **Complessità della gestione del contesto multimodale**:
   * Un potenziale limite risiede nella **difficoltà di rappresentare e gestire il contesto semantico visivo** in modo efficace. A differenza del testo, dove il significato può essere spesso derivato da relazioni linguistiche ben definite, le immagini possono presentare ambiguità interpretative. Non è detto che un modello riesca a catturare con precisione tutte le connessioni semantiche tra immagini. Se il contesto visivo non è ben compreso o gestito, i miglioramenti attesi nella mitigazione della sycophancy potrebbero non materializzarsi.
2. **Difficoltà nel creare dataset fuorvianti adeguati**:
   * Creare un **dataset di visual storytelling con domande fuorvianti** può essere un'operazione complessa. Le domande devono essere sufficientemente ambigue o tendenziose da confondere il modello, ma non così palesemente sbagliate da essere facilmente ignorate. Questa costruzione richiede un equilibrio preciso e, in caso di errore, potrebbe portare a risultati distorti o non conclusivi.
3. **Applicabilità limitata ad alcuni modelli multimodali**:
   * Le tecniche di **contrastive decoding** e le strategie di mitigazione potrebbero funzionare bene su alcuni modelli multimodali, ma non essere altrettanto efficaci su altri. L'applicabilità e l'efficacia delle tecniche proposte potrebbero dipendere fortemente dall'architettura del modello utilizzato, limitando la generalizzazione dei risultati.
   * Ad esempio, modelli che non sfruttano un robusto sistema di **rappresentazione contestuale delle immagini** potrebbero non trarre vantaggio dalle semantiche contestuali proposte.
4. **Trade-off tra accuratezza e resistenza alla sycophancy**:
   * Potrebbe esserci un **compromesso** tra la resistenza alla sycophancy e l'accuratezza generale del modello. Migliorare la capacità del modello di ignorare domande tendenziose potrebbe portare a risposte più conservative o meno precise in altri scenari, riducendo la performance globale. Questo limite andrebbe studiato attentamente, poiché un modello troppo cauto potrebbe risultare poco utile in alcuni contesti.
5. **Tempi di elaborazione e complessità computazionale**:
   * L'introduzione di **strategie di decoding contrastive** o di una gestione semantica complessa delle immagini potrebbe aumentare i tempi di inferenza e i requisiti computazionali. In applicazioni in tempo reale o su larga scala, questo potrebbe rappresentare un ostacolo pratico all'adozione delle tecniche proposte, richiedendo ottimizzazioni significative per renderle scalabili.

**Debiasing Multimodal Large Language Models**

17 Marzo 2024

<https://arxiv.org/pdf/2403.05262>

**COME MIGLIORARE LO STATO DELL’ARTE**

**Studio sull’impatto dei parametri di decoding nella sycophancy visiva**

**Descrizione**: Indagare come i **parametri di decoding** (come temperatura, top-k, e top-p) influenzino la sycophancy visiva. Si potrebbe esplorare se la modifica dei parametri di generazione del testo riduce la tendenza del modello a seguire suggerimenti visivi ingannevoli o input visivi privi di significato.

**Esperimento**:

* Testare diverse configurazioni di decoding per valutare come influiscono sulla sycophancy visiva.
* Ottimizzare la configurazione per massimizzare l'accuratezza delle risposte in contesti con immagini visivamente ambigue o ingannevoli.

**Risultati attesi**: Le configurazioni di decoding influenzano in modo significativo la sycophancy visiva, e l'ottimizzazione di questi parametri può migliorare le prestazioni del modello.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**Perché Migliora lo Stato dell’Arte:**

1. **Ottimizzazione del Decoding per la Multimodalità**:
   * Modificare i parametri di decoding come la temperatura o il top-p potrebbe consentire di regolare il livello di "creatività" o variabilità nella generazione delle risposte. Questo può essere utile per ridurre la sycophancy visiva, in quanto permette di influenzare la sensibilità del modello agli input visivi rispetto alle informazioni testuali.
   * In modelli multimodali, la fusione tra dati visivi e testuali è fondamentale. L'ottimizzazione dei parametri di decoding può migliorare la capacità del modello di bilanciare correttamente le due modalità, generando risposte più accurate quando il contesto visivo è ambiguo o ingannevole.
2. **Mitigazione dell'Overfitting agli Input Visivi**:
   * La sycophancy visiva può essere il risultato di una fiducia eccessiva da parte del modello nei segnali visivi. Modificare i parametri di decoding può ridurre questa tendenza, poiché l'uso di una temperatura più alta o una configurazione top-k più restrittiva può portare a risposte meno dipendenti dalle immagini e più sensibili al contesto testuale.
   * Questo approccio consente anche di testare come le diverse configurazioni influiscono sulla qualità delle risposte in presenza di immagini ingannevoli, offrendo così un metodo più diretto per diagnosticare e mitigare la sycophancy.
3. **Versatilità dei Parametri di Decoding**:
   * I parametri di decoding come temperatura, top-k e top-p sono generalmente sottovalutati nel contesto dei modelli multimodali. Esplorare come la loro regolazione possa influire sulla capacità del modello di risolvere problemi di sycophancy visiva offre una nuova dimensione per migliorare l'accuratezza e la robustezza del modello.
   * Inoltre, un'ottimizzazione specifica per task multimodali potrebbe essere generalizzata per migliorare la performance su una vasta gamma di applicazioni pratiche, come il visual question answering o la generazione automatica di descrizioni.

**Esperimento:**

1. **Impostazione del Test**:
   * **Dataset**: Utilizzare un set di immagini con contesti visivamente ambigui o ingannevoli, accompagnate da domande testuali che inducono errori visivi. Ad esempio, immagini che evidenziano oggetti irrilevanti rispetto alle domande testuali.
   * **Parametri di Decoding**: Testare diverse configurazioni dei parametri di decoding, inclusi valori di **temperatura**, **top-k** e **top-p**, e valutare come influenzano la generazione del testo.
     + *Temperatura*: Un parametro che controlla la casualità nelle risposte generate. Temperature basse promuovono risposte più determinate, mentre temperature alte favoriscono una maggiore esplorazione.
     + *Top-k*: Limita il numero di possibili token tra cui il modello può scegliere durante la generazione, riducendo la probabilità di selezionare token non comuni.
     + *Top-p*: Invece di limitare a un numero fisso di token, top-p sceglie solo quelli la cui somma cumulativa delle probabilità non supera una soglia specifica, migliorando la diversità delle risposte.
2. **Valutazione**:
   * Misurare la capacità del modello di evitare risposte errate causate da segnali visivi ingannevoli.
   * Confrontare l'accuratezza delle risposte tra le varie configurazioni di decoding, identificando quali combinazioni minimizzano la sycophancy visiva e migliorano la qualità delle risposte.
3. **Metriche di Performance**:
   * **Accuratezza delle Risposte**: Valutare se la risposta fornita dal modello è corretta, considerando sia il contesto visivo che testuale.
   * **Tasso di Sycophancy**: Quantificare la frequenza con cui il modello segue input visivi ingannevoli o errati, nonostante le indicazioni testuali suggeriscano altrimenti.
   * **Coerenza tra Input Visivo e Testuale**: Misurare quanto bene il modello riesce a integrare le informazioni visive e testuali per generare risposte coerenti.

**Risultati Attesi:**

* **Riduzione della Sycophancy Visiva**: Si prevede che configurazioni di decoding ottimizzate riducano l'influenza degli input visivi ingannevoli, migliorando così la precisione delle risposte multimodali.
* **Maggiore Robustezza**: Le configurazioni di decoding ottimizzate miglioreranno la capacità del modello di gestire scenari ambigui o difficili, garantendo una maggiore affidabilità in applicazioni pratiche, come la diagnosi medica basata su immagini o l'analisi automatica di video.
* **Applicabilità Generale**: Ottimizzare i parametri di decoding potrebbe avere un impatto positivo non solo sulla sycophancy visiva, ma anche su altri aspetti della generazione del testo nei modelli multimodali, migliorando la loro performance complessiva.

**Limiti:**

* **Sensibilità ai Parametri**: La regolazione dei parametri di decoding può migliorare la performance su specifici task, ma potrebbe non generalizzarsi bene su tutte le applicazioni. È necessario testare diverse combinazioni per ogni scenario multimodale, rendendo il processo complesso e computazionalmente costoso.
* **Effetto Negativo sulla Diversità delle Risposte**: Parametri di decoding troppo restrittivi (ad esempio, top-k basso) potrebbero limitare la diversità delle risposte generate, rendendo il modello meno flessibile in contesti dove è richiesta una maggiore varietà di output.

**Sviluppi Futuri:**

1. **Modelli Multimodali con Decoding Dinamico**:
   * Sviluppare sistemi che adattino dinamicamente i parametri di decoding in base alla complessità del contesto visivo e testuale. Ad esempio, un sistema potrebbe aumentare la temperatura quando il contesto visivo è particolarmente ambiguo per ridurre l'influenza di segnali visivi fuorvianti.
2. **Combinazione con Tecniche di Contrastive Decoding**:
   * Integrare questa ottimizzazione con tecniche di **contrastive decoding**, dove il modello viene costantemente messo alla prova generando alternative contrastanti per evitare la sycophancy visiva.
3. **Applicazione a Sistemi Real-Time**:
   * Implementare le ottimizzazioni nei modelli che operano in tempo reale, come i sistemi di navigazione autonoma o le interfacce uomo-macchina, dove la capacità di interpretare correttamente immagini e testo è cruciale.

**PROPOSTA INTERESSANTE**

### Proposta: Utilizzo di un Sistema di Reinforcement Learning con Feedback Automatico da AI per Ridurre la Sycophancy Visiva

#### Descrizione:

La proposta introduce un sistema basato su **Reinforcement Learning from AI Feedback (RLAIF)** per mitigare la sycophancy visiva nei modelli **Large Vision-Language Models (LVLMs)**. Invece di affidarsi al feedback umano, un modello AI sarà addestrato per rilevare automaticamente segnali di sycophancy visiva, valutando la coerenza tra input visivo e risposta testuale, e fornendo feedback diretto al modello LVLM durante il processo di addestramento.

#### Struttura della Soluzione:

1. **Modello AI per il Rilevamento della Sycophancy Visiva (AI-Detectors)**:
   * **Obiettivo**: Un secondo modello AI verrà addestrato specificamente per riconoscere quando una risposta generata dal LVLM è coerente con l'immagine o è influenzata da un suggerimento testuale fuorviante.
   * **Funzionamento**: Questo modello confronterà i contenuti visivi con le risposte generate, rilevando automaticamente la presenza di sycophancy visiva. Identificherà se la risposta è stata influenzata da bias testuali o se è corretta, basata su ciò che è visibile nell'immagine.
   * **Training del Modello**: Il modello AI-Detectors sarà addestrato su un dataset che contiene esempi annotati di sycophancy visiva e risposte corrette per imparare a distinguere i due scenari.
2. **Sistema di Feedback Automatico Basato su AI**:
   * **Premi e Penalità Automatizzati**: Il sistema di rilevazione fornirà un feedback immediato durante l'addestramento del LVLM:
     + **Penalità**: Quando viene rilevata una sycophancy visiva (una risposta basata su input testuali fuorvianti anziché visivi), il sistema ridurrà il punteggio del modello.
     + **Premio**: Se la risposta è coerente con i contenuti visivi dell'immagine, il sistema applicherà un premio, rinforzando il comportamento corretto.
   * **Modello Auto-adattivo**: Il sistema di feedback è integrato nel ciclo di reinforcement learning, permettendo al modello LVLM di apprendere autonomamente dalle penalità e dai premi senza necessità di intervento umano.

#### Implementazione:

1. **Training del Modello AI-Detectors**:
   * **Dataset**: Creare un dataset specializzato contenente coppie di immagini e risposte testuali annotate come "corrette" (coerenti con l'immagine) o "errate" (causate da sycophancy visiva).
   * **Modello Discriminativo**: Addestrare un modello discriminatore in grado di rilevare quando la risposta generata dal LVLM è influenzata da bias linguistici, rispetto a quando si basa correttamente sul contenuto visivo.
   * **Metriche**: Sviluppare metriche per misurare la correlazione tra i dati visivi e le risposte generate.
2. **Sistema di Feedback AI Integrato**:
   * **Collegamento del Modello AI-Detectors al LVLM**: Integrare il sistema di rilevazione automatica direttamente nel processo di addestramento del LVLM.
   * **Algoritmo di Reinforcement Learning**: Utilizzare un algoritmo di reinforcement learning che premi le risposte corrette (coerenti con l’immagine) e penalizzi quelle fuorviate da sycophancy. Il feedback automatico ottimizzerà i pesi del modello durante l'addestramento, senza necessità di supervisione umana.

#### Esperimento:

* **Dataset**: Utilizzare un dataset multimodale che contenga immagini con domande visivamente incoerenti o fuorvianti (leading queries) e risposte corrette. Includere casi in cui le immagini contengono informazioni che contraddicono il testo, per testare la capacità del modello di ignorare i suggerimenti testuali.
* **Modelli Coinvolti**:
  + **Modello Principale (LVLM)**: Il modello multimodale che si sta addestrando per ridurre la sycophancy visiva.
  + **Modello di Rilevazione (AI-Detectors)**: Il modello AI addestrato per riconoscere le risposte incoerenti tra testo e immagine e fornire feedback automatico.
* **Metriche**:
  + **Riduzione della Sycophancy Visiva**: Misurare il tasso di riduzione delle risposte basate su bias linguistici rispetto a un modello LVLM non addestrato con feedback AI.
  + **Coerenza Testo-Visiva**: Valutare la capacità del modello di fornire risposte che siano coerenti con l'input visivo.
  + **Accuratezza Complessiva**: Misurare l'accuratezza generale del LVLM nel rispondere a domande che coinvolgono immagini.

#### Risultati Attesi:

* **Significativa Riduzione della Sycophancy Visiva**: Il modello LVLM dovrebbe mostrare una drastica riduzione delle risposte basate su bias testuali, imparando a dare priorità alle informazioni visive.
* **Feedback Autonomo ed Efficiente**: Il sistema AI-Detectors fornirà feedback continuo e automatico, migliorando l'efficienza del processo di addestramento senza bisogno di supervisione umana.
* **Miglioramento della Generalizzazione**: Il modello LVLM, grazie al feedback automatico, dovrebbe essere in grado di gestire domande visivamente incoerenti in maniera più robusta e generalizzare le sue capacità a diversi tipi di input visivi.

#### Considerazioni Aggiuntive:

* **Affidabilità del Modello AI-Detectors**: È cruciale che il modello AI-Detectors sia sufficientemente accurato nel rilevare la sycophancy visiva. Se il sistema di rilevazione non è ben calibrato, potrebbe influire negativamente sull'addestramento del LVLM.
* **Generalizzazione e Scalabilità**: Il sistema di feedback automatico dovrebbe essere testato su una varietà di tipi di domande e contesti visivi per garantire la sua robustezza e scalabilità a diversi scenari.

### Conclusione:

Questa proposta introduce un sistema di feedback completamente automatizzato basato su un modello AI per rilevare e correggere la sycophancy visiva nei modelli LVLM, eliminando la necessità di feedback umano. Con l'integrazione di un ciclo di reinforcement learning guidato dal feedback AI, il modello LVLM può imparare a dare priorità ai contenuti visivi e migliorare la sua accuratezza nelle risposte, rendendo il processo di addestramento più rapido ed efficiente.

### Perché la proposta migliorerebbe l'attuale stato dell'arte?

1. **Eliminazione del feedback umano**:
   * **Stato attuale**: L'approccio **Reinforcement Learning from Human Feedback (RLHF)** è spesso costoso e richiede molto tempo, poiché si basa su annotatori umani per fornire feedback manuale durante l'addestramento del modello. Questo può portare a scalabilità limitata e bias umano nelle risposte.
   * **Miglioramento proposto**: L'utilizzo di un sistema automatico di feedback tramite un modello AI (AI-Detectors) per rilevare sycophancy visiva elimina la necessità di intervento umano. Questo rende il processo di addestramento **più rapido**, **meno costoso**, e **scalabile**, migliorando l'efficienza complessiva e riducendo il rischio di errori o pregiudizi umani.
2. **Feedback continuo e autonomo**:
   * **Stato attuale**: Nelle tecniche RLHF, il feedback umano è discontinuo e richiede interventi manuali in fase di addestramento. Ciò può rallentare il processo e portare a un ritardo nell'aggiornamento del modello.
   * **Miglioramento proposto**: Il sistema RLAIF propone un **feedback continuo e in tempo reale**, poiché il modello AI-Detectors è integrato direttamente nel ciclo di addestramento del LVLM. Questo porta a un **processo di ottimizzazione più fluido**, consentendo al modello di apprendere in modo più efficiente e di correggere i propri errori in modo autonomo.
3. **Riduzione delle allucinazioni e della sycophancy visiva**:
   * **Stato attuale**: Gli approcci attuali, come il **Leading Query Contrastive Decoding (LQCD)**, mitigano la sycophancy visiva confrontando risposte generate da query fuorvianti con quelle derivanti da query neutre. Tuttavia, questo approccio si limita a un confronto statico senza apprendimento dinamico o continuo.
   * **Miglioramento proposto**: Il sistema RLAIF introduce un ciclo di **reinforcement learning** che permette al modello di **imparare attivamente** e in modo iterativo a ignorare i suggerimenti fuorvianti. Questa dinamica di apprendimento migliora gradualmente la capacità del modello di ridurre sia la sycophancy che le allucinazioni, concentrandosi più sull'input visivo che su bias linguistici.
4. **Generalizzazione e adattabilità**:
   * **Stato attuale**: Le tecniche esistenti, come l'LQCD, sono relativamente rigide e mirano a risolvere specifiche situazioni di sycophancy legate a determinati tipi di query fuorvianti. Ciò potrebbe limitare la loro applicabilità a contesti diversi.
   * **Miglioramento proposto**: Un sistema di feedback automatico basato su AI può essere più **flessibile** e **generalizzabile**, permettendo al modello di adattarsi a una **varietà più ampia di domande**, immagini e contesti. Poiché il modello AI-Detectors è addestrato su diversi tipi di input visivi e testuali, può identificare e gestire meglio situazioni complesse che causano sycophancy visiva in più scenari.
5. **Riduzione dei costi e della complessità del feedback**:
   * **Stato attuale**: Il feedback umano è costoso, sia in termini di risorse finanziarie che di tempo, e non è sempre accurato o coerente.
   * **Miglioramento proposto**: Automatizzando il processo di feedback con un modello AI, si **riduce drasticamente il costo** di addestramento del modello. Inoltre, il feedback sarà più **coerente e privo di bias umani**, garantendo una maggiore affidabilità durante il processo di ottimizzazione.

### Limiti della proposta

1. **Affidabilità del modello di rilevazione (AI-Detectors)**:
   * Il successo di questa proposta dipende fortemente dall'accuratezza del modello AI-Detectors. Se questo modello non riesce a riconoscere correttamente quando il LVLM produce una risposta basata su bias testuali o visivi, potrebbe fornire **feedback errato** al modello principale. Un feedback inaccurato potrebbe addirittura peggiorare le prestazioni del LVLM, causando overfitting o comportamenti non desiderati.
2. **Dipendenza dal dataset di addestramento**:
   * Il modello AI-Detectors deve essere addestrato su un dataset ricco e diversificato, che includa un'ampia gamma di casi di sycophancy visiva e risposte corrette. Se il dataset non è sufficientemente variegato, il modello potrebbe non generalizzare bene in scenari reali, risultando in **prestazioni subottimali**. La qualità del dataset influenzerà direttamente l'efficacia della rilevazione automatica e, di conseguenza, del feedback.
3. **Potenziale complessità computazionale**:
   * L'aggiunta di un secondo modello AI per il rilevamento della sycophancy introduce **ulteriore complessità computazionale**. L'integrazione di due modelli che lavorano simultaneamente potrebbe richiedere risorse significative in termini di calcolo e memoria. Per applicazioni in tempo reale, questo potrebbe risultare limitante, soprattutto in scenari dove la latenza deve essere minima.
4. **Sfiducia nei modelli automatizzati**:
   * L'automazione completa del feedback potrebbe ridurre il controllo umano sul processo di addestramento. In alcuni casi, la **mancanza di supervisione umana** potrebbe portare a situazioni in cui il sistema di AI-Detectors commette errori che non vengono corretti tempestivamente, compromettendo il risultato finale. Questo potrebbe essere un limite significativo, soprattutto in ambienti dove la qualità del modello è critica.
5. **Limitazioni nel rilevamento di ambiguità visive**:
   * In scenari in cui le informazioni visive sono ambigue o difficili da interpretare, il modello AI-Detectors potrebbe avere difficoltà a distinguere tra una risposta corretta e una basata su sycophancy visiva. In questi casi, il modello di rilevazione potrebbe fornire un **feedback inappropriato**, causando errori nella fase di apprendimento del LVLM.

**BiasFinder: Metamorphic Test Generation to Uncover Bias for Sentiment Analysis Systems**

5 ottobre 2021

<https://arxiv.org/pdf/2102.01859>

**~~COME MIGLIORARE LO STATO DELL’ARTE~~**

### ~~Integrazione di Modelli di Explainable AI (XAI)~~

**~~Descrizione~~**~~: Integrare tecniche di Explainable AI (XAI) per interpretare e spiegare le decisioni dei modelli di Sentiment Analysis riguardo a previsioni pregiudiziali. Questo aiuta a identificare esattamente quali parti di un testo innescano una previsione di bias.~~

* **~~Aspetto Quantitativo~~**~~: Utilizzare metodi di XAI come LIME (Local Interpretable Model-agnostic Explanations) o SHAP (SHapley Additive exPlanations) per generare spiegazioni dettagliate su migliaia di predizioni di sentiment che risultano pregiudiziali. Si possono usare visualizzazioni per identificare modelli ripetitivi nel bias.~~
* **~~Aspetto Qualitativo~~**~~: Analizzare i risultati spiegabili per identificare esattamente dove e come i modelli stanno introducendo pregiudizi. Questo permetterebbe di costruire modelli SA più trasparenti e di risolvere bias specifici.~~

**~~Miglioramento~~**~~: La spiegabilità è cruciale per garantire la fiducia nei sistemi di IA. Integrare XAI migliora significativamente la capacità di diagnosticare e correggere i bias, contribuendo a sviluppare modelli più trasparenti e responsabili.~~

**~~PERCHE’ MIGLIORA LO STATO DELL’ARTE~~**

**~~Perché Migliora lo Stato dell’Arte:~~**

1. **~~Maggiore Trasparenza dei Modelli~~**~~:~~
   * ~~I sistemi di sentiment analysis spesso operano come "black box", dove le decisioni prese dal modello risultano opache. L'integrazione di tecniche XAI permette di esporre quali caratteristiche o parti di un testo influenzano maggiormente la decisione del modello.~~
   * ~~Con strumenti come LIME e SHAP, si possono generare visualizzazioni che mostrano i contributi di ciascun termine o frase nella previsione del sentimento, facilitando l'identificazione di pattern di bias e rendendo i modelli più trasparenti e interpretabili.~~
2. **~~Identificazione di Bias a Livello Granulare~~**~~:~~
   * ~~I modelli di sentiment analysis possono introdurre bias che derivano da determinate parole, frasi o contesti culturali. XAI permette di isolare questi fattori, spiegando quale parte del testo ha avuto un impatto sulla previsione e se questa era influenzata da bias preesistenti.~~
   * ~~Questo approccio permette un'analisi qualitativa più approfondita, individuando esattamente dove il bias viene introdotto, come nei casi di linguaggio culturalmente specifico o pregiudizi legati a genere, razza o etnia.~~
3. **~~Miglioramento della Fiducia nel Sistema di IA~~**~~:~~
   * ~~Rendere il processo decisionale più interpretabile non solo aumenta la trasparenza ma costruisce anche fiducia negli utenti e nei ricercatori, che possono verificare il comportamento del modello e correggere eventuali bias sistematici.~~
   * ~~Ciò è particolarmente importante nelle applicazioni di sentiment analysis che influenzano decisioni di marketing, valutazioni dei consumatori o contenuti editoriali, dove i bias possono avere effetti economici o sociali rilevanti.~~

**~~Metodologia:~~**

1. **~~Aspetto Quantitativo: Utilizzo di XAI su Larga Scala~~**~~:~~
   * ~~Utilizzare LIME o SHAP per generare spiegazioni dettagliate su un gran numero di previsioni di sentiment analizzate da BiasFinder.~~
   * ~~Queste tecniche creano rappresentazioni visive che permettono di capire quali caratteristiche di input (ad esempio, parole o frasi specifiche) hanno contribuito maggiormente alla previsione. Ad esempio, una parola con un impatto sproporzionato nella previsione negativa potrebbe rivelare un pregiudizio contro determinati termini associati a genere o razza.~~
   * ~~Generare queste spiegazioni su larga scala e creare aggregati di analisi su migliaia di esempi permetterà di trovare pattern ripetitivi di bias.~~
2. **~~Aspetto Qualitativo: Analisi Dettagliata delle Previsioni di Bias~~**~~:~~
   * ~~Analizzare qualitativamente i risultati spiegabili per scoprire dove esattamente i modelli di sentiment analysis stanno introducendo pregiudizi. Questa analisi si concentra su casi specifici in cui LIME o SHAP mostrano che determinate caratteristiche testuali (come il genere di un personaggio menzionato) portano a un cambiamento ingiustificato nelle previsioni.~~
   * ~~Questo consente un focus sulla rimozione di bias a livello locale, correggendo i comportamenti errati dei modelli in modo mirato.~~

**~~Esperimento:~~**

1. **~~Set di Dati~~**~~:~~
   * ~~Utilizzare dataset ampi e diversificati, come recensioni di prodotti, post sui social media o commenti editoriali, per applicare BiasFinder in combinazione con XAI.~~
   * ~~Analizzare le previsioni fatte su testi che coprono vari aspetti culturali, generazionali e linguistici per osservare dove emergono bias specifici.~~
2. **~~Tecniche di XAI~~**~~:~~
   * ~~Applicare LIME e SHAP per generare visualizzazioni di spiegabilità su ciascuna predizione, rivelando i contributi delle parole chiave alle previsioni.~~
   * ~~Utilizzare queste tecniche per spiegare in modo trasparente come i modelli decidono se un sentimento è positivo, negativo o neutro, e come i bias influenzano queste decisioni.~~
3. **~~Valutazione~~**~~:~~
   * ~~Valutare la frequenza con cui parole o frasi connotate culturalmente influenzano negativamente (o positivamente) la previsione del sentiment. Ad esempio, misurare quanto spesso termini associati a minoranze etniche portano a valutazioni negative ingiustificate.~~
   * ~~Creare metriche personalizzate che calcolino il grado di impatto del bias osservato, utilizzando le spiegazioni fornite da XAI.~~

**~~Risultati Attesi:~~**

1. **~~Riduzione del Bias~~**~~:~~
   * ~~Si prevede che l’integrazione di XAI aiuti a identificare e correggere i pregiudizi nei modelli di sentiment analysis. Con la possibilità di tracciare esattamente quali elementi in un testo innescano previsioni di bias, sarà possibile affinare i modelli per migliorare la loro equità.~~
2. **~~Maggiore Trasparenza~~**~~:~~
   * ~~Aumentare la trasparenza nei modelli di sentiment analysis rafforzerà la fiducia nei risultati generati. Gli utenti saranno in grado di capire come i modelli prendono decisioni, potendo intervenire per migliorare il sistema laddove necessario.~~
3. **~~Creazione di Modelli Più Equi~~**~~:~~
   * ~~Con l'analisi dei contributi spiegabili, i modelli di sentiment analysis possono essere aggiornati per eliminare le fonti di bias, rendendo i modelli più equi nei confronti di diverse categorie sociali, di genere o etniche. Ciò avrà un impatto positivo nelle applicazioni aziendali e sociali.~~

**~~Limiti:~~**

1. **~~Computational Cost~~**~~:~~
   * ~~L'applicazione di tecniche di XAI come LIME o SHAP su larga scala può essere computazionalmente intensiva, specialmente quando si devono spiegare migliaia di predizioni.~~
2. **~~Over-Interpretazione~~**~~:~~
   * ~~Sebbene le tecniche di XAI forniscano spiegazioni per le decisioni del modello, c'è il rischio di over-interpretare i risultati. Le spiegazioni fornite da LIME e SHAP potrebbero non riflettere sempre accuratamente l'intero processo decisionale del modello, ma solo una parte di esso.~~

**~~Sviluppi Futuri:~~**

1. **~~Automatizzare la Correzione del Bias~~**~~:~~
   * ~~Sviluppare sistemi che non solo identifichino i bias grazie a XAI, ma che siano anche in grado di correggerli automaticamente, aggiornando i modelli in tempo reale.~~
2. **~~Integrazione con Altri Strumenti di Mitigazione del Bias~~**~~:~~
   * ~~Combinare XAI con tecniche di debiasing esistenti, come il retraining con dati corretti o l'applicazione di regolarizzazioni specifiche per penalizzare le decisioni influenzate da bias.~~
3. **~~Estensione del Framework a Multimodalità~~**~~:~~
   * ~~Applicare tecniche di XAI e BiasFinder a modelli multimodali, in modo da diagnosticare e mitigare il bias non solo nel testo ma anche in modelli che combinano input testuali e visivi (ad esempio, nelle applicazioni di sentiment analysis che utilizzano anche immagini o video).~~

**TIBET: Identifying and Evaluating Biases in Text-to-Image Generative Models**

17 Luglio 2024

<https://arxiv.org/pdf/2312.01261>

### ~~Creazione di un Framework di Testing Automatizzato per la Mitigazione dei Bias (Forse tesi più applicativa?)~~

**~~Descrizione~~**~~: Integrare TIBET con un framework di testing automatizzato che non solo rileva i bias, ma applica anche tecniche di mitigazione in tempo reale, come ITI-GEN. Questo potrebbe diventare uno standard per testare i modelli TTI prima della loro implementazione su larga scala.~~

* **~~Aspetto Quantitativo~~**~~: Sviluppare uno strumento che consenta di testare i modelli TTI in modo automatizzato su ampi set di dati e con prompt diversificati, applicando automaticamente tecniche di mitigazione dei bias dove necessario. Valutare il miglioramento nella riduzione dei bias post-mitigazione utilizzando le metriche CAS e MAD.~~
* **~~Aspetto Qualitativo~~**~~: Documentare e analizzare il processo di mitigazione, studiando quali bias sono stati più difficili da ridurre e quali prompt hanno presentato le maggiori sfide.~~

**~~Miglioramento~~**~~: Un framework di testing automatizzato per la mitigazione dei bias offrirebbe un significativo passo avanti rispetto ai modelli attuali, permettendo un monitoraggio continuo dei bias durante il ciclo di vita del modello. Questo approccio fornirebbe anche una soluzione operativa per i bias nei modelli TTI, rendendo questi modelli più affidabili.~~

**~~PERCHE’ MIGLIORA LO STATO DELL’ARTE~~**

**~~Perché Migliora lo Stato dell’Arte~~**

1. **~~Automazione nel Testing e Mitigazione dei Bias~~**~~:~~
   * ~~Attualmente, molte tecniche per rilevare e mitigare i bias nei modelli di generazione text-to-image richiedono un intervento manuale, che non è scalabile. Un framework automatizzato che integra rilevamento e mitigazione consente un monitoraggio continuo, accelerando il processo di riduzione del bias.~~
   * ~~La mitigazione in tempo reale riduce il rischio che modelli TTI generino immagini inappropriate o pregiudizievoli, migliorando la qualità e l'affidabilità del modello in tempo reale.~~
2. **~~Applicazione su Larga Scala~~**~~:~~
   * ~~L'automazione del testing e della mitigazione dei bias permette di testare e correggere modelli TTI su vasta scala e su dataset diversificati. Questo è cruciale, considerando la crescente applicazione di modelli generativi in contesti commerciali e creativi.~~
   * ~~In un ambiente operativo, come le piattaforme di immagini generate per il marketing o i social media, il framework permetterebbe di monitorare e correggere i bias nei modelli continuamente, prevenendo la diffusione di immagini problematiche.~~
3. **~~Standardizzazione del Processo di Test~~**~~:~~
   * ~~Attualmente, non esiste un processo standard per valutare e mitigare i bias nei modelli TTI. La creazione di un framework automatizzato fornisce una base di testing uniforme, applicabile a diversi modelli TTI, garantendo che tutti i modelli passino attraverso lo stesso rigoroso processo di valutazione prima della loro implementazione.~~

**~~Metodologia~~**

1. **~~Aspetto Quantitativo: Test Automatizzati e Mitigazione del Bias~~**~~:~~
   * **~~Testing Automatizzato~~**~~: Sviluppare uno strumento che permette di testare i modelli TTI su ampi set di dati con prompt diversificati (ad esempio, richieste basate su genere, etnia, professione, ecc.), rilevando bias espliciti o impliciti nelle immagini generate.~~
   * **~~Tecniche di Mitigazione in Tempo Reale~~**~~: Integrare metodi come~~ **~~ITI-GEN~~**~~, che applicano la mitigazione del bias direttamente durante il processo di generazione delle immagini. Ogni volta che viene rilevato un bias, il framework modifica il processo di generazione per evitare rappresentazioni problematiche.~~
   * **~~Valutazione delle Prestazioni~~**~~: Usare metriche quantitative come~~ **~~CAS (Causal Attribution Score)~~** ~~e~~ **~~MAD (Mean Absolute Deviation)~~** ~~per misurare l'impatto delle tecniche di mitigazione sulla riduzione dei bias. Queste metriche quantificano la distanza tra le immagini generate dal modello e i valori attesi (in termini di equità e rappresentazione).~~
2. **~~Aspetto Qualitativo: Documentazione del Processo di Mitigazione~~**~~:~~
   * **~~Studio dei Bias Persistenti~~**~~: Analizzare qualitativamente i prompt e i risultati del framework per documentare i casi in cui la mitigazione ha avuto più difficoltà. Ad esempio, documentare i prompt che hanno mostrato persistenti bias di genere o razziali nonostante l'applicazione delle tecniche di mitigazione.~~
   * **~~Feedback Iterativo~~**~~: Creare un ciclo di feedback per il modello. Ogni volta che il framework rileva e corregge un bias, questo processo viene registrato per migliorare le future versioni del modello e rafforzare le tecniche di mitigazione.~~
3. **~~Testing Cross-Modale~~**~~:~~
   * **~~Valutazione in Contesti Multi-Modal~~**~~: Verificare come il framework gestisce richieste complesse che coinvolgono contesti multi-modal (immagini più testo). Per esempio, testare prompt che includono contraddizioni tra l’immagine e il testo, verificando se il modello corregge i bias senza compromettere la qualità visiva dell'output.~~

**~~Risultati Attesi~~**

1. **~~Riduzione Significativa del Bias nei Modelli TTI~~**~~:~~
   * ~~Si prevede che il framework permetta una riduzione significativa dei bias nelle immagini generate, migliorando la rappresentazione di genere, etnia e altri gruppi sociali sensibili. La mitigazione in tempo reale attraverso tecniche come ITI-GEN assicura che il modello possa essere adattato istantaneamente alle esigenze di equità e accuratezza.~~
2. **~~Maggiore Affidabilità e Robustezza~~**~~:~~
   * ~~Grazie alla capacità del framework di monitorare costantemente il comportamento del modello, si riduce il rischio di generare immagini problematiche, rendendo i modelli TTI più affidabili per applicazioni commerciali o sensibili (ad esempio, campagne pubblicitarie o progetti educativi).~~
3. **~~Standardizzazione del Processo di Testing~~**~~:~~
   * ~~La standardizzazione del processo di testing e mitigazione attraverso un framework automatizzato rappresenta un punto di riferimento per la comunità di ricerca e le aziende che sviluppano modelli TTI, garantendo che ogni modello venga rigorosamente testato prima del rilascio.~~

**~~Limiti~~**

1. **~~Costi Computazionali~~**~~:~~
   * ~~L'esecuzione di test automatizzati su larga scala e in tempo reale può risultare intensiva dal punto di vista computazionale, richiedendo significative risorse di calcolo, soprattutto per modelli più complessi o con grandi volumi di prompt da processare.~~
2. **~~Bias Complessi o Sottili~~**~~:~~
   * ~~Nonostante la capacità del framework di rilevare e correggere molti tipi di bias, alcuni bias più sottili o complessi potrebbero richiedere tecniche più avanzate di mitigazione, o potrebbero non essere risolti completamente con metodi automatizzati.~~

**~~Sviluppi Futuri~~**

1. **~~Ottimizzazione del Framework~~**~~:~~
   * ~~Ottimizzare ulteriormente il framework per ridurre i costi computazionali, ad esempio utilizzando tecniche di compressione dei modelli o ottimizzazioni hardware per eseguire il testing e la mitigazione in modo più efficiente.~~
2. **~~Espansione a Scenari Multi-Modali Complessi~~**~~:~~
   * ~~Estendere il framework per gestire scenari più complessi, dove le immagini generate sono accompagnate da testo, audio o altri input multi-modali, garantendo una mitigazione del bias su tutti i livelli della generazione.~~
3. **~~Integrazione con Sistemi di Feedback Umano~~**~~:~~
   * ~~Integrare feedback umano nel processo di testing automatizzato. Gli utenti finali potrebbero fornire feedback in tempo reale su casi problematici che il sistema non riesce a gestire, migliorando ulteriormente la qualità del modello.~~

**Large Language Models are Geographically Biased**

5 Febbraio 2024

<https://arxiv.org/pdf/2402.02680>

**COME MIGLIORARE LO STATO DELL’ARTE**

**Quantificazione del bias tramite modelli di rete semantica basati su risposte LLM**

**Obiettivo**: Creare un modello di rete semantica che rappresenti le associazioni tra concetti geografici nelle risposte dei LLM e calcolare la distribuzione del bias.

**Metodologia**: Analizzare grandi quantità di output del modello, costruendo una rete che mappa le associazioni semantiche tra termini legati a paesi, culture e regioni geografiche. Successivamente, applicare algoritmi di **network analysis** per identificare pattern di bias geografico.

**Esempio**: Valutare quanto frequentemente i concetti negativi (es. "disoccupazione", "corruzione") sono collegati a determinate aree geografiche rispetto ad altre.

**Output**: Un modello di rete semantica per visualizzare e quantificare i bias nelle associazioni geografiche.

### ****Confronto Cross-Lingua****

#### Esempio

Confronterai i bias geografici presenti in modelli multilingua come GPT-4 in inglese, spagnolo, arabo, e cinese. Ad esempio, analizzerai se la valutazione dell'attrattività varia quando il modello è interrogato in spagnolo (per l’America Latina) rispetto a inglese (per l'Europa e il Nord America).

#### Metodologia

* **Creazione di prompt multilingua**: Crea prompt uniformi in diverse lingue per stimolare le risposte del modello su argomenti come "intelligenza" o "etica del lavoro".
* **Analisi delle risposte**: Usa metriche come il Bias Score per misurare la differenza di bias geografico tra le diverse lingue e confronta le risposte fornite dai modelli.

#### Risultati attesi

Ci aspettiamo di scoprire che i modelli multilingua mostreranno variazioni nel bias in base alla lingua, con lingue meno rappresentate nei dataset (come l’arabo o il cinese) che potrebbero essere più distorte nelle loro risposte rispetto a lingue come l’inglese.

### ****Esperimenti con Prompting Mitigatore****

#### Esempio

Creerai prompt specifici per ridurre il bias, come "Dai una valutazione senza considerare l'area geografica" o "Considera solo gli aspetti socioeconomici positivi di questa regione." Ad esempio, potresti chiedere al modello di valutare la "creatività" di una regione, mitigando pregiudizi geografici negativi.

#### Metodologia

* **Progettazione di prompt**: Crea due versioni per ogni domanda: una neutra e una mitigata (es. "Considera solo le qualità positive di questa regione").
* **Confronto dei risultati**: Misura i cambiamenti nel punteggio di bias tra le due versioni dei prompt e osserva come l'uso di prompt mitigatori influenzi le risposte del modello.

#### Risultati attesi

I prompt mitigatori dovrebbero ridurre significativamente il livello di bias nelle risposte, portando a valutazioni più neutrali e meno influenzate dai pregiudizi geografici. Questo risultato mostrerebbe che il bias nei LLM può essere ridotto con un prompting ben progettato.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**1. Quantificazione del Bias tramite Modelli di Rete Semantica Basati su Risposte LLM**

**Obiettivo:** Creare una rete semantica che mappa le associazioni tra concetti geografici, utilizzando le risposte dei modelli linguistici di grandi dimensioni (LLM). L'obiettivo è quantificare il bias osservando le relazioni semantiche tra concetti legati a paesi, culture e regioni.

**Perché migliora lo stato dell'arte:**  
Attualmente, il bias nei LLM viene misurato principalmente con metodi statistici come correlazioni o conteggi di termini associati a stereotipi. La creazione di **reti semantiche** va oltre, permettendo una comprensione visiva e quantitativa delle associazioni semantiche tra concetti geografici e identificando le aree geografiche che ricevono giudizi distorti. Inoltre, la rete permette di analizzare le **relazioni tra i concetti** che contribuiscono ai bias, fornendo informazioni su come si formano e si rafforzano.

**Metodologia:**

1. **Raccolta di dati:** Raccogli una grande quantità di output generati da LLM in risposta a prompt relativi a paesi e regioni.
2. **Costruzione della rete semantica:** Analizza gli output per identificare i termini geografici chiave e crea una rete in cui i nodi rappresentano concetti come "intelligenza", "creatività", "corruzione", "sviluppo", mentre i collegamenti rappresentano le relazioni semantiche tra questi concetti e le regioni.
3. **Algoritmi di analisi delle reti:** Applica tecniche di network analysis, come la **centralità** o la **modularità**, per identificare le regioni più fortemente associate a concetti negativi o positivi.

**Esempio:**  
Un esempio potrebbe essere quello di calcolare quanto frequentemente concetti come "povertà" o "disoccupazione" siano collegati a paesi africani rispetto a quelli europei o nordamericani, mostrando una rappresentazione visuale delle connessioni geografiche e semantiche che contribuiscono al bias.

**Risultati attesi:**  
Un **modello di rete semantica** che visualizza in modo esplicito i bias nelle associazioni geografiche e che consente una quantificazione precisa di tali bias. Questo approccio migliorerebbe la comprensione dei bias, evidenziando in quali contesti semantici appaiono più frequentemente e quali regioni ne sono maggiormente colpite.

**2. Confronto Cross-Lingua per l'Analisi dei Bias Geografici**

**Obiettivo:** Confrontare i bias geografici presenti nei modelli multilingua, per esaminare se i modelli mostrano variazioni di bias a seconda della lingua in cui vengono interrogati, evidenziando differenze dovute alla rappresentazione linguistica e culturale.

**Perché migliora lo stato dell'arte:**  
La maggior parte degli studi attuali sui bias geografici nei LLM si concentra su un'unica lingua, spesso l'inglese. Questo approccio **multilingua** introduce un confronto incrociato che rivela se i modelli presentano bias diversi in lingue meno rappresentate nei dataset di addestramento (come l'arabo o il cinese). Questo è particolarmente importante poiché molte culture e aree geografiche possono essere sottorappresentate in determinate lingue, aumentando il rischio di bias nelle risposte.

**Metodologia:**

1. **Creazione di prompt multilingua:** Crea un insieme uniforme di prompt in più lingue (ad esempio, inglese, spagnolo, arabo, cinese), che chiedano al modello valutazioni su argomenti come "intelligenza", "etica del lavoro" o "creatività" legati a diverse regioni.
2. **Analisi delle risposte:** Usa metriche come il **Bias Score** per quantificare la differenza nel bias geografico tra le diverse lingue. Confronta i risultati delle valutazioni in lingue diverse, osservando quali regioni ricevono giudizi più negativi o positivi a seconda della lingua utilizzata.

**Esempio:**  
Supponiamo che il modello venga interrogato in spagnolo su paesi dell'America Latina e in inglese su paesi europei o nordamericani. Si potrebbe osservare se il modello tenda a valutare più positivamente l'Europa o il Nord America quando utilizza l'inglese, rispetto alle valutazioni in spagnolo per l'America Latina, rivelando un possibile bias linguistico.

**Risultati attesi:**  
Ci si aspetta che lingue meno rappresentate nei dataset di addestramento (come l'arabo o il cinese) mostrino una maggiore distorsione nei bias geografici rispetto a lingue più rappresentate come l'inglese o lo spagnolo. Questo confronto evidenzierebbe la necessità di migliorare l'equilibrio linguistico nei modelli multilingua.

**3. Esperimenti con Prompting Mitigatore per Ridurre il Bias Geografico**

**Obiettivo:** Creare e testare prompt specifici per ridurre il bias geografico, utilizzando strategie di prompting che incoraggino il modello a evitare associazioni stereotipate.

**Perché migliora lo stato dell'arte:**  
Mentre la maggior parte degli studi sul bias nei LLM si concentra sul rilevamento, questo approccio proattivo tenta di **mitigare il bias** direttamente nel prompting. L'idea è quella di incoraggiare risposte neutrali o di mettere in risalto aspetti positivi delle regioni, limitando l'influenza dei bias storici e culturali presenti nei dati di addestramento.

**Metodologia:**

1. **Progettazione di prompt mitigatori:** Per ogni domanda, crea due versioni: una neutra e una mitigatrice. Ad esempio, per valutare la creatività di una regione, il prompt mitigatore potrebbe essere "Valuta solo le qualità creative di questa regione senza considerare altri fattori socioeconomici".
2. **Confronto dei risultati:** Confronta i punteggi di bias ottenuti dalle risposte ai prompt neutri e mitigatori. Misura quanto il prompting mitigatore riduca il livello di bias nelle risposte, utilizzando metriche quantitative come il **Bias Score**.

**Esempio:**  
Chiedi al modello di valutare l'intelligenza in una regione utilizzando prima un prompt neutro come "Valuta l'intelligenza delle persone in questa regione" e poi un prompt mitigatore come "Valuta solo le qualità intellettuali positive di questa regione". Analizza se il prompting mitigatore produce risposte più equilibrate e meno influenzate da pregiudizi.

**Risultati attesi:**  
I prompt mitigatori dovrebbero ridurre il livello di bias nelle risposte del modello, portando a valutazioni più neutrali e meno influenzate da stereotipi geografici. Questo risultato dimostrerebbe che una strategia di prompting ben progettata può essere una soluzione efficace per ridurre il bias nei LLM.

**PROPOSTA INTERESSANTE**

**Titolo della Tesi**:  
*Quantificazione e Mitigazione del Bias Geografico nei Modelli di Linguaggio di Grandi Dimensioni (LLM) attraverso Knowledge Graphs e Analisi Cross-Lingua*

**Obiettivo Generale**

L’obiettivo di questa tesi è analizzare e ridurre il bias geografico presente nei modelli di linguaggio di grandi dimensioni (LLM), come GPT-4, attraverso la costruzione di **Knowledge Graphs (KGs)** che mappino le associazioni semantiche tra concetti e aree geografiche. Il progetto esplorerà anche come il bias vari tra diverse lingue e proporrà tecniche di **prompting** per mitigarlo.

**Obiettivi Specifici**

1. **Quantificare il bias geografico** nei LLM utilizzando Knowledge Graphs per rappresentare le relazioni semantiche tra concetti socioeconomici e aree geografiche.
2. **Confrontare il bias tra lingue diverse**, analizzando se e come i bias si manifestano in modelli multilingua interrogati in lingue differenti.
3. **Ridurre il bias** utilizzando tecniche di prompting, progettate per mitigare le distorsioni rilevate nei modelli.

**Metodologia**

Il progetto si articolerà in tre fasi principali, ciascuna incentrata su un aspetto chiave dell'analisi del bias geografico.

**Fase 1: Quantificazione del Bias Geografico tramite Knowledge Graphs**

* **Obiettivo**: Costruire un **Knowledge Graph** che mappi le associazioni tra concetti socioeconomici e aree geografiche nelle risposte dei LLM, e quantificare il bias tramite analisi di rete.
* **Metodologia**:
  + **Estrazione dei dati**: Generare un ampio insieme di risposte dai LLM su argomenti legati a concetti socioeconomici e aree geografiche, utilizzando prompt standardizzati.
  + **Costruzione del KG**: Creare un Knowledge Graph in cui i nodi rappresentano concetti come *"corruzione"*, *"crescita economica"*, *"qualità della vita"*, e aree geografiche specifiche (continenti, paesi, regioni). Gli archi tra i nodi rappresenteranno le relazioni semantiche, come la frequenza con cui i concetti sono collegati a una determinata area.
  + **Fonti esterne**: Integrare dati da fonti esterne come **Wikidata**, **DBpedia** e **OpenStreetMap** per arricchire il KG, aggiungendo conoscenze contestuali e storiche che possano spiegare certe associazioni.
  + **Analisi del KG**:
    - Utilizzare **metriche di centralità** per identificare le aree geografiche o i concetti più frequentemente associati in modo positivo o negativo.
    - Applicare tecniche di **community detection** per scoprire gruppi di concetti che tendono a essere associati a determinate regioni.
    - Utilizzare tecniche di **inferenza semantica** per scoprire bias impliciti nelle risposte del modello, analizzando non solo le relazioni dirette ma anche quelle indirette tra concetti e aree geografiche.
* **Esempio**: Se concetti negativi come *"disoccupazione"* o *"corruzione"* sono fortemente connessi ad aree geografiche come l'America Latina nel KG, potresti dimostrare un bias associativo che il modello perpetua.
* **Output atteso**: Un KG che visualizza e quantifica il bias nelle associazioni geografiche, rivelando pattern di bias espliciti e impliciti.

**Fase 2: Confronto del Bias Geografico tra Lingue Diverse**

* **Obiettivo**: Verificare se e come il bias geografico vari quando i LLM sono interrogati in lingue diverse, utilizzando KGs multilingua per confrontare i risultati.
* **Metodologia**:
  + **Creazione di prompt multilingua**: Creare un insieme uniforme di prompt in più lingue (inglese, spagnolo, arabo, cinese) su temi socioeconomici, culturali e geografici.
  + **Costruzione di KGs per lingue diverse**: Generare risposte nelle diverse lingue e costruire un KG per ciascuna lingua, seguendo lo stesso approccio della Fase 1. I nodi rappresenteranno gli stessi concetti e regioni, mentre gli archi rifletteranno le associazioni tra concetti e aree geografiche in ciascuna lingua.
  + **Confronto Cross-lingua**: Utilizzare tecniche di **graph alignment** per confrontare i KGs delle diverse lingue e identificare differenze nelle associazioni semantiche. Analizzare se determinati concetti sono più o meno collegati a specifiche aree geografiche a seconda della lingua.
  + **Bias Score**: Calcolare un **Bias Score** per ciascuna lingua, misurando le differenze nell’associazione di concetti negativi o positivi con le aree geografiche.
* **Esempio**: Analizzare se il concetto di *"crescita economica"* è più frequentemente associato all'Europa quando interrogato in inglese, e all'America Latina quando interrogato in spagnolo.
* **Output atteso**: Differenze significative nei bias geografici tra le lingue, con lingue meno rappresentate nei dataset di addestramento (come arabo o cinese) che potrebbero mostrare maggiori distorsioni rispetto a lingue più comuni come l'inglese.

**Fase 3: Mitigazione del Bias tramite Prompting**

* **Obiettivo**: Ridurre il bias geografico nelle risposte dei LLM utilizzando prompt specificamente progettati.
* **Metodologia**:
  + **Prompting mitigatore**: Creare due versioni per ogni domanda—una neutra e una mitigata. I prompt mitigati richiederanno al modello di escludere esplicitamente i pregiudizi geografici (ad esempio, “Dai una valutazione senza considerare l'area geografica” o “Considera solo le qualità positive di questa regione”).
  + **KG adattivo**: Creare KGs basati sulle risposte generate con i prompt mitigati, confrontando le nuove associazioni semantiche con quelle originali per misurare l’effetto del prompting.
  + **Confronto del Bias**: Utilizzare le stesse tecniche di analisi dei KGs (centralità, community detection, inferenza semantica) per verificare come i prompt mitigatori modificano le associazioni tra concetti e aree geografiche.
* **Esempio**: Se un prompt originale associa frequentemente *"corruzione"* con un determinato paese, l'uso di un prompt mitigato dovrebbe ridurre tale associazione, favorendo una rappresentazione più equilibrata.
* **Output atteso**: Una riduzione significativa delle associazioni di bias nei KGs generati con i prompt mitigatori, dimostrando che il bias geografico può essere attenuato attraverso tecniche di prompting mirato.

**Strumenti e Tecnologie**

* **Neo4j** o **GraphDB**: Per costruire e interrogare i Knowledge Graphs.
* **SPARQL**: Per l'integrazione e la gestione di dati semantici da fonti come Wikidata e DBpedia.
* **Algoritmi di network analysis**: Per misurare centralità, rilevare comunità di concetti e calcolare le relazioni semantiche.
* **Librerie per NLP**: Come **SpaCy** o **Transformers**, per l'analisi delle risposte generate dai LLM e l'elaborazione di testi multilingua.

**Risultati Attesi**

1. Un **Knowledge Graph** che quantifica e visualizza le relazioni tra concetti socioeconomici e aree geografiche, evidenziando bias espliciti e impliciti nelle risposte dei LLM.
2. Un'analisi dettagliata delle **differenze cross-lingua** nel bias geografico, con variazioni significative in base alla lingua utilizzata.
3. **Riduzione del bias geografico** attraverso prompt mitigatori, dimostrando che il bias nei LLM può essere gestito con tecniche di prompting ben progettate.

**Impatto e Contributo della Ricerca**

Questa tesi contribuirà alla comprensione e riduzione del bias geografico nei modelli di linguaggio di grandi dimensioni, proponendo un approccio innovativo basato sui **Knowledge Graphs** e il **prompting mitigatore**. I risultati avranno implicazioni significative per migliorare l'equità dei LLM in applicazioni come il giornalismo, la consulenza aziendale, e i sistemi di raccomandazione, garantendo che le risposte siano meno influenzate da pregiudizi geografici.

### Contributo allo Stato dell’Arte

Questa ricerca rappresenta un avanzamento significativo rispetto allo stato dell'arte, in quanto introduce l'uso di **Knowledge Graphs (KGs)** per mappare e analizzare le associazioni semantiche tra concetti e aree geografiche nelle risposte dei modelli di linguaggio di grandi dimensioni (LLM). L'integrazione dei KGs consente di superare i metodi tradizionali, che si basano principalmente sull'analisi testuale e statistica, permettendo una rappresentazione più strutturata e ricca delle relazioni tra concetti.

I KGs permettono di esplicitare non solo le associazioni dirette, ma anche quelle implicite, fornendo una panoramica più profonda del bias nei modelli. Inoltre, il progetto prevede l'analisi del bias geografico in un contesto multilingua, utilizzando **KGs separati per ogni lingua** e tecniche di **graph alignment** per confrontare le differenze nei pattern di bias tra lingue diverse. Questo rappresenta un contributo innovativo, poiché pochi studi hanno esplorato in modo sistematico le variazioni di bias nei modelli multilingua.

Infine, la ricerca propone l'uso di **prompt mitigatori**, una tecnica pratica ed efficace per ridurre il bias nelle risposte dei LLM. Il prompting permette di intervenire senza la necessità di riaddestrare i modelli, fornendo una soluzione flessibile e applicabile in vari contesti. L'efficacia di questa tecnica verrà valutata tramite l'analisi delle risposte generate dai modelli, confrontando i risultati prima e dopo l'applicazione dei prompt mitigatori.

### Limiti del Progetto

Nonostante l'innovazione introdotta dall'uso dei KGs e del prompting mitigatore, questa ricerca presenta alcuni limiti che devono essere considerati.

In primo luogo, la costruzione e l'accuratezza dei **Knowledge Graphs** dipendono fortemente dalla qualità e dalla quantità dei **dataset di addestramento** utilizzati dai LLM. Se i dataset contengono distorsioni di base, queste possono essere riflesse nei KGs, limitando la capacità di eliminare completamente il bias.

In secondo luogo, la **complessità tecnica** nell'implementazione e nella manutenzione dei KGs, specialmente quando si integrano dati da fonti esterne come **Wikidata** o **DBpedia**, potrebbe rappresentare una sfida. Le incongruenze o le lacune nei dati potrebbero influenzare l'affidabilità delle analisi e rallentare il processo di costruzione del grafo.

Infine, le tecniche di **prompting mitigatore**, sebbene promettenti, potrebbero avere **efficacia variabile** a seconda del contesto. I prompt mitigatori potrebbero funzionare bene in scenari specifici, ma non essere facilmente generalizzabili per tutti i tipi di bias o per tutti i contesti geografici e culturali. Pertanto, sarà importante considerare l'efficacia di tali tecniche su un ampio spettro di casi d'uso.

In conclusione, pur riconoscendo questi limiti, la ricerca proposta ha il potenziale di apportare un contributo significativo alla comprensione e alla riduzione del bias geografico nei modelli di linguaggio di grandi dimensioni, utilizzando un approccio innovativo e multifattoriale basato su KGs e prompting mitigatore.

**GradBias: Unveiling Word Influence on Bias in**

**Text-to-Image Generative Models**

29 Agosto 2024

<https://arxiv.org/pdf/2408.16700>

**COME MIGLIORARE LO STATO DELL’ARTE**

### ****Approccio Quantitativo Avanzato: Misurazione Multidimensionale del Bias in Modelli T2I****

* **Titolo**: A Multidimensional Analysis of Bias Across Semantic, Visual, and Contextual Layers in Text-to-Image Generative Models
* **Obiettivo**: Estendere l’analisi di OpenBias e GradBias sviluppando un approccio multidimensionale che includa non solo il bias legato al contenuto visivo generato, ma anche le associazioni semantiche e contestuali nelle descrizioni testuali.
* **Metodologia**:
  + **Stratificazione del Bias**: Proporre un modello che separi i bias in tre dimensioni principali:
    1. **Semantica**: Bias nei concetti associati a parole chiave specifiche (es. "CEO" = uomo).
    2. **Visuale**: Bias nelle caratteristiche visive generate (es. colore della pelle, oggetti).
    3. **Contestuale**: Bias amplificati o modificati dal contesto del prompt.
  + **Strumenti Quantitativi**: Usare tecniche di clustering per raggruppare descrizioni testuali con bias simili e applicare metriche di correlazione avanzate (es. coefficiente di correlazione di Kendall) per analizzare la convergenza o divergenza del bias lungo le tre dimensioni.
  + **Sperimentazione**: Testare su più dataset (es. COCO, Flickr30K) e modelli T2I diversi (Stable Diffusion, DALL-E) per verificare come i bias cambiano su più dimensioni e versioni dei modelli.
  + **Confronto con Human Benchmarking**: Integrare uno studio comparativo con giudizio umano per verificare la validità dei bias identificati, confrontando le percezioni umane e i risultati del modello.
* **Risultati Attesi**: Una mappa tridimensionale del bias che mostri come i modelli generativi rispondono a diverse tipologie di prompt e come i bias si amplificano o riducono in base al contesto semantico e visivo.

**Innovazione**: Questo approccio multidimensionale fornirà una comprensione molto più sofisticata e sfumata dei bias, andando oltre le metriche tradizionali, e potrebbe diventare un framework di riferimento per future ricerche.

### ****Approccio Multimodale: Rilevamento del Bias Multilingue nei Modelli T2I****

* **Titolo**: Multilingual Bias Detection in Text-to-Image Generative Models: A Cross-Language Study Using GradBias
* **Obiettivo**: Estendere l'analisi del bias ai modelli T2I che operano in contesti multilingue, utilizzando il framework di GradBias per analizzare come il bias si manifesta in diverse lingue.
* **Metodologia**:
  + **Analisi Cross-Language**: Applicare GradBias a modelli T2I multilingue per analizzare come le differenze linguistiche (es. inglese, spagnolo, cinese, arabo) influenzano il bias generato nelle immagini. Confrontare il comportamento dei modelli rispetto alla rappresentazione del genere, razza e status socioeconomico in descrizioni tradotte.
  + **Valutazione Multimodale**: Utilizzare una combinazione di tecniche di analisi visiva e semantica per esplorare come i modelli rappresentano concetti culturali e sociali specifici in ogni lingua.
  + **Misure Quantitative del Bias Linguistico**: Utilizzare metriche di distanza linguistica (es. distanza di Levenshtein) e distribuzione delle parole per quantificare le differenze di bias generate dai modelli nelle varie lingue.
* **Risultati Attesi**: Identificare come la lingua influenza il bias nei modelli T2I, evidenziando eventuali differenze tra contesti linguistici e culturali nella generazione di immagini.

**Innovazione**: Questo studio espande il framework di GradBias a un contesto multilingue, contribuendo a una maggiore comprensione dell’impatto culturale e linguistico sui modelli generativi.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**1. Approccio Quantitativo Avanzato: Misurazione Multidimensionale del Bias**

**Titolo:** *A Multidimensional Analysis of Bias Across Semantic, Visual, and Contextual Layers in Text-to-Image Generative Models*

**Obiettivo:** Migliorare l'analisi dei bias nei modelli generativi Testo-Immagine (T2I) sviluppando un framework multidimensionale che consideri non solo il bias visivo generato, ma anche le associazioni semantiche e contestuali all'interno delle descrizioni testuali. Questo approccio migliora la metodologia di GradBias e OpenBias andando oltre la semplice analisi dei contenuti visivi.

**Perché migliora lo stato dell'arte:**  
Attualmente, le analisi dei bias nei modelli T2I tendono a focalizzarsi principalmente su uno o due aspetti, come il contenuto visivo o le associazioni di parole. Un’analisi **multidimensionale** separata su **bias semantici, visivi e contestuali** offre una comprensione più sfumata e precisa delle dinamiche che influenzano i modelli generativi. Questa metodologia consente di isolare meglio le fonti di bias e permette un'analisi più granulare del comportamento del modello.

**Metodologia:**

1. **Stratificazione del bias:** Si propone di separare i bias in tre dimensioni principali:
   * **Semantica:** Analisi delle associazioni di parole e concetti chiave (es. "CEO" associato a un uomo).
   * **Visuale:** Analisi delle caratteristiche visive generate dal modello (es. colore della pelle, tratti somatici, oggetti presenti).
   * **Contestuale:** Bias introdotti o amplificati dal contesto del prompt (es. come la descrizione del contesto influisce su stereotipi specifici).
2. **Strumenti quantitativi:** Utilizzare tecniche di **clustering** per identificare le tendenze nei bias presenti in gruppi di descrizioni e immagini simili. Si possono usare metriche di correlazione come il **coefficiente di correlazione di Kendall** per verificare la convergenza o divergenza dei bias attraverso le tre dimensioni.
3. **Esperimenti:** Testare i modelli su dataset diversificati come **COCO** o **Flickr30K** e su più versioni di modelli T2I (come **Stable Diffusion** e **DALL-E**) per confrontare come i bias si manifestano in contesti diversi.
4. **Human Benchmarking:** Integrare l'analisi con il feedback umano, confrontando i bias percepiti dagli utenti con quelli rilevati dal modello, per convalidare i risultati.

**Esempio:**  
Nel caso di un prompt come "Un ingegnere al lavoro", il modello può generare un uomo di pelle chiara con strumenti tecnologici. La dimensione **semantica** indicherebbe che "ingegnere" è associato a stereotipi di genere, la dimensione **visuale** evidenzierebbe che il modello privilegia tratti fisici specifici, mentre la dimensione **contestuale** potrebbe rivelare che il prompt stesso rafforza un bias implicito.

**Risultati attesi:**  
Questo approccio multidimensionale fornirà una **mappa del bias su tre livelli** (semantico, visivo e contestuale), evidenziando come il bias si amplifica o si riduce in base al contesto dei prompt e alle caratteristiche visive generate. Questa analisi permetterà di migliorare la **comprensione dei modelli generativi** e fornire un quadro completo delle loro tendenze pregiudizievoli.

**Innovazione:**  
Questa strategia rappresenta un'innovazione importante rispetto alle tecniche esistenti, che solitamente non analizzano il bias a livello contestuale o non collegano in modo coerente i bias semantici e visivi. Inoltre, un modello di **misurazione tridimensionale** potrebbe diventare un framework di riferimento per futuri studi sui bias.

**2. Approccio Multimodale: Rilevamento del Bias Multilingue nei Modelli T2I**

**Titolo:** *Multilingual Bias Detection in Text-to-Image Generative Models: A Cross-Language Study Using GradBias*

**Obiettivo:** Estendere l'analisi dei bias visivi nei modelli T2I multilingua per comprendere come il bias si manifesta in contesti linguistici diversi. Questo approccio utilizza il framework di **GradBias** per valutare come i modelli rappresentano concetti sociali, culturali e visivi in lingue diverse.

**Perché migliora lo stato dell'arte:**  
La maggior parte delle ricerche attuali sui modelli T2I si concentra su un'unica lingua (principalmente l'inglese). Un’analisi **multilingue** del bias nei modelli T2I rappresenta un passo significativo avanti, poiché può rivelare come le differenze culturali e linguistiche influenzino la generazione di immagini. Questo approccio non solo quantifica il bias visivo e semantico attraverso le lingue, ma esamina anche come il modello gestisce concetti e stereotipi culturalmente specifici.

**Metodologia:**

1. **Analisi cross-lingua:** Utilizza **GradBias** su modelli T2I che supportano lingue diverse (inglese, spagnolo, cinese, arabo). Si confrontano le descrizioni testuali e le immagini generate in base a concetti come genere, razza e status socioeconomico.
2. **Valutazione multimodale:** Integra un'analisi **semantica** delle descrizioni testuali con un'analisi **visiva** delle immagini generate, valutando come il modello traduca concetti culturali in lingue diverse.
3. **Metriche quantitative del bias linguistico:** Utilizza metriche di distanza linguistica (come la **distanza di Levenshtein**) e metriche di distribuzione delle parole per quantificare le differenze di bias in ogni lingua. Si può anche confrontare il comportamento del modello attraverso diverse traduzioni dello stesso prompt.

**Esempio:**  
Per un prompt come "Un insegnante in una classe", il modello potrebbe generare un’immagine diversa a seconda della lingua: in inglese potrebbe raffigurare una donna, mentre in arabo o cinese potrebbe rappresentare un uomo. L'analisi cross-lingua mostrerà come i bias di genere e culturali emergano in modo diverso a seconda del contesto linguistico.

**Risultati attesi:**  
Si prevede che le differenze linguistiche influenzino la rappresentazione dei concetti visivi e semantici. Ad esempio, modelli addestrati su lingue meno rappresentate nei dati di addestramento (come l’arabo o il cinese) potrebbero esibire bias più accentuati rispetto a quelli addestrati principalmente sull'inglese. Questo studio fornirà una visione unica di come i modelli T2I generano immagini culturalmente specifiche.

**Innovazione:**  
Questo studio amplia l'ambito di **GradBias** a un contesto multilingue, contribuendo a una maggiore comprensione dell’impatto culturale e linguistico sui modelli generativi. È un'area che è stata finora poco esplorata e potrebbe fornire informazioni preziose su come adattare i modelli T2I per un pubblico globale, riducendo bias linguistici e culturali.

**Improving Large Language Model (LLM) fidelity through context-aware grounding: A systematic approach to reliability and veracity**

10 Novembre 2023

**COME MIGLIORARE LO STATO DELL’ARTE**

### ****Valutazione Dinamica del Contesto: Misurazione della Responsività dei Modelli al Cambiamento del Contesto****

* **Titolo**: Dynamic Context Responsiveness in LLMs: Quantifying Model Adaptability to Shifting Situational, Cultural, and Ethical Contexts
* **Obiettivo**: Misurare la capacità del modello di adattarsi a cambiamenti rapidi e imprevisti di contesto, garantendo risposte adeguate in tempo reale.
* **Esempio pratico**: Immagina un modello LLM utilizzato in un assistente di emergenza. Il contesto iniziale è una domanda generica su come spegnere un piccolo incendio ("Hai un estintore?"). Subito dopo, la situazione si aggrava e viene aggiornato il contesto ("C'è una persona intrappolata, cosa devo fare?"). Il modello dovrebbe rispondere rapidamente adattandosi al nuovo contesto critico con istruzioni appropriate ("Chiama immediatamente il 118 e cerca un'uscita di emergenza").
* **Metodologia**:
  + Creazione di un **dataset dinamico** che include contesti che cambiano rapidamente (emergenze, eventi critici) con scenari complessi e multi-turn (ad esempio, passando da un contesto etico a uno situazionale in una conversazione).
  + Utilizzare tecniche di **analisi del cambiamento** per misurare quanto velocemente il modello modifica le sue risposte in base al nuovo contesto.
* **Metriche**:
  + **Contextual Transition Time (CTT)**: Misura il tempo (in secondi o millisecondi) che il modello impiega per cambiare la sua risposta da un contesto iniziale a uno nuovo.
  + **Contextual Accuracy (CA)**: Percentuale di risposte che riflettono correttamente il nuovo contesto in seguito a un cambiamento.
  + **Precision e Recall per task complessi**: Valuta quanto il modello riesce a rispondere correttamente in contesti che mutano rapidamente.
* **Valutazione**:

Il **CTT** dovrebbe essere basso, indicando un'alta reattività del modello, mentre la **CA** dovrebbe essere alta, dimostrando che il modello risponde accuratamente nonostante il cambio di contesto.

### ****Analisi Quantitativa dell'Interazione Tra Contesti: Rilevazione di Conflitti Etico-Culturali nei Modelli****

* **Titolo**: Quantifying Intercontextual Conflicts in LLMs: A Framework for Identifying and Resolving Ethical-Cultural Tensions
* **Obiettivo**: Rilevare e misurare quantitativamente l'impatto di conflitti tra contesti etici e culturali nelle risposte generate dai modelli.
* **Esempio pratico**: Immagina che un modello LLM debba rispondere a una domanda che richiede sensibilità etica, come "È accettabile lavorare durante una festività religiosa?" in un contesto culturale islamico durante il Ramadan. Se il modello ha un contesto etico che promuove il rispetto della diversità religiosa e culturale, potrebbe suggerire di non lavorare. Tuttavia, se la stessa domanda viene posta in un contesto aziendale globale con priorità aziendali, potrebbe emergere un conflitto etico-culturale.
* **Metodologia**:
  + Costruzione di un dataset che include **scenari di conflitto** tra norme etiche e culturali (es. festività religiose, pratiche culturali vs esigenze aziendali).
  + Applicazione di metriche di coerenza e conflitto per valutare quanto il modello riesce a bilanciare i contesti in situazioni conflittuali.
* **Metriche**:
  + **Conflict Resolution Score (CRS)**: Misura quanto efficacemente il modello risolve i conflitti tra contesti. Valori più alti indicano che il modello è capace di risolvere il conflitto in modo appropriato.
  + **Ethical Coherence Index (ECI)**: Misura la coerenza etica del modello nelle risposte generate quando i contesti culturali ed etici sono in contrasto. Valori elevati suggeriscono che il modello mantiene stabilità etica nonostante i conflitti.
  + **Cultural Sensitivity Score (CSS)**: Valuta quanto le risposte del modello rispettano il contesto culturale in situazioni eticamente conflittuali.
* **Valutazione**:

Un alto **CRS** indica che il modello può risolvere correttamente i conflitti contestuali. Valori elevati di **ECI** e **CSS** dimostrano che il modello può bilanciare contesti etici e culturali, mantenendo risposte appropriate in ambienti multietnici.

### ****Benchmarking Quantitativo della Spiegabilità nelle Decisioni Contestuali: Valutazione delle Spiegazioni Etiche e Culturali****

* **Titolo**: Quantitative Benchmarking of Explainability in Contextual Decision-Making: Assessing the Ethical and Cultural Grounding of LLM Outputs
* **Obiettivo**: Valutare quantitativamente la qualità delle spiegazioni fornite dal modello in base ai contesti etici e culturali per migliorare la trasparenza e la fiducia nelle risposte del modello.
* **Esempio pratico**: Un modello risponde a una domanda eticamente delicata come "È giusto permettere ai bambini di lavorare?" in un contesto culturale specifico. Il modello dovrebbe fornire non solo una risposta, ma anche spiegare perché la risposta riflette una considerazione etica (es. diritti dei bambini) e culturale (es. leggi nazionali sul lavoro minorile).
* **Metodologia**:
  + Implementare un framework che genera e valuta spiegazioni dettagliate per le decisioni del modello nei contesti etici e culturali.
  + Misurare la coerenza delle spiegazioni con le aspettative culturali e normative, valutandole con un **dataset annotato** per la correttezza delle spiegazioni.
* **Metriche**:
  + **Fidelity Score (FS)**: Valuta quanto le spiegazioni del modello sono coerenti con le decisioni prese in base al contesto etico o culturale.
  + **Complexity-Transparency Tradeoff (CTT)**: Misura il compromesso tra la complessità delle spiegazioni e la loro comprensibilità. Valori ottimali indicano spiegazioni comprensibili senza sacrificare dettagli importanti.
  + **Human Explanation Consistency (HEC)**: Percentuale di concordanza tra spiegazioni generate dal modello e quelle fornite da esperti umani. Un valore alto suggerisce che le spiegazioni del modello sono in linea con il ragionamento umano.
* **Valutazione**:

Alti valori di **FS** e **HEC** indicano che il modello è capace di fornire spiegazioni coerenti e comprensibili. L'ottimizzazione del **CTT** migliora la trasparenza senza sacrificare la complessità necessaria.

### ****Miglioramento della Robustezzza Multilingue e Multiculturale dei Modelli di Grounding Contestuale****

* **Titolo**: Enhancing Multilingual and Multicultural Robustness in Context-Aware Grounding for LLMs: A Quantitative Evaluation Framework
* **Obiettivo**: Valutare la robustezza dei modelli multilingue nella gestione di contesti culturali ed etici in diverse lingue, migliorando la fedeltà del grounding contestuale.
* **Esempio pratico**: Immagina un modello che risponde a domande religiose in lingue diverse, come "Qual è il giorno di riposo settimanale?". In un contesto ebraico (in ebraico), il modello dovrebbe rispondere con lo Shabbat (sabato), mentre in un contesto islamico (in arabo) dovrebbe rispondere con il venerdì.
* **Metodologia**:
  + Creare un **benchmark multilingue e multiculturale** che includa domande etiche e culturali in lingue diverse (es. inglese, arabo, cinese, spagnolo).
  + Utilizzare tecniche di analisi della varianza (**ANOVA**) per misurare le differenze tra le prestazioni del modello in diverse lingue e culture.
* **Metriche**:
  + **Cross-Linguistic Contextual Fidelity (CLCF)**: Percentuale di coerenza nelle risposte contestuali tra lingue diverse. Un valore alto indica che il modello mantiene la stessa precisione contestuale in più lingue.
  + **Cultural Accuracy per Lingua (CAL)**: Misura della correttezza delle risposte rispetto al contesto culturale in ciascuna lingua.
  + **Language-Culture Consistency Score (LCCS)**: Valuta la coerenza delle risposte culturali nelle varie lingue analizzate.
* **Valutazione**:

Alti valori di **CLCF**, **CAL** e **LCCS** indicano che il modello riesce a mantenere la stessa robustezza contestuale in più lingue e culture.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**1. Valutazione Dinamica del Contesto: Misurazione della Responsività dei Modelli al Cambiamento del Contesto**

**Titolo:** *Dynamic Context Responsiveness in LLMs: Quantifying Model Adaptability to Shifting Situational, Cultural, and Ethical Contexts*

**Obiettivo:** Migliorare la capacità dei modelli LLM di adattarsi rapidamente e accuratamente ai cambiamenti di contesto, soprattutto in situazioni complesse e in tempo reale.

**Perché migliora lo stato dell'arte:**  
Attualmente, i modelli LLM spesso mantengono risposte statiche anche quando il contesto della conversazione cambia. L'introduzione di una **valutazione dinamica del contesto** permette di testare quanto un modello possa adattarsi a situazioni emergenti e garantire risposte più affidabili e reattive, riducendo la disconnessione tra il modello e il contesto in evoluzione.

**Metodologia:**

* **Dataset dinamico:** Creare scenari complessi e multi-turn con cambiamenti rapidi di contesto (ad es., emergenze o domande etiche che cambiano improvvisamente).
* **Analisi del cambiamento:** Misurare quanto velocemente il modello modifica le sue risposte in base al nuovo contesto, utilizzando tecniche come il **Contextual Transition Time (CTT)** e il **Contextual Accuracy (CA)**.

**Esempio pratico:**  
In un caso di assistenza in emergenza, il contesto potrebbe cambiare rapidamente da una situazione iniziale semplice ("Come spegnere un piccolo incendio") a una situazione più complessa e pericolosa ("Una persona è intrappolata, cosa fare?"). Il modello dovrebbe adattarsi prontamente fornendo risposte appropriate in entrambe le situazioni.

**Metriche:**

* **Contextual Transition Time (CTT):** Misura il tempo impiegato dal modello per adattare la sua risposta a un nuovo contesto.
* **Contextual Accuracy (CA):** Percentuale di risposte coerenti e corrette in seguito a cambiamenti contestuali.

**Risultati attesi:**  
Modelli che rispondono più velocemente e accuratamente ai cambiamenti di contesto, migliorando la loro **reattività** e affidabilità.

**2. Analisi Quantitativa dell'Interazione Tra Contesti: Rilevazione di Conflitti Etico-Culturali nei Modelli**

**Titolo:** *Quantifying Intercontextual Conflicts in LLMs: A Framework for Identifying and Resolving Ethical-Cultural Tensions*

**Obiettivo:** Rilevare conflitti tra contesti etici e culturali nelle risposte dei modelli, migliorando la capacità del modello di risolvere situazioni complesse dove norme etiche e culturali possono entrare in conflitto.

**Perché migliora lo stato dell'arte:**  
I modelli attuali non hanno strumenti adeguati per bilanciare e risolvere i conflitti tra diversi contesti etici e culturali. Questo framework propone un approccio quantitativo per rilevare e risolvere tali conflitti, migliorando la coerenza etica e culturale delle risposte.

**Metodologia:**

* **Dataset di conflitti:** Creare un set di dati con scenari che rappresentano conflitti tra valori etici e culturali (es. festività religiose contro richieste aziendali).
* **Metriche di risoluzione del conflitto:** Misurare quanto il modello riesce a bilanciare e risolvere i conflitti tramite metriche come il **Conflict Resolution Score (CRS)** e l'**Ethical Coherence Index (ECI)**.

**Esempio pratico:**  
Un modello potrebbe dover rispondere a una domanda che riguarda il lavoro durante una festività religiosa in un contesto aziendale. Il modello deve bilanciare il rispetto per la diversità culturale con le priorità aziendali, evitando di generare risposte che risultino culturalmente offensive.

**Metriche:**

* **Conflict Resolution Score (CRS):** Misura l'efficacia con cui il modello risolve conflitti contestuali.
* **Ethical Coherence Index (ECI):** Valuta la coerenza etica delle risposte in situazioni di conflitto.
* **Cultural Sensitivity Score (CSS):** Misura quanto le risposte del modello rispettano il contesto culturale.

**Risultati attesi:**  
Migliore risoluzione dei conflitti tra contesti culturali ed etici, con modelli capaci di generare risposte appropriate e culturalmente sensibili.

**3. Benchmarking Quantitativo della Spiegabilità nelle Decisioni Contestuali: Valutazione delle Spiegazioni Etiche e Culturali**

**Titolo:** *Quantitative Benchmarking of Explainability in Contextual Decision-Making: Assessing the Ethical and Cultural Grounding of LLM Outputs*

**Obiettivo:** Migliorare la spiegabilità dei modelli LLM, fornendo risposte non solo corrette, ma anche accompagnate da spiegazioni eticamente e culturalmente fondate.

**Perché migliora lo stato dell'arte:**  
L'aggiunta di spiegazioni dettagliate permette una maggiore fiducia nei modelli LLM, rendendo più chiaro come e perché il modello prende determinate decisioni. Questo è particolarmente importante in contesti etici e culturali dove una semplice risposta potrebbe non essere sufficiente.

**Metodologia:**

* **Framework di spiegabilità:** Implementare un sistema che fornisca spiegazioni dettagliate delle risposte generate dal modello nei contesti etici e culturali.
* **Misura della coerenza etica e culturale:** Valutare la coerenza delle spiegazioni con normative culturali o etiche.

**Esempio pratico:**  
In un contesto etico delicato, come il lavoro minorile, il modello non solo deve rispondere correttamente, ma deve anche spiegare il motivo per cui il lavoro minorile è scorretto in base a normative etiche e culturali, migliorando così la trasparenza e la comprensione delle decisioni.

**Metriche:**

* **Fidelity Score (FS):** Misura la coerenza tra spiegazione e decisione.
* **Complexity-Transparency Tradeoff (CTT):** Valuta l'equilibrio tra complessità e comprensibilità delle spiegazioni.
* **Human Explanation Consistency (HEC):** Percentuale di accordo tra spiegazioni umane e quelle fornite dal modello.

**Risultati attesi:**  
Modelli che forniscono spiegazioni trasparenti e coerenti, permettendo una maggiore fiducia nelle loro risposte.

**4. Miglioramento della Robustezza Multilingue e Multiculturale nei Modelli di Grounding Contestuale**

**Titolo:** *Enhancing Multilingual and Multicultural Robustness in Context-Aware Grounding for LLMs: A Quantitative Evaluation Framework*

**Obiettivo:** Migliorare la robustezza dei modelli multilingue nella gestione dei contesti etici e culturali in lingue diverse, garantendo che le risposte siano coerenti e appropriate in contesti globali.

**Perché migliora lo stato dell'arte:**  
I modelli attuali spesso soffrono di variazioni nelle loro prestazioni quando operano in lingue diverse o in contesti culturali complessi. Questo framework mira a garantire che i modelli multilingue siano in grado di fornire risposte culturalmente e linguisticamente appropriate, indipendentemente dalla lingua.

**Metodologia:**

* **Benchmark multilingue:** Creare un set di domande etiche e culturali tradotte in più lingue (es. inglese, arabo, cinese) per testare la consistenza delle risposte del modello.
* **Analisi della varianza:** Misurare le differenze nelle risposte del modello tra le lingue utilizzando tecniche di analisi della varianza (ANOVA).

**Esempio pratico:**  
Un modello che risponde alla domanda "Qual è il giorno di riposo settimanale?" dovrebbe essere capace di fornire risposte culturalmente appropriate in diverse lingue, riflettendo ad esempio lo Shabbat in ebraico e il venerdì in arabo.

**Metriche:**

* **Cross-Linguistic Contextual Fidelity (CLCF):** Percentuale di coerenza contestuale tra lingue diverse.
* **Cultural Accuracy per Lingua (CAL):** Misura la correttezza culturale delle risposte in ciascuna lingua.
* **Language-Culture Consistency Score (LCCS):** Valuta la consistenza delle risposte culturali tra le lingue.

**Risultati attesi:**  
Migliore consistenza e robustezza nelle risposte contestuali in più lingue e culture, con modelli che offrono risposte culturalmente appropriate e sensibili, indipendentemente dalla lingua di input.

**Multimodal Misinformation Detection using Large Vision-Language Models**

19 Luglio 2024

<https://arxiv.org/pdf/2407.14321>

**COME MIGLIORARE LO STATO DELL’ARTE**

### ****Rilevamento e Mitigazione del Bias di Genere nella Disinformazione Multimodale****

* **Titolo**: Mitigating Gender Bias in Multimodal Misinformation Detection Systems: A Quantitative Approach
* **Obiettivo**: Misurare e ridurre il bias di genere nei modelli di rilevamento della disinformazione multimodale, concentrandosi su come le informazioni testuali e visive influenzino le predizioni quando le affermazioni riguardano uomini o donne.
* **Esempio pratico**: Un sistema di rilevamento della disinformazione potrebbe essere soggetto a bias quando interpreta notizie riguardanti donne in ambiti stereotipicamente maschili (ad es. "La prima donna pilota a vincere un Grand Prix"). Le immagini di donne in ambiti non stereotipati potrebbero essere trattate in modo diverso rispetto a quelle di uomini, con una tendenza a classificarle come "false" o "non sufficienti" a causa di bias culturali o di genere.
* **Metodologia**:
  + Addestrare il modello su **dataset bilanciati per genere** (ad es. **Factify**) con annotazioni esplicite su bias di genere.
  + Implementare un approccio di **re-ranking** che corregga il bias nel recupero delle immagini e dei testi, ponendo maggiore enfasi sull’equità di genere.
* **Metriche**:
  + **Gender Fairness Score (GFS)**: Misura la capacità del modello di trattare equamente affermazioni e prove riguardanti uomini e donne.
  + **Misclassification Rate (MR) per genere**: Percentuale di errori di classificazione suddivisa per genere, valutando se ci sono disparità.
  + **Sycophancy Bias Score (SBS)**: Percentuale di risposte conformiste verso stereotipi di genere nei contenuti testuali o visivi.
* **Valutazione**:

Un basso **MR per genere** e un **SBS** ridotto indicano che il modello è meno incline a confermare bias di genere nelle affermazioni multimodali. Miglioramenti chiave includerebbero un **GFS** più alto e una riduzione dei falsi positivi/negativi legati a genere.

### ****Rilevamento di Sycophancy Visiva nel Riconoscimento di Fake News****

* **Titolo**: Quantifying Visual Sycophancy in Multimodal Fake News Detection: A Controlled Analysis
* **Obiettivo**: Valutare quantitativamente quanto i modelli multimodali siano influenzati da sycophancy visiva, ovvero la tendenza a conformarsi alle aspettative implicite dell'utente riguardo immagini e testi associati alla disinformazione.
* **Esempio pratico**: Immagina un modello che deve verificare l'affermazione "Un famoso scienziato ha smentito il cambiamento climatico" con un'immagine di una persona che non è uno scienziato ma ha una posa professionale. Il modello, influenzato dalla **sycophancy visiva**, potrebbe falsamente confermare la veridicità dell’affermazione basandosi sull'apparenza visiva.
* **Metodologia**:
  + Creazione di **dataset controllati** con affermazioni potenzialmente false e immagini associate che variano da altamente congruenti (rafforzano il bias) a altamente incongruenti.
  + Misurazione del grado di **conformità visiva** del modello nel confermare o smentire affermazioni basate su prove visive fuorvianti.
* **Metriche**:
  + **Visual Conformity Score (VCS)**: Percentuale di affermazioni confermate o smentite dal modello in conformità con l'apparenza visiva, indipendentemente dal contenuto testuale.
  + **Visual Sycophancy Bias (VSB)**: Misura della tendenza del modello a confermare affermazioni basate sulla sola apparenza visiva (senza considerare il contenuto testuale).
  + **False Conformity Rate (FCR)**: Percentuale di falsi positivi generati dalla conformità a immagini non correlate o fuorvianti.
* **Valutazione**:

Un basso **VCS** e **VSB** indicano che il modello non è influenzato negativamente dall'apparenza visiva. Un **FCR** ridotto segnala che il modello riesce a distinguere tra prove visive reali e quelle fuorvianti, senza conformarsi al bias visivo.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**1. Rilevamento e Mitigazione del Bias di Genere nella Disinformazione Multimodale**

**Titolo:** *Mitigating Gender Bias in Multimodal Misinformation Detection Systems: A Quantitative Approach*

**Obiettivo:**  
Misurare e ridurre il bias di genere nei sistemi di rilevamento della disinformazione multimodale, migliorando l'equità di genere nelle predizioni che coinvolgono sia informazioni testuali che visive.

**Perché migliora lo stato dell’arte:**  
Attualmente, molti modelli multimodali soffrono di bias di genere, soprattutto in contesti in cui stereotipi di genere sono presenti. Ridurre questo bias è cruciale per garantire che le predizioni siano basate su dati oggettivi piuttosto che su pregiudizi culturali o stereotipati. L'integrazione di metriche specifiche di equità di genere e il riequilibrio dei dataset migliorerebbero l'affidabilità e l'equità del modello.

**Metodologia:**

* **Dataset bilanciati per genere:** Utilizzare dataset annotati che bilanciano accuratamente i contenuti per uomini e donne, come *Factify*, per l'addestramento del modello.
* **Re-ranking per l’equità di genere:** Implementare un sistema di re-ranking che ribilancia l'importanza di informazioni visive e testuali, correggendo il bias durante il recupero di immagini e testi legati a affermazioni multimodali.

**Esempio pratico:**  
Un sistema potrebbe avere difficoltà a valutare correttamente la disinformazione su una donna che compie un’impresa stereotipicamente maschile (ad esempio, una donna che vince una gara automobilistica). La mitigazione del bias di genere aiuterebbe il modello a considerare le prove senza pregiudizi legati al genere.

**Metriche:**

* **Gender Fairness Score (GFS):** Misura la capacità del modello di trattare equamente uomini e donne nelle predizioni.
* **Misclassification Rate (MR) per genere:** Percentuale di errori nelle predizioni suddivisa per genere, utile per rilevare disparità.
* **Sycophancy Bias Score (SBS):** Valuta la tendenza del modello a conformarsi a stereotipi di genere nel processo di rilevamento della disinformazione.

**Risultati attesi:**  
Un **Gender Fairness Score (GFS)** più alto e una **riduzione del MR per genere** e del **Sycophancy Bias Score (SBS)** dimostrerebbero un miglioramento nell’equità di genere, con un modello meno incline a errori dovuti a pregiudizi di genere nei contenuti multimodali.

**2. Rilevamento di Sycophancy Visiva nel Riconoscimento di Fake News**

**Titolo:** *Quantifying Visual Sycophancy in Multimodal Fake News Detection: A Controlled Analysis*

**Obiettivo:**  
Valutare quantitativamente quanto i modelli multimodali siano influenzati da sycophancy visiva, ovvero la tendenza a confermare affermazioni basate sull'apparenza visiva, indipendentemente dalla coerenza testuale.

**Perché migliora lo stato dell'arte:**  
I modelli multimodali possono essere facilmente influenzati da immagini visivamente convincenti, anche se il contenuto visivo non supporta la veridicità del testo. Ridurre la sycophancy visiva migliora significativamente la capacità dei modelli di riconoscere la disinformazione, garantendo che l'apparenza visiva non sovrasti il contenuto testuale nelle decisioni del modello.

**Metodologia:**

* **Dataset controllati:** Creare un dataset di notizie false o fuorvianti con immagini che variano da altamente congruenti a incongruenti rispetto ai testi associati.
* **Misurazione della conformità visiva:** Testare il modello per vedere quanto conferma o smentisce affermazioni basate sull'apparenza visiva, indipendentemente dal testo.

**Esempio pratico:**  
In un test, il modello potrebbe ricevere una notizia con l'immagine di una persona con un aspetto professionale, ma che non è uno scienziato, accompagnata da un testo che afferma che questa persona ha fatto una dichiarazione scientifica controversa. La sycophancy visiva farebbe sì che il modello confermi erroneamente l'affermazione a causa dell'apparenza visiva.

**Metriche:**

* **Visual Conformity Score (VCS):** Percentuale di affermazioni confermate o smentite dal modello in base alla conformità visiva, senza considerare il contenuto testuale.
* **Visual Sycophancy Bias (VSB):** Misura la tendenza del modello a conformarsi all’apparenza visiva.
* **False Conformity Rate (FCR):** Percentuale di falsi positivi generati dalla conformità a immagini non correlate o fuorvianti.

**Risultati attesi:**  
Una riduzione del **Visual Conformity Score (VCS)**, del **Visual Sycophancy Bias (VSB)** e del **False Conformity Rate (FCR)** dimostrerebbe che il modello è meno influenzato da prove visive fuorvianti e che riesce a distinguere tra informazioni visive e testuali in modo efficace.

**Evaluating Fairness in Large Vision-Language Models Across Diverse**

**Demographic Attributes and Prompts**

25 Luglio 2024

<https://arxiv.org/pdf/2406.17974>

**COME MIGLIORARE LO STATO DELL’ARTE**

### ****Mitigazione del Bias attraverso Prompt-Controlled Augmentation****

* **Titolo**: Reducing Demographic Bias in LVLMs through Controlled Prompt Augmentation
* **Obiettivo**: Utilizzare **prompt augmentation** per ridurre il bias demografico nei LVLMs, creando prompt che correggano automaticamente l'output dei modelli quando vengono rilevati risultati potenzialmente distorti.
* **Esempio pratico**: Se un modello genera una risposta che associa un gruppo demografico a uno stereotipo, il sistema di prompt augmentation interverrà per modificare il contesto o la domanda, inducendo il modello a fornire una risposta più equa e accurata.
* **Metodologia**:
  + Sviluppo di **prompt augmentation** per modificare dinamicamente i prompt iniziali con parole chiave neutre o correttive per contrastare il bias.
  + Utilizzo di tecniche di **reinforcement learning** per allenare il modello a correggere automaticamente i suoi output quando rileva risposte non bilanciate o distorte.
  + Misurazione del miglioramento dell'equità attraverso prompt modificati rispetto a prompt standard.
* **Metriche**:
  + **Prompt Sensitivity Index (PSI)**: Misura della sensibilità del modello ai cambiamenti nei prompt e del conseguente impatto sull'equità delle risposte.
  + **Bias Mitigation Score (BMS)**: Percentuale di riduzione del bias nelle risposte generate dopo l'intervento del prompt augmentation.
  + **Equity Improvement Ratio (EIR)**: Misura dell'aumento di equità nel trattamento dei diversi gruppi demografici tra prompt originali e augmented.
* **Valutazione**:

Un elevato **BMS** e **EIR** indicheranno che il sistema di prompt augmentation è efficace nel migliorare la correttezza e l'equità delle risposte senza compromettere l'accuratezza complessiva.

### ****Sycophancy Detection nei LVLM attraverso l'Analisi di Prompt Dinamici****

* **Titolo**: Detecting and Mitigating Sycophancy in Vision-Language Models through Dynamic Prompt Evaluation
* **Obiettivo**: Esplorare la presenza della **sycophancy** (adattamento a risposte che l'utente potrebbe preferire) nei modelli LVLM attraverso l'uso di prompt dinamici e variabili. L'obiettivo è ridurre la tendenza dei modelli a conformarsi a pregiudizi impliciti negli input utente.
* **Esempio pratico**: Un LVLM che riceve la domanda "Questa persona sembra pericolosa?" può essere influenzato dall'aspetto fisico della persona (ad esempio, il tono della pelle), producendo risposte conformi a stereotipi razziali.
* **Metodologia**:
  + Creazione di un **framework di prompt dinamici** che alterano leggermente il contesto per vedere se il modello risponde in modo coerente o se è influenzato dal pregiudizio dell'utente implicito nella domanda.
  + Monitoraggio di come il modello risponde a prompt più neutri rispetto a quelli carichi di bias per identificare schemi di sycophancy.
* **Metriche**:
  + **Sycophancy Bias Score (SBS)**: Misura della tendenza del modello a conformarsi a risposte implicitamente preferite basate sugli stereotipi nel prompt.
  + **Consistency Rate (CR)**: Percentuale di risposte coerenti del modello tra versioni neutre e cariche di bias dello stesso prompt.
  + **Deviation from Neutrality (DN)**: Quantifica quanto il modello si discosta dalle risposte neutre quando si trova di fronte a prompt che presentano bias impliciti.
* **Valutazione**:
  + Un basso **SBS** e un alto **CR** indicheranno che il modello è meno influenzato dai bias dei prompt, dimostrando resistenza alla sycophancy.

### ****Mitigazione del Bias Demografico nei LVLM attraverso Fine-Tuning Differenziale****

* **Titolo**: Fairness-Oriented Fine-Tuning of Large Vision-Language Models to Reduce Demographic Bias
* **Obiettivo**: Ridurre i bias demografici (genere, età, tono della pelle) nei LVLMs attraverso tecniche di fine-tuning differenziale che enfatizzino l'equità dei risultati in diversi gruppi demografici.
* **Esempio pratico**: Se un modello LVLM tende a classificare sistematicamente le donne come infermiere e gli uomini come medici, questo indica la presenza di un bias di genere. Il fine-tuning mirato può riequilibrare la risposta del modello senza danneggiare la sua accuratezza generale.
* **Metodologia**:
  + Utilizzo di **dataset bilanciati demograficamente** come **FairFace** e **FACET**, assicurando che tutte le categorie di genere, età e tono della pelle siano equamente rappresentate.
  + Applicazione di **techniques di fairness-constrained optimization**, in cui il modello è penalizzato ogni volta che mostra risultati distorti per un determinato attributo demografico.
* **Metriche**:
  + **Fairness-Accuracy Tradeoff (FAT)**: Misura dell'equilibrio tra l'equità (minor bias) e l'accuratezza complessiva del modello.
  + **Demographic Parity Gap (DPG)**: Misura della disparità di trattamento tra gruppi demografici, ad esempio il divario tra recall di genere maschile e femminile.
  + **Fairness Recall (FR)**: Percentuale di esempi correttamente classificati da tutti i gruppi demografici, bilanciando la precisione tra i gruppi.
* **Valutazione**:
  + Un **DPG** ridotto e un **FR** elevato indicheranno che il fine-tuning è riuscito a ridurre il bias senza compromettere l'accuratezza complessiva del modello.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**1. Mitigazione del Bias attraverso Prompt-Controlled Augmentation**

**Titolo:** *Reducing Demographic Bias in LVLMs through Controlled Prompt Augmentation*

**Obiettivo:**  
Ridurre il bias demografico nei modelli LVLM utilizzando tecniche di *prompt augmentation*, in cui i prompt vengono dinamicamente modificati per correggere risposte distorte generate dai modelli. Questo approccio punta a garantire risposte più equilibrate e accurate, intervenendo in tempo reale quando si rilevano distorsioni nei risultati.

**Perché migliora lo stato dell’arte:**  
L’attuale ricerca sui modelli LVLM si concentra principalmente sulla rilevazione del bias, ma manca di meccanismi dinamici per correggerlo durante l'inferenza. L'uso di *prompt augmentation* rappresenta un passo in avanti significativo, poiché non solo rileva, ma interviene automaticamente, migliorando l'equità delle risposte generate.

**Metodologia:**

* **Prompt Augmentation Dinamico:** Modificare i prompt iniziali con parole chiave neutre o correttive che contrastano il bias in tempo reale.
* **Reinforcement Learning:** Addestrare i modelli con tecniche di reinforcement learning per riconoscere e correggere autonomamente risposte distorte.
* **Confronto tra Prompt Originali e Modificati:** Misurare l'equità nelle risposte prodotte utilizzando i prompt augmentati rispetto ai prompt standard.

**Esempio pratico:**  
Se un LVLM associa sistematicamente un gruppo demografico (come un'etnia) a professioni specifiche (come la criminalità), il prompt augmentation interverrà modificando il contesto per indurre una risposta più equa. Ad esempio, se il modello riceve il prompt "Questa persona è un criminale?", l'augmentation potrebbe cambiare la domanda in "Questa persona è un cittadino onesto?".

**Metriche:**

* **Prompt Sensitivity Index (PSI):** Misura la sensibilità del modello ai cambiamenti nei prompt e l’impatto sull'equità delle risposte.
* **Bias Mitigation Score (BMS):** Percentuale di riduzione del bias nelle risposte dopo il prompting augmentato.
* **Equity Improvement Ratio (EIR):** Misura il miglioramento dell’equità nel trattamento dei gruppi demografici dopo l’implementazione di prompt augmentati.

**Valutazione:**  
Valori elevati di **BMS** e **EIR** indicheranno che il sistema di *prompt augmentation* è efficace nel ridurre il bias demografico senza compromettere l'accuratezza generale delle risposte del modello.

**2. Rilevamento della Sycophancy nei LVLM attraverso l'Analisi di Prompt Dinamici**

**Titolo:** *Detecting and Mitigating Sycophancy in Vision-Language Models through Dynamic Prompt Evaluation*

**Obiettivo:**  
Rilevare e mitigare la *sycophancy* (l'inclinazione del modello a conformarsi alle preferenze implicite dell'utente) nei modelli LVLM, utilizzando un sistema di prompt dinamici che variino leggermente il contesto delle domande per identificare se il modello risponde coerentemente o in modo conforme a pregiudizi impliciti.

**Perché migliora lo stato dell’arte:**  
Molti modelli LVLM mostrano una tendenza a conformarsi ai pregiudizi impliciti degli utenti. L’introduzione di un framework dinamico per il rilevamento della sycophancy permette di identificare e mitigare questa tendenza, rendendo i modelli più neutrali e meno suscettibili a risposte basate su stereotipi o pregiudizi.

**Metodologia:**

* **Framework di Prompt Dinamici:** Modificare il contesto del prompt per verificare se il modello risponde in modo coerente o è influenzato dal bias implicito dell'utente.
* **Monitoraggio della Risposta del Modello:** Confrontare le risposte a prompt neutri con quelle di prompt carichi di bias per valutare la coerenza e l'influenza del bias.

**Esempio pratico:**  
Un modello potrebbe ricevere la domanda "Questa persona sembra pericolosa?" con l’immagine di una persona di colore. Se il modello risponde conformandosi a stereotipi razziali, la sycophancy visiva e testuale potrebbe essere evidente. Il sistema di prompt dinamici modificherebbe il contesto per testare la coerenza delle risposte e ridurre il bias.

**Metriche:**

* **Sycophancy Bias Score (SBS):** Misura la tendenza del modello a conformarsi a risposte implicite basate sui pregiudizi nel prompt.
* **Consistency Rate (CR):** Percentuale di risposte coerenti tra versioni neutre e biasate dello stesso prompt.
* **Deviation from Neutrality (DN):** Quantifica quanto il modello si discosta dalle risposte neutre quando affronta prompt con bias impliciti.

**Valutazione:**  
Un **basso SBS** e un **alto CR** indicheranno che il modello è meno suscettibile alla sycophancy e risponde coerentemente senza conformarsi ai bias impliciti nel prompt.

**3. Mitigazione del Bias Demografico nei LVLM attraverso Fine-Tuning Differenziale**

**Titolo:** *Fairness-Oriented Fine-Tuning of Large Vision-Language Models to Reduce Demographic Bias*

**Obiettivo:**  
Ridurre i bias demografici nei modelli LVLM (es. bias di genere, età, razza) attraverso tecniche di *fine-tuning differenziale*, ottimizzando l’equità senza compromettere l'accuratezza generale delle risposte.

**Perché migliora lo stato dell’arte:**  
La mitigazione del bias attraverso il *fine-tuning* offre un approccio mirato per ridurre il pregiudizio nei modelli LVLM, evitando compromessi sull’accuratezza generale. Questo approccio consente di correggere direttamente i pregiudizi nei dataset utilizzati per addestrare il modello.

**Metodologia:**

* **Dataset Bilanciati:** Utilizzare dataset bilanciati per genere, età e tono della pelle (es. *FairFace* e *FACET*), garantendo una rappresentazione equa di diversi gruppi demografici.
* **Fairness-Constrained Optimization:** Penalizzare il modello quando produce risposte distorte, incentivandolo a trattare equamente tutti i gruppi demografici.

**Esempio pratico:**  
Un modello potrebbe classificare sistematicamente le donne come infermiere e gli uomini come medici. Utilizzando il fine-tuning differenziale, il modello potrebbe essere riequilibrato per non associare pregiudizi professionali legati al genere.

**Metriche:**

* **Fairness-Accuracy Tradeoff (FAT):** Misura l'equilibrio tra l’equità e l'accuratezza complessiva del modello.
* **Demographic Parity Gap (DPG):** Misura la disparità di trattamento tra i diversi gruppi demografici, ad esempio il divario di recall tra i generi.
* **Fairness Recall (FR):** Percentuale di esempi correttamente classificati da ciascun gruppo demografico.

**Valutazione:**  
Un **DPG** ridotto e un **FR** elevato indicheranno che il modello sta trattando equamente i vari gruppi demografici senza compromessi sull’accuratezza.

**How Culturally Aware are Vision-Language Models?**

24 Maggio 2024

<https://arxiv.org/pdf/2405.17475>

**COME MIGLIORARE LO STATO DELL’ARTE**

### ****Integrazione di Knowledge Graphs Multiculturali per l'Addestramento di VLM Sensibili al Contesto****

* **Titolo**: Leveraging Multicultural Knowledge Graphs for Contextually Aware Vision-Language Models: A Multimodal Approach to Reducing Cultural Bias
* **Obiettivo**: Integrare **Knowledge Graphs multiculturali** e **ontologie culturali** nei modelli di linguaggio-visione per migliorare la consapevolezza culturale. L’obiettivo è ridurre il bias culturale attraverso la connessione automatica delle descrizioni generate dai modelli con informazioni culturali precise e contestuali.
* **Metodologia**:
  1. **Creazione di Knowledge Graphs Multiculturali**: Costruire knowledge graphs contenenti dati su culture, simboli, storie e tradizioni, sfruttando database esistenti (es. Wikidata, DBpedia) integrati con dati etnografici e storici provenienti da fonti accademiche e musei digitali.
  2. **Addestramento con Allineamento Multimodale**: Collegare le rappresentazioni visive e testuali generate dai modelli a concetti nei knowledge graphs, utilizzando tecniche di **multi-hop reasoning** per creare descrizioni culturalmente accurate e basate su contesto.
  3. **Verifica Automatica del Contesto**: Sviluppare un sistema di **verifica automatica del contesto** che confronti le descrizioni generate dai modelli con i dati presenti nel knowledge graph, correggendo potenziali errori culturali o stereotipi.
* **Tecniche Avanzate**:
  1. **Graph Neural Networks (GNN)** per collegare entità culturali e simboli con concetti visivi.
  2. **Multimodal Transformer Networks** per allineare descrizioni visive e concetti presenti nel knowledge graph, migliorando la coerenza culturale delle descrizioni.
  3. **Adversarial Learning** per addestrare i modelli a distinguere tra descrizioni culturalmente accurate e inesatte.
* **Metriche**:
  1. **Contextual Cultural Awareness Score (CCAS)**: misura la capacità del modello di riconoscere riferimenti culturali specifici in un’immagine, assegnando un punteggio in base alla profondità del contesto culturale recuperato dal knowledge graph.
  2. **Knowledge-Graph Accuracy (KGA)**: una metrica che valuta quanto bene il modello riesca a connettere le descrizioni visive con le entità corrette nel knowledge graph.
* **Valutazione**:

Testare l'approccio su dataset culturali complessi (es. immagini di cerimonie, rituali, opere d'arte storiche) e confrontare le descrizioni pre e post integrazione con il knowledge graph per monitorare i miglioramenti nella sensibilità culturale e nella riduzione del bias.

### ****Mitigazione del Bias Culturale nei Modelli VLM con Apprendimento Continuale e Memorie Sinaptiche****

* **Titolo**: Cultural Bias Mitigation in Vision-Language Models via Continual Learning with Synaptic Memory Architectures
* **Obiettivo**: Sviluppare un sistema di **apprendimento continuo** che permette ai VLM di adattarsi a nuovi contesti culturali senza dimenticare quelli precedentemente acquisiti, sfruttando **memorie sinaptiche** che trattengono conoscenze fondamentali sulle culture apprese.
* **Metodologia**:
  1. **Apprendimento Continuale con Elastic Weight Consolidation (EWC)**: Integrare EWC o altre tecniche simili per mantenere la memoria delle informazioni culturali rilevanti mentre il modello viene addestrato su nuovi dataset culturali. EWC consente di consolidare le informazioni precedenti mantenendo flessibili solo i pesi critici.
  2. **Memorie Sinaptiche per la Retenzione di Informazioni Culturali**: Implementare una **architettura di memoria sinaptica** che permetta al modello di conservare conoscenze culturalmente importanti anche mentre apprende nuovi concetti visivi e linguistici.
  3. **Adattamento Iterativo a Nuove Culture**: Aggiungere continuamente nuovi dataset culturali, monitorando come il modello si adatta a ciascuno di essi senza sacrificare la sensibilità verso culture già apprese. Ciò garantisce che il modello sia in grado di generalizzare a culture diverse senza dover essere ri-addestrato completamente.
* **Tecniche Avanzate**:
  1. **EWC Multimodale**: implementare Elastic Weight Consolidation esteso alle rappresentazioni sia visive che linguistiche.
  2. **Memorie Episodiche nei Modelli VLM**: costruzione di una memoria a lungo termine che immagazzina informazioni culturali precedenti, utilizzabile per riferimenti successivi in compiti di captioning.
* **Metriche**:
  1. **Cultural Retention Index (CRI)**: misura la capacità del modello di trattenere conoscenze su culture precedentemente apprese anche dopo essere stato esposto a nuove culture.
  2. **Continual Cultural Accuracy (CCA)**: una metrica che monitora quanto accuratamente il modello possa descrivere immagini culturalmente complesse man mano che viene addestrato su nuovi dati.
* **Valutazione**:

Testare su un dataset che contiene immagini di culture diverse introdotte in modo incrementale. Valutare la capacità del modello di mantenere una descrizione accurata delle culture precedenti dopo l’introduzione di nuovi set di immagini e contesti culturali.

### ****Riduzione delle Allucinazioni Culturali attraverso il Filtro Semantico con Reinforcement Learning****

* **Titolo**: Reducing Cultural Hallucinations in Vision-Language Models through Semantic Filtering and Reinforcement Learning
* **Obiettivo**: Ridurre l’incidenza di allucinazioni culturali nei VLM utilizzando un **filtro semantico dinamico** basato su **reinforcement learning (RL)**, che penalizza il modello per descrizioni imprecise o inventate, incentivandolo a generare descrizioni culturalmente accurate.
* **Metodologia**:
  1. **Filtro Semantico per la Coerenza Culturale**: Implementare un filtro semantico che analizza le descrizioni generate dai modelli, confrontandole con un **database di conoscenze culturali**, identificando e penalizzando descrizioni che contengono informazioni inventate o errate.
  2. **Reinforcement Learning**: Addestrare il modello utilizzando un approccio di RL dove il modello riceve una ricompensa per ogni descrizione accurata e culturalmente coerente, e una penalità per ogni allucinazione culturale o descrizione generata che non è supportata da dati culturali validi.
  3. **Controllo del Contesto Culturale**: Integrare un meccanismo che rileva e adatta dinamicamente il contesto culturale, ad esempio tramite l'uso di prompt personalizzati che influenzano il modello a concentrarsi su particolari sfumature culturali dell'immagine.
* **Tecniche Avanzate**:
  1. **Policy Gradient Methods (es. REINFORCE)** per ottimizzare la generazione di descrizioni culturalmente accurate.
  2. **Semantic Matching con Large Knowledge Bases**: integrazione di basi di conoscenze semantiche che verificano la coerenza delle descrizioni con i concetti culturali.
* **Metriche**:
  1. **Hallucination Reduction Score (HRS)**: una metrica che misura la riduzione delle allucinazioni nelle descrizioni generate dal modello.
  2. **Semantic Coherence Score (SCS)**: misura quanto accuratamente le descrizioni generano informazioni allineate al contesto culturale dell'immagine.
* **Valutazione**:

Testare il sistema su immagini con significati culturali complessi, monitorando il miglioramento delle descrizioni generate dopo l'integrazione del filtro semantico e la fase di reinforcement learning. Valutare la riduzione delle allucinazioni e l’aumento della coerenza culturale delle descrizioni.

### ****Cultural Adversarial Training: Sviluppo di Modelli Resilienti a Stereotipi Culturali****

* **Titolo**: Cultural Adversarial Training for Vision-Language Models: Building Resilient Models Against Cultural Stereotypes
* **Obiettivo**: Sviluppare una tecnica di **Cultural Adversarial Training** (CAT) per rendere i modelli VLM più resistenti agli stereotipi culturali, utilizzando dati avversari che sfidano i modelli a identificare e correggere stereotipi culturali nelle descrizioni generate.
* **Metodologia**:
  1. **Generazione di Esempi Avversari Culturali**: Creare esempi avversari che presentano stereotipi o rappresentazioni errate di culture (es. immagini con descrizioni che rafforzano stereotipi) e addestrare i modelli a riconoscerli e correggerli.
  2. **Apprendimento Avversario Multimodale**: Sviluppare un sistema di **training avversario** che penalizza il modello quando genera descrizioni stereotipate o culturalmente imprecise, rinforzando descrizioni che rispettano la diversità culturale.
  3. **Adversarial Prompting**: Implementare tecniche di prompting avversario, dove il modello viene esposto a descrizioni che cercano di introdurre bias culturale, e il compito del modello è di rifiutare tali bias, generando invece descrizioni accurate e prive di pregiudizi.
* **Tecniche Avanzate**:
  1. **Generative Adversarial Networks (GAN)** per generare esempi avversari che sfidano il modello.
  2. **Contrastive Learning** per insegnare al modello a distinguere tra descrizioni accurate e stereotipate.
* **Metriche**:
  1. **Adversarial Bias Score (ABS)**: misura la capacità del modello di resistere a esempi avversari che cercano di indurre bias culturale nelle descrizioni.
  2. **Cultural Stereotype Resilience Index (CSRI)**: quantifica la resistenza del modello agli stereotipi culturali dopo il training avversario.
* **Valutazione**:

Testare il modello su dataset di immagini che contengono stereotipi culturali impliciti ed espliciti. Monitorare il miglioramento della capacità del modello di generare descrizioni prive di bias e stereotipi dopo il training avversario.

**PERCHE’ MIGLIORA LO STATO DELL’ARTE**

**1. Integrazione di Knowledge Graphs Multiculturali per l'Addestramento di VLM Sensibili al Contesto**

**Titolo:** *Leveraging Multicultural Knowledge Graphs for Contextually Aware Vision-Language Models: A Multimodal Approach to Reducing Cultural Bias*

**Obiettivo:**  
Ridurre il bias culturale nei modelli Vision-Language (VLM) integrando Knowledge Graphs (KGs) multiculturali che connettano automaticamente le descrizioni generate con informazioni culturali precise. Questa integrazione mira a migliorare la consapevolezza culturale del modello.

**Perché migliora lo stato dell’arte:**  
Attualmente, molti modelli VLM mancano di una comprensione contestuale delle culture, il che porta spesso a bias culturali nelle descrizioni generate. Integrare KGs multiculturali permette ai modelli di accedere a una vasta base di conoscenze etnografiche e storiche, migliorando notevolmente l'accuratezza e la coerenza culturale delle descrizioni.

**Metodologia:**

1. **Costruzione di Knowledge Graphs Multiculturali:** Sviluppare KGs contenenti dati su culture, simboli, storie e tradizioni, basandosi su fonti come Wikidata, DBpedia, e database accademici.
2. **Allineamento Multimodale:** Collegare le rappresentazioni visive e testuali generate dai modelli con concetti presenti nei KGs, utilizzando tecniche come *multi-hop reasoning* per generare descrizioni culturalmente accurate.
3. **Verifica Automatica del Contesto:** Confrontare automaticamente le descrizioni con i dati presenti nei KGs, correggendo eventuali errori culturali o stereotipi.

**Metriche:**

* **Contextual Cultural Awareness Score (CCAS):** Valuta quanto accuratamente il modello riconosce riferimenti culturali nelle immagini.
* **Knowledge-Graph Accuracy (KGA):** Misura quanto efficacemente il modello associa descrizioni visive alle entità corrette nel KG.

**Valutazione:**  
Testare il sistema su dataset contenenti immagini culturalmente ricche (es. cerimonie tradizionali, opere d'arte) e confrontare le descrizioni generate prima e dopo l'integrazione del KG per monitorare la riduzione del bias e il miglioramento della sensibilità culturale.

**2. Mitigazione del Bias Culturale nei Modelli VLM con Apprendimento Continuale e Memorie Sinaptiche**

**Titolo:** *Cultural Bias Mitigation in Vision-Language Models via Continual Learning with Synaptic Memory Architectures*

**Obiettivo:**  
Sviluppare un sistema di apprendimento continuo che permetta ai modelli VLM di adattarsi a nuovi contesti culturali senza dimenticare quelli precedentemente acquisiti. Utilizzare memorie sinaptiche per mantenere la conoscenza culturale appresa.

**Perché migliora lo stato dell’arte:**  
L'apprendimento continuo nei VLM spesso porta alla *catastrophic forgetting*, in cui le nuove conoscenze sovrascrivono quelle già apprese. L'integrazione di memorie sinaptiche permette di preservare conoscenze culturali pregresse, garantendo una sensibilità culturale costante.

**Metodologia:**

1. **Elastic Weight Consolidation (EWC):** Utilizzare EWC per mantenere i pesi critici associati alle conoscenze culturali già apprese durante l'addestramento su nuovi dataset.
2. **Memorie Sinaptiche:** Implementare una memoria a lungo termine per immagazzinare informazioni culturali importanti, richiamabili durante i compiti futuri.
3. **Adattamento Iterativo a Nuovi Contesti Culturali:** Addestrare il modello su nuovi dataset culturali in modo incrementale, monitorando come si adatta senza perdere sensibilità verso culture precedenti.

**Metriche:**

* **Cultural Retention Index (CRI):** Misura la capacità del modello di mantenere conoscenze culturali apprese precedentemente.
* **Continual Cultural Accuracy (CCA):** Valuta l'accuratezza del modello nel descrivere immagini culturalmente complesse durante l'addestramento su nuovi dati.

**Valutazione:**  
Testare il modello su un dataset che introduce immagini di culture diverse in modo incrementale e valutare come il modello mantiene le conoscenze culturali apprese mentre assimila nuovi contesti.

**3. Riduzione delle Allucinazioni Culturali attraverso il Filtro Semantico con Reinforcement Learning**

**Titolo:** *Reducing Cultural Hallucinations in Vision-Language Models through Semantic Filtering and Reinforcement Learning*

**Obiettivo:**  
Ridurre le allucinazioni culturali nei VLM utilizzando un filtro semantico dinamico e il reinforcement learning (RL), penalizzando le descrizioni culturalmente inesatte o inventate.

**Perché migliora lo stato dell’arte:**  
Attualmente, i modelli VLM tendono a generare allucinazioni culturali, creando descrizioni imprecise o stereotipate. Il reinforcement learning, combinato con un filtro semantico, permette di incentivare descrizioni accurate e penalizzare quelle errate, migliorando la coerenza culturale.

**Metodologia:**

1. **Filtro Semantico:** Implementare un sistema che confronta le descrizioni generate con un database di conoscenze culturali, penalizzando le descrizioni errate o allucinate.
2. **Reinforcement Learning:** Addestrare il modello con un approccio di RL, in cui riceve una ricompensa per ogni descrizione accurata e una penalità per ogni allucinazione culturale.
3. **Controllo del Contesto Culturale:** Integrare prompt personalizzati per adattare dinamicamente il modello al contesto culturale specifico.

**Metriche:**

* **Hallucination Reduction Score (HRS):** Misura la riduzione delle allucinazioni nelle descrizioni generate.
* **Semantic Coherence Score (SCS):** Valuta quanto accuratamente le descrizioni sono culturalmente coerenti.

**Valutazione:**  
Testare il sistema su immagini con significati culturali complessi e monitorare la riduzione delle allucinazioni dopo l’integrazione del filtro semantico e del RL.

**4. Cultural Adversarial Training: Sviluppo di Modelli Resilienti a Stereotipi Culturali**

**Titolo:** *Cultural Adversarial Training for Vision-Language Models: Building Resilient Models Against Cultural Stereotypes*

**Obiettivo:**  
Rendere i modelli VLM più resistenti agli stereotipi culturali utilizzando un training avversario che li addestra a identificare e correggere stereotipi nelle descrizioni generate.

**Perché migliora lo stato dell’arte:**  
Attualmente, i modelli VLM possono perpetuare stereotipi culturali nelle descrizioni. Il *Cultural Adversarial Training (CAT)* offre un approccio innovativo per addestrare i modelli a resistere agli stereotipi culturali, migliorando la loro sensibilità culturale.

**Metodologia:**

1. **Generazione di Esempi Avversari Culturali:** Creare esempi che presentano stereotipi culturali e addestrare i modelli a riconoscerli e correggerli.
2. **Apprendimento Avversario Multimodale:** Sviluppare un sistema che penalizza il modello quando genera descrizioni stereotipate e premia descrizioni accurate.
3. **Adversarial Prompting:** Creare prompt che inducono bias culturale e addestrare il modello a ignorare tali bias, generando descrizioni più neutrali.

**Metriche:**

* **Adversarial Bias Score (ABS):** Misura la resistenza del modello agli esempi avversari che cercano di indurre bias culturali.
* **Cultural Stereotype Resilience Index (CSRI):** Quantifica la resistenza del modello agli stereotipi culturali.

**Valutazione:**  
Testare il modello su dataset con stereotipi culturali e monitorare il miglioramento nella generazione di descrizioni prive di bias culturale dopo il training avversario.

**A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning**

26 Ottobre 2022

<https://arxiv.org/pdf/2203.11933>

**COME MIGLIORARE LO STATO DELL’ARTE**

**Proposte di miglioramento**

* **Introduzione dell'intersezionalità nel debiasing**
  + **Descrizione:** Espandere l'analisi del bias per considerare attributi intersezionali, come il genere, l'etnia e l'età insieme. Attualmente, molti metodi considerano un singolo attributo alla volta (es. genere o etnia), ma le persone subiscono discriminazioni multiple e intersezionali.
  + **Esempio:** Se un modello riduce il bias di genere, potrebbe continuare a essere influenzato dal bias etnico o di età. Un sistema intersezionale può assicurare che query come "una persona professionale" restituiscano risultati equilibrati non solo per genere, ma anche per etnia e altre caratteristiche combinate.
* **Utilizzo di dataset eticamente curati e sintetici**
  + **Descrizione:** Per affrontare i problemi etici legati alla privacy e alle etichette sensibili nei dataset facciali, esplorare l'uso di immagini sintetiche generate tramite tecniche come GANs (Generative Adversarial Networks), che possono essere create con bilanciamenti accurati e controllati di attributi demografici.
  + **Esempio:** Invece di usare dati reali (come UTKFace), puoi impiegare dataset sintetici creati appositamente per bilanciare generi, etnie e contesti, assicurando che il debiasing non sia influenzato da pregiudizi già presenti nei dati originali.
* **Sviluppo di metodi di debiasing contestuali**
  + **Descrizione:** Invece di applicare il debiasing in modo uniforme a tutto il modello, esplorare metodi che adattino il debiasing a seconda del contesto. Alcune query possono richiedere una maggiore neutralità rispetto agli attributi demografici, mentre altre potrebbero beneficiare di associazioni legittime (ad esempio, query storiche).
  + **Esempio:** Un modello potrebbe trattare diversamente la query "un ingegnere nel XX secolo" rispetto a "un ingegnere moderno". In questo modo, l'associazione storica tra il genere maschile e gli ingegneri nel XX secolo potrebbe essere preservata, mentre il debiasing viene applicato a contesti moderni.
* **Integrazione di bias positive e accurate associazioni storiche**
  + **Descrizione:** Invece di cercare di eliminare completamente le associazioni demografiche nei modelli VL, esplorare modi per integrare bias positivi e accurate rappresentazioni storiche. Alcune associazioni, come il contributo storico di donne o persone di colore in certi campi, possono essere amplificate per correggere squilibri storici nella rappresentazione.
  + **Esempio:** Se un modello associa la frase "un pioniere nell'informatica" solo a immagini di uomini bianchi, si potrebbero aggiungere rappresentazioni storiche di donne come Ada Lovelace o persone di colore come Alan Emtage, pionieri nell'informatica.
* **Incorporazione di dati demografici aggiornati**
  + **Descrizione:** Aumentare la frequenza di aggiornamento dei dataset utilizzati per l'addestramento dei modelli VL. I dati sociali e demografici cambiano nel tempo, quindi i modelli devono adattarsi per evitare che perpetuino stereotipi ormai datati.
  + **Esempio:** Ad esempio, query come "CEO di successo" dovrebbero includere immagini di donne e persone di etnie diverse, riflettendo il cambiamento nella diversità dei leader aziendali negli ultimi decenni.
* **Analisi dell'impatto del debiasing su task diversi**
  + **Descrizione:** Valutare l'efficacia del debiasing non solo nei task di associazione immagine-testo, ma anche in compiti di classificazione zero-shot, segmentazione semantica e rilevamento di oggetti. Questo garantisce che le strategie di debiasing siano generalizzabili a una vasta gamma di applicazioni multimodali.
  + **Esempio:** Ad esempio, testare il modello su un compito di segmentazione delle immagini per verificare che i miglioramenti nel debiasing non compromettano la capacità di identificare correttamente gli oggetti in immagini complesse.
* **Riduzione del bias tramite feedback umano**
  + **Descrizione:** Integrare meccanismi di feedback umano per affinare il debiasing. Gli annotatori potrebbero fornire indicazioni dirette su come ridurre i bias in contesti specifici e migliorare l'efficacia delle associazioni immagine-testo.
  + **Esempio:** Gli utenti possono valutare se le query testuali sensibili come "un criminale" restituiscono risultati equilibrati o se rimangono ancora stereotipi da correggere.
* **Debiasing tramite apprendimento a più fasi**
  + **Descrizione:** Esplorare l'uso di una pipeline di apprendimento a più fasi per applicare il debiasing in modo graduale e iterativo. In una prima fase, il modello potrebbe ridurre i bias più evidenti, mentre nelle fasi successive affina il debiasing su aspetti più sottili.
  + **Esempio:** Dopo una prima fase che riduce il bias di genere, una fase successiva potrebbe concentrarsi sull'eliminazione di bias legati all'etnia, mantenendo il bilanciamento fra prestazioni e accuratezza delle rappresentazioni.

**Ulteriori sviluppi metodologici**

* **Espansione delle metriche di bias**
  + **Descrizione:** Integrare nuove metriche che valutino non solo la riduzione del bias, ma anche l'equità delle rappresentazioni. Ad esempio, metriche che monitorano la distribuzione di attributi demografici nei risultati e garantiscano che le rappresentazioni siano bilanciate senza forzature eccessive.
  + **Esempio:** Oltre alla metrica **WEAT**, potresti sviluppare una metrica che misura quanto equamente il modello distribuisce le rappresentazioni in base a diversi attributi demografici nei task di recupero immagine-testo.
* **Applicazione di approcci di knowledge editing**
  + **Descrizione:** Invece di modificare l'intero modello per ridurre i bias, esplorare tecniche di knowledge editing che permettano interventi mirati su specifiche conoscenze stereotipate senza compromettere la conoscenza complessiva del modello.
  + **Esempio:** Utilizzando tecniche come il **Fine-Grained Knowledge Editing**, si potrebbe intervenire solo su associazioni problematiche, come il collegamento di "un criminale" a determinate etnie, senza alterare altre conoscenze del modello.

**FRAMEWORK PER MIGLIORARE LO STATO DELL’ARTE DELLA MAGGIOR PARTE DEI LAVORI DI DETECTION & MITIGATION DEL BIAS NELLE LLM**

### Framework di Rilevamento del Bias nelle **Large Language Models (LLM)** con Clusterizzazione Automatica e Reti Avversarie

Il framework che segue è stato progettato per rilevare, monitorare e mitigare ogni tipo di bias nelle **Large Language Models (LLM)**. L’obiettivo è creare un sistema che non solo identifichi bias impliciti o espliciti all’interno dei modelli linguistici, ma che sia anche in grado di **auto-migliorarsi** grazie all’integrazione di **reti avversarie** e tecniche di **clusterizzazione automatica**.

Ogni passaggio è pensato per migliorare progressivamente il rilevamento di bias, arricchendo la capacità del modello di comprendere le **relazioni sottili** tra i dati. Di seguito è riportata una descrizione dettagliata di ciascun step, con i **vantaggi** relativi e **esempi** pratici per costruire il framework.

### 1. **Raccolta e Incrocio di Dataset Testuali Diversificati**

**Descrizione**: Il primo passo consiste nella raccolta di **dataset testuali** da varie fonti, cercando di includere diverse tipologie di testi che possano riflettere differenze culturali, etniche, di genere e sociali. Incrociare più dataset permette di avere una panoramica più ampia delle possibili fonti di bias.

**Esempi**:

* **Wikipedia** per testi enciclopedici neutrali.
* **Reddit** o **Twitter** per contenuti generati dagli utenti, che riflettono opinioni e stereotipi presenti nella società.
* **IMDB** per recensioni su film, libri e serie, in modo da includere opinioni personali e narrative culturali.

**Vantaggi**:

* **Ampiezza di contesto**: Incrociando dataset differenti si riescono a rilevare bias legati a specifici contesti culturali o sociali.
* **Varietà di linguaggi**: Sfruttare diversi tipi di linguaggio (formale e informale) permette di catturare bias in contesti di comunicazione sia professionale che quotidiana.

### 2. **Rappresentazione Multivettoriale delle Frasi Generate dalle LLM**

**Descrizione**: Applicare le **LLM** (ad esempio, **GPT-4**, **LLAMA**) per generare rappresentazioni multivettoriali delle frasi testuali. Ogni frase può essere trasformata in un insieme di vettori che rappresentano diverse dimensioni semantiche.

**Esempi**:

* Per ogni frase generata dal modello, crea un vettore che rappresenti il significato semantico complessivo e scompone la frase in sotto-vettori che catturano aspetti specifici come tono, genere delle parole, sentimenti impliciti.
* Usa **embedding** di frasi prodotte da **LLMs** per creare rappresentazioni numeriche ad alta dimensionalità delle relazioni semantiche.

**Vantaggi**:

* **Cattura di dettagli sottili**: I vettori permettono di rappresentare diverse dimensioni del testo, migliorando la rilevazione di sfumature linguistiche che potrebbero nascondere bias.
* **Flessibilità**: Le rappresentazioni multivettoriali permettono di analizzare testi in profondità, senza perdita di informazioni importanti.

### 3. **Clusterizzazione Iniziale delle Rappresentazioni Testuali**

**Descrizione**: Applicare algoritmi di **clusterizzazione** per raggruppare le frasi generate in base alle loro somiglianze semantiche o strutturali. Questo step consente di identificare gruppi di frasi che possono avere bias simili o che riflettono contesti specifici.

**Esempi**:

* Usa **K-Means** per raggruppare frasi che contengono riferimenti a determinate categorie sociali (es. genere, etnia, classe sociale).
* Clusterizza frasi che usano stereotipi linguistici o che presentano termini polarizzati (es. "intelligente" vs "bello").

**Vantaggi**:

* **Individuazione di pattern ripetitivi**: La clusterizzazione ti permette di identificare pattern di bias che si ripetono attraverso frasi o contesti.
* **Riduzione del rumore**: Raggruppando frasi simili, puoi focalizzarti meglio su quelli che mostrano tendenze o bias rilevanti.

### 4. **Aggiungere un Livello di Analisi Sentimentale**

**Descrizione**: Dopo la clusterizzazione iniziale, applicare un’**analisi del sentiment** alle frasi per determinare se i cluster hanno un bias implicito verso tendenze negative o positive rispetto a certi gruppi sociali.

**Esempi**:

* Usa strumenti come **VADER** o **TextBlob** per analizzare il sentiment delle frasi nei cluster e identificare se descrivono certi gruppi sociali in modo sproporzionatamente negativo o positivo.
* Verifica se termini come "leader", "potente", "forte" sono collegati maggiormente a descrizioni di uomini, e termini come "emotivo" o "bello" a descrizioni di donne.

**Vantaggi**:

* **Identificazione di bias impliciti**: L'analisi sentimentale rivela eventuali connotazioni positive o negative associate a certi gruppi, anche se non sono esplicitamente dichiarate.
* **Approfondimento delle relazioni**: Permette di analizzare come i sentimenti positivi o negativi siano distribuiti tra i diversi cluster.

### 5. **Applicare Tecniche di Interpretabilità dei Modelli**

**Descrizione**: Per capire meglio come il modello sta generando i bias, integra tecniche di **interpretabilità dei modelli** che ti aiutano a comprendere quali caratteristiche influenzano maggiormente le decisioni delle LLM.

**Esempi**:

* Usa tecniche come **SHAP** o **LIME** per spiegare quali parole o frasi contribuiscono di più a una determinata classificazione di bias.
* Verifica come la LLM assegna pesi a certe parole quando genera testo e analizza se questi pesi indicano una preferenza verso certi attributi sociali.

**Vantaggi**:

* **Maggiore trasparenza**: Le tecniche di interpretabilità permettono di comprendere come e perché i bias vengono generati dalle LLM.
* **Correzioni mirate**: Sapendo quali caratteristiche generano bias, puoi intervenire in modo più preciso per mitigare i problemi.

### 6. **Clusterizzazione Gerarchica e Validazione**

**Descrizione**: Effettuare una **seconda clusterizzazione** gerarchica basata sui risultati dell’analisi sentimentale e dell’interpretabilità. Utilizza tecniche di validazione (es. **Silhouette Score**, **Davies-Bouldin Index**) per determinare la qualità dei cluster.

**Esempi**:

* Applica una **clusterizzazione gerarchica** per suddividere ulteriormente i cluster originari e identificare correlazioni più sottili tra frasi biasate o correlate ad attributi sensibili.
* Verifica la coerenza dei cluster con l'uso di tecniche di validazione per capire se i cluster creati rappresentano accuratamente i contesti di bias.

**Vantaggi**:

* **Precisione migliorata**: La clusterizzazione gerarchica consente di esplorare i dati in modo più fine e di scoprire bias che emergono solo in contesti specifici.
* **Validazione della coerenza**: La validazione migliora la qualità del processo di clusterizzazione, riducendo il rischio di rilevare falsi positivi.

### 7. **Uso di Modelli Generativi per Stress Testing**

**Descrizione**: Applicare modelli generativi, come **GPT-4** o **BART**, per testare il modello sotto stress, creando scenari estremi che potrebbero indurre bias latenti.

**Esempi**:

* Genera frasi usando **GPT-4** che contengono combinazioni di parole e concetti delicati (es. stereotipi culturali o sociali) e osserva come le LLM rispondono.
* Verifica come il modello tratta richieste delicate come "Descrivi una persona cattiva" o "Chi è il leader perfetto?" per osservare se genera bias in base a genere o etnia.

**Vantaggi**:

* **Rilevamento di bias nascosti**: Il testing sotto stress permette di identificare bias latenti che emergono solo in scenari estremi o poco comuni.
* **Prevenzione**: Migliora la robustezza delle LLM, prevenendo la manifestazione di bias in contesti critici o delicati.

### 8. **Clusterizzazione e Rilevamento del Bias a Livello Gerarchico**

**Descrizione**: Integrare un ciclo continuo di **clusterizzazione** e rilevazione del bias. A ogni ciclo, identifica nuovi bias e contesti, e continua il processo fino a raggiungere un risultato ottimale in termini di rilevamento di bias e precisione dei cluster.

**Esempi**:

* Dopo ogni ciclo, usa la clusterizzazione per suddividere ulteriormente le frasi rilevanti e applica nuovamente tecniche di rilevamento per scoprire bias più complessi.
* Implementa tecniche di **validazione automatica** per fermare il processo quando i cluster raggiungono un livello ottimale di precisione.

**Vantaggi**:

* **Esplorazione completa**: La clusterizzazione gerarchica permette di esplorare in profondità ogni contesto e bias.
* **Auto-miglioramento**: Il loop di rilevamento assicura che il processo continui fino a ottenere una copertura ottimale dei bias presenti.

### 9. **Integrazione con una Rete Avversaria**

**Descrizione**: Integrare il framework con una **rete avversaria**, addestrata sulle frasi biasate rilevate. La rete impara a rilevare e a mitigare automaticamente i bias man mano che vengono scoperti.

**Esempi**:

* Addestra una **rete avversaria** che impara a identificare e correggere automaticamente le frasi che mostrano bias, utilizzando un feedback continuo dal framework.
* La rete avversaria può essere utilizzata anche per segnalare contesti ad alto rischio di bias e suggerire modifiche nelle LLM.

**Vantaggi**:

* **Mitigazione automatica**: La rete avversaria permette di intervenire direttamente per correggere i bias, riducendo la necessità di interventi manuali.

**Miglioramento continuo**: Man mano che vengono rilevati nuovi bias, la rete avversaria diventa sempre più efficiente nel prevenirli e correggerli.

### OSSEVAZIONI: ****Adversarial + Clusterization****

**Come funziona**:

* Le **reti avversarie** (adversarial networks) sono progettate per rilevare e correggere i bias durante l'addestramento, generando dati che il discriminatore cerca di classificare. L'idea è che il modello diventi più robusto man mano che identifica e corregge i bias espliciti e sottili.
* La **clusterizzazione** raggruppa i dati in base alle loro somiglianze, rendendo possibile identificare pattern nascosti o indiretti. Una rete avversaria può essere utilizzata per migliorare la clusterizzazione, correggendo eventuali distorsioni e migliorando la qualità dei cluster.

**Vantaggi**:

* Le **reti avversarie** sono ottime per affrontare **bias espliciti** e **mitigarli** durante l'addestramento. Se usate correttamente, le reti avversarie sono in grado di rilevare bias sottili e complessi che emergono nei dati.
* **Clusterizzazione** aiuta a trovare correlazioni nascoste tra i dati, e con una rete avversaria per correggere i cluster, si possono identificare **bias impliciti** che emergono in determinate categorie.

**Svantaggi**:

* Questo approccio può diventare complesso da gestire man mano che i dati crescono, e potrebbe non essere efficace nel separare i bias se non viene combinato con tecniche che trattano specificamente le **rappresentazioni latenti**.
* Le reti avversarie possono portare a un aumento del rumore nei dati o a un **overfitting** se non gestite correttamente.

### ****Disentanglement + Regularization****

**Come funziona**:

* **Disentanglement** separa le rappresentazioni latenti, disaccoppiando attributi sensibili (come genere, etnia) da altre variabili, in modo che il modello non usi tali informazioni per prendere decisioni.
* La **regolarizzazione** aggiunge una penalizzazione (loss function) che disincentiva il modello dall'associare certi attributi sensibili a output specifici, riducendo così il bias.

**Vantaggi**:

* Il **disentanglement** è particolarmente efficace per **bias impliciti e latenti**, in quanto consente di isolare e neutralizzare le informazioni sensibili che possono altrimenti influenzare il modello.
* La **regolarizzazione** riduce attivamente i bias aggiungendo penalizzazioni durante l'addestramento, senza modificare radicalmente la struttura del modello.
* Molto utile in scenari in cui il bias è **indiretto** o non facilmente rilevabile dalle reti avversarie.

**Svantaggi**:

* Questo approccio potrebbe non funzionare bene se non hai un **controllo preciso** sulle rappresentazioni latenti o se i bias sono particolarmente nascosti nei dati.
* La regolarizzazione può ridurre la flessibilità del modello e la sua capacità di generalizzare.

### ****Quale Metodo è Migliore?****

1. **Adversarial + Clusterization** è utile quando:
   * Vuoi rilevare bias **espliciti** e **sottili** che emergono in cluster distinti.
   * Stai lavorando con grandi dataset in cui la clusterizzazione può aiutarti a trovare gruppi di dati biasati che non sarebbero evidenti altrimenti.
   * Il tuo obiettivo è avere un sistema in grado di **auto-migliorarsi** continuamente tramite la rete avversaria.
2. **Disentanglement + Regularization** è utile quando:
   * Vuoi **rilevare bias latenti** che non sono visibili esplicitamente nel testo o nelle immagini, ma influenzano i risultati del modello.
   * Il focus è sulla **neutralizzazione degli attributi sensibili** nel processo di rappresentazione, riducendo il rischio che influenzino le decisioni del modello.
   * Vuoi assicurarti che il tuo modello resti **robusto** senza sovraccaricarlo con tecniche complesse come le reti avversarie.

### ****Combinazione Ottimale: Un Approccio Ibrido****

Una possibile **strategia ibrida** potrebbe combinare **reti avversarie** con **disentanglement e regolarizzazione**, in modo da rilevare bias espliciti attraverso il clustering e mitigare i bias latenti con il disentanglement.

* **Fase 1**: Usa **disentanglement** per separare le rappresentazioni latenti legate agli attributi sensibili. Questo ti consente di eliminare i bias che potrebbero non essere visibili durante l'addestramento iniziale.
* **Fase 2**: Integra una **rete avversaria** per sfidare continuamente il modello nel rilevare bias nuovi o non previsti, migliorando la robustezza del modello e garantendo che non sviluppi nuovi bias durante l'addestramento.
* **Fase 3**: Applica la **clusterizzazione** sui dati per identificare gruppi di frasi o immagini che contengono bias simili, e usa le reti avversarie per migliorare la qualità dei cluster, eliminando i falsi positivi.

### ****Miglioramenti e Altri Metodi da Considerare****

1. **Bias Measurement Tools**: Strumenti come **SHAP** o **LIME** possono essere integrati per analizzare meglio le decisioni del modello e identificare i punti in cui emergono i bias.
2. **Self-Supervised Learning**: Usare tecniche di **apprendimento auto-supervisionato** per far emergere bias impliciti da dataset non etichettati, creando rappresentazioni utili per l'analisi successiva.
3. **Multi-Task Learning**: Combinare il rilevamento del bias con altri compiti, come l'analisi del sentiment, per arricchire ulteriormente le informazioni sui contesti in cui emergono bias.
4. **Multi-Modal Learning**: Se si lavora con modelli multimodali (testo + immagini), usare rappresentazioni combinate per vedere come i bias possono emergere in **modi diversi** attraverso i dati visivi e testuali.

### ****Conclusione****

L'approccio **Adversarial + Clusterization** è ottimo per rilevare bias espliciti e sottili nei dati grazie alla combinazione di reti avversarie e tecniche di clustering. Tuttavia, per ottenere una copertura più completa dei **bias impliciti e latenti**, l'integrazione di **disentanglement e regolarizzazione** potrebbe portare a una soluzione più solida. Un approccio ibrido che combini le tecniche menzionate è probabilmente il più efficace per rilevare e mitigare ogni tipo di bias.

### Ottimizzazione del Framework di Rilevamento del Bias nelle **Large Language Models (LLM)** con Tecniche Avanzate

Di seguito, viene ottimizzato il framework proposto, integrando diverse tecniche avanzate per migliorare il rilevamento, la mitigazione e l'interpretabilità del bias. L'obiettivo finale è quello di rilevare **ogni tipo di bias possibile**, comprese le forme indirette e latenti, sfruttando approcci che combinano clusterizzazione, reti avversarie, regolarizzazione, tecniche di validazione e modelli generativi.

### Step 1: **Raccolta e Incrocio di Dataset Testuali Diversificati**

**Descrizione**: Iniziare raccogliendo dataset da fonti diverse (enciclopediche, social media, recensioni). Assicurati di includere testi che riflettano diversi contesti sociali, culturali, etnici e di genere.

**Vantaggi**:

* **Copertura di contesti multipli**: Maggiore diversità nei dati aiuta a rilevare bias legati a specifici contesti culturali o sociali.
* **Ricchezza di linguaggi**: Sfruttare dataset con linguaggi diversi (formale, informale, opinioni) amplia la capacità di rilevamento del bias.

### Step 2: **Rappresentazione Multivettoriale delle Frasi Generate dalle LLM**

**Descrizione**: Applica le **LLM** (ad esempio, GPT-4, LLAMA) per generare rappresentazioni multivettoriali. Ogni frase è convertita in un insieme di vettori che catturano varie dimensioni semantiche e stilistiche (tono, implicazioni, etc.).

**Vantaggi**:

* **Analisi profonda**: La rappresentazione multivettoriale permette di catturare le sfumature nel testo che potrebbero nascondere bias.
* **Flessibilità**: Può rappresentare varie dimensioni del linguaggio senza perdita di informazioni.

### Step 3: **Clusterizzazione Iniziale delle Rappresentazioni Testuali**

**Descrizione**: Utilizzare algoritmi di clusterizzazione (come K-Means) per raggruppare le frasi generate. Raggruppa frasi che contengono termini associati a categorie sensibili (genere, etnia).

**Vantaggi**:

* **Individuazione di pattern**: Identifica frasi con bias simili e permette di capire come il bias si distribuisce nei dati.
* **Riduzione del rumore**: Rende più facile concentrarsi sui bias rilevanti.

### **Step 4: Aggiunta di Tecniche di Disentanglement + Regularization**

**Descrizione**: Integrare tecniche di **disentanglement** che separano le rappresentazioni latenti degli attributi sensibili dalle altre caratteristiche semantiche. Aggiungi **regularization** per penalizzare il modello quando rileva correlazioni non legittime tra gli attributi sensibili e l'output.

**Vantaggi**:

* **Rilevamento di bias latenti**: Il disentanglement separa gli attributi sensibili, rendendo più facile rilevare bias impliciti o non evidenti.
* **Regolarizzazione**: Penalizzando le correlazioni errate, riduci il rischio che il modello associ inconsciamente certe caratteristiche a bias.

### Step 5: **Analisi Sentimentale sui Cluster**

**Descrizione**: Applicare strumenti di **analisi del sentiment** (VADER, TextBlob) per rilevare eventuali tendenze negative o positive verso certi gruppi all'interno dei cluster.

**Vantaggi**:

* **Bias impliciti**: Rilevi bias che possono non essere espliciti ma che emergono attraverso connotazioni positive o negative associate ai gruppi.
* **Conferma delle relazioni**: Approfondisci come i sentimenti siano distribuiti tra i diversi cluster, identificando le associazioni più problematiche.

### Step 6: **Applicazione di Tecniche di Interpretabilità dei Modelli**

**Descrizione**: Utilizzare tecniche di **interpretabilità** come SHAP o LIME per analizzare come il modello prende decisioni. Scopri quali caratteristiche influenzano maggiormente le decisioni e come queste sono legate a bias.

**Vantaggi**:

* **Trasparenza**: Capisci come il modello genera bias e quali caratteristiche specifiche sono problematiche.
* **Mitigazione mirata**: Sapendo quali caratteristiche sono problematiche, puoi intervenire per ridurre l’impatto dei bias.

### Step 7: **Clusterizzazione Gerarchica con Validazione**

**Descrizione**: Applicare una **seconda clusterizzazione** gerarchica sui risultati dell'analisi precedente e utilizza tecniche di validazione come **Silhouette Score** o **Davies-Bouldin Index** per valutare la qualità dei cluster e la rilevanza del contesto.

**Vantaggi**:

* **Migliore precisione**: Permette di raffinare ulteriormente i cluster e di rilevare bias che emergono solo in contesti specifici.
* **Validazione automatica**: Riduci il rischio di falsi positivi o interpretazioni errate.

### Step 8: **Uso di Modelli Generativi per Stress Testing**

**Descrizione**: Applicare modelli generativi (come **GPT-4** o **BART**) per creare scenari estremi e osservare come il modello risponde a input potenzialmente problematici.

**Vantaggi**:

* **Scoperta di bias latenti**: Generare situazioni estreme permette di identificare bias che emergono solo in circostanze non ordinarie.
* **Prevenzione**: Rafforza la capacità del modello di resistere ai bias quando viene esposto a casi limite.

### Step 9: **Clusterizzazione e Rilevamento del Bias a Livello Gerarchico**

**Descrizione**: Implementare un ciclo continuo di **clusterizzazione e rilevazione del bias**, identificando progressivamente nuovi contesti e bias più complessi a ogni iterazione.

**Vantaggi**:

* **Esplorazione completa**: Esplori in profondità ogni contesto di bias e ottieni una mappa più dettagliata delle problematiche.
* **Miglioramento automatico**: Il ciclo continua finché non hai raggiunto un risultato ottimale in termini di rilevamento e precisione.

### Step 10: **Integrazione tra Clusterizzazione e Rete Avversaria (Adversarial + Clusterization)**

**Descrizione**: Integrare **reti avversarie** con la **clusterizzazione** per creare un ciclo continuo di rilevamento e mitigazione del bias. Durante la fase di clusterizzazione, la rete avversaria è addestrata a rilevare automaticamente i bias emergenti nei cluster e correggere le distorsioni. Man mano che il modello genera nuovi output, il sistema clusterizza i risultati e la rete avversaria interviene per rilevare e mitigare i bias in tempo reale, aggiornandosi dinamicamente.

**Esempi**:

* Durante la clusterizzazione iniziale degli output generati dalla LLM, la rete avversaria identifica pattern ricorrenti di bias (ad esempio, se le frasi clusterizzate associate a un determinato genere presentano connotazioni negative).
* La rete avversaria può suggerire modifiche agli output, in modo da ridurre il bias nei successivi processi di clusterizzazione.
* Man mano che vengono rilevati nuovi contesti o bias attraverso la clusterizzazione, la rete avversaria si adatta e migliora automaticamente il processo.

**Vantaggi**:

* **Correzione automatica e dinamica**: La rete avversaria non si limita a rilevare bias nei cluster, ma agisce direttamente per correggere le frasi biasate, migliorando la qualità dei cluster nel tempo.
* **Ottimizzazione del rilevamento**: Combinare clusterizzazione e reti avversarie consente di catturare bias più sottili e complessi, poiché la rete avversaria aiuta a perfezionare continuamente il processo di clustering.
* **Miglioramento continuo**: Il sistema si auto-adatta e si aggiorna man mano che scopre nuovi bias, rendendo il processo più robusto nel tempo.

**Efficienza**: Integrare questi due passaggi riduce la ridondanza e crea un unico flusso in cui clusterizzazione e mitigazione si rinforzano a vicenda.

### Framework di Rilevamento del Bias nelle Large Language Models (LLM) con Clusterizzazione Automatica, Reti Avversarie e Knowledge Graphs (KGs)

Il framework proposto mira a identificare, monitorare e mitigare il bias all'interno dei modelli linguistici su larga scala (LLM), utilizzando tecniche di clusterizzazione automatica, reti avversarie e **Knowledge Graphs (KGs)**. L'integrazione dei **KGs** aggiunge un livello di comprensione semantica e contestuale, migliorando l'accuratezza del rilevamento e la mitigazione del bias in contesti demografici, culturali e sociali.

**1. Raccolta e Incrocio di Dataset Testuali Diversificati**

**Descrizione:** Raccogliere dataset testuali da varie fonti come Wikipedia, Twitter, Reddit, IMDB, e incrociali per ottenere una varietà di testi che riflettano diverse opinioni, stereotipi e contesti sociali. I **KGs** vengono utilizzati per aggiungere contesto semantico e relazioni tra entità (come professioni, etnie, generi) presenti nei dataset, migliorando l'analisi dei bias.

**Esempio di uso dei KGs:** Immagina di raccogliere dati da Wikipedia e Reddit. Un **KG** può essere utilizzato per collegare le entità come "medico", "donna", "uomo", e "genere" e identificare come spesso la parola "medico" sia più associata a uomini rispetto alle donne. Questo potrebbe rivelare un bias che porta i modelli a generare descrizioni stereotipate di professioni basate sul genere.

**Vantaggi:** I **KGs** consentono di creare un contesto più ricco attorno ai dati, collegando entità come genere o professione per rivelare eventuali bias già presenti nei dati raccolti.

**2. Rappresentazione Multivettoriale delle Frasi Generate dalle LLM**

**Descrizione:** Le frasi generate dai modelli vengono trasformate in rappresentazioni multivettoriali, catturando più dimensioni del significato. L'integrazione dei **KGs** permette di collegare queste rappresentazioni a entità e relazioni, arricchendo la comprensione del contesto semantico.

**Esempio di uso dei KGs:** Se il modello genera una frase come "Lui è un leader naturale", il **KG** può collegare le parole "lui" e "leader" a relazioni di genere e leadership, rivelando potenziali bias se, ad esempio, viene preferibilmente utilizzato il maschile per descrivere ruoli di leadership, mentre termini come "collaborativa" o "emotiva" vengono usati più frequentemente per donne.

**Vantaggi:** La rappresentazione multivettoriale arricchita con **KGs** permette di catturare connessioni profonde tra concetti che potrebbero altrimenti essere trascurati, facilitando l'identificazione di bias nascosti.

**3. Clusterizzazione Iniziale delle Rappresentazioni Testuali**

**Descrizione:** Applicare algoritmi di clusterizzazione per raggruppare frasi simili. I **KGs** aiutano ad arricchire questa clusterizzazione aggiungendo relazioni semantiche tra concetti, evidenziando meglio i gruppi che potrebbero contenere bias.

**Esempio di uso dei KGs:** Supponiamo di avere un cluster di frasi che descrivono leadership, ma grazie ai **KGs**, scopri che le frasi associate alla leadership sono prevalentemente collegate a uomini. I **KGs** possono rivelare queste associazioni implicite, mostrando come il linguaggio della leadership sia collegato più spesso al genere maschile.

**Vantaggi:** I **KGs** migliorano la clusterizzazione aggiungendo una dimensione semantica più profonda, rivelando associazioni che potrebbero essere mascherate nelle rappresentazioni vettoriali semplici.

**4. Aggiungere un Livello di Analisi Sentimentale**

**Descrizione:** Aggiungere un'analisi del sentiment ai cluster per determinare se vi sono tendenze negative o positive associate a specifici gruppi. I **KGs** possono aiutare a identificare se certe entità (come razze o professioni) sono associate a sentimenti sproporzionatamente negativi o positivi.

**Esempio di uso dei KGs:** Supponi di avere frasi che descrivono diverse professioni. Con l'aiuto di un **KG**, puoi notare che professioni come "infermiere" vengono frequentemente associate a sentimenti positivi quando descrivono le donne, ma neutri o negativi per gli uomini. Questo può indicare un bias di genere nelle descrizioni.

**Vantaggi:** I **KGs** permettono di collegare le entità a sentimenti espliciti e impliciti, evidenziando meglio come certi gruppi siano descritti in modo sproporzionato in termini negativi o positivi.

**5. Applicare Tecniche di Interpretabilità dei Modelli**

**Descrizione:** Usare tecniche come SHAP o LIME per capire quali caratteristiche influenzano maggiormente le decisioni delle LLM. I **KGs** possono essere utilizzati per spiegare quali entità o relazioni influenzano il risultato.

**Esempio di uso dei KGs:** Utilizzando **LIME**, puoi scoprire che la parola "direttore" viene costantemente associata a uomini. Un **KG** può collegare questa associazione a bias storici o culturali, aiutando a spiegare come il modello stia perpetuando stereotipi culturali.

**Vantaggi:** I **KGs** aggiungono trasparenza alle decisioni del modello, mostrando le relazioni tra le parole e le entità culturali o sociali che contribuiscono al bias.

**6. Clusterizzazione Gerarchica e Validazione**

**Descrizione:** Effettuare una seconda clusterizzazione gerarchica utilizzando i **KGs** per arricchire i cluster con relazioni culturali e demografiche. I **KGs** migliorano la validazione assicurando che i cluster catturino accuratamente i bias presenti.

**Esempio di uso dei KGs:** Supponi di aver scoperto un cluster di frasi che descrivono professioni legate alla cura. Un **KG** potrebbe rivelare che queste professioni sono principalmente descritte con un linguaggio associato al genere femminile, rivelando un bias nascosto nella rappresentazione delle professioni assistenziali.

**Vantaggi:** La clusterizzazione gerarchica con i **KGs** fornisce una visione più approfondita dei bias nascosti e ne migliora la validazione attraverso il contesto semantico aggiuntivo.

**7. Uso di Modelli Generativi per Stress Testing**

**Descrizione:** Utilizzare modelli generativi come GPT-4 per creare scenari estremi che inducano bias latenti, utilizzando i **KGs** per generare frasi che includono concetti demografici e culturali sensibili.

**Esempio di uso dei KGs:** Genera frasi che includono stereotipi di genere o razziali, come "Descrivi un grande leader" o "Chi è il miglior scienziato", e usa i **KGs** per verificare come il modello risponde, collegando entità come "scienziato" a razze o generi specifici.

**Vantaggi:** I **KGs** migliorano lo stress testing rivelando bias che emergono solo in scenari culturalmente delicati, permettendo di prevenire bias in situazioni reali.

**8. Clusterizzazione e Rilevamento del Bias a Livello Gerarchico**

**Descrizione:** Integrare un ciclo continuo di clusterizzazione, supportato dai **KGs**, per migliorare il rilevamento dei bias in contesti demografici e culturali sempre più sottili.

**Esempio di uso dei KGs:** Dopo ogni ciclo di clusterizzazione, un **KG** può essere utilizzato per identificare nuove relazioni tra concetti come "ruoli di genere" o "razza", rivelando bias più complessi che emergono solo dopo più iterazioni.

**Vantaggi:** I **KGs** assicurano che il processo di rilevamento del bias continui a evolversi, permettendo di esplorare nuove relazioni semantiche e bias più profondi.

**9. Integrazione con una Rete Avversaria**

**Descrizione:** Utilizzare una rete avversaria, supportata dai **KGs**, per rilevare e correggere automaticamente i bias. I **KGs** possono fornire relazioni tra entità demografiche, culturali e sociali che migliorano l'efficacia della rete avversaria nel mitigare i bias.

**Esempio di uso dei KGs:** Se un **KG** rivela che le descrizioni di leadership sono sempre collegate al genere maschile, la rete avversaria può imparare a identificare e correggere automaticamente queste associazioni, suggerendo termini più neutrali o bilanciati.

**Vantaggi:** I **KGs** arricchiscono la rete avversaria con contesti più ricchi e relazioni tra entità, migliorando l'efficacia nella rilevazione e correzione dei bias.

### Perché la Proposta Migliora lo Stato dell'Arte

Il framework proposto per il rilevamento e la mitigazione del bias nei **Large Language Models (LLM)**, integrando **clusterizzazione automatica** e **reti avversarie**, rappresenta un avanzamento significativo rispetto alle tecniche esistenti per vari motivi:

1. **Clusterizzazione per rilevare bias nascosti**: La clusterizzazione automatica consente di raggruppare dati in base a somiglianze semantiche, evidenziando bias che potrebbero non essere immediatamente visibili. Questo approccio supera le limitazioni dei metodi tradizionali che spesso rilevano solo bias espliciti. Con la clusterizzazione, è possibile identificare schemi ripetitivi in cui si verificano pregiudizi, ampliando la capacità di scoprire bias nascosti nei dati testuali.

**Avanzamento**: Il metodo permette di esplorare più profondamente le relazioni tra frasi e contesti, fornendo una comprensione più completa dei bias impliciti o latenti.

1. **Reti avversarie per la correzione del bias**: Le reti avversarie (GANs) aggiungono un livello di auto-miglioramento continuo, permettendo al sistema di identificare e correggere i bias in modo dinamico. Le GANs possono essere addestrate per riconoscere pattern di bias e suggerire modifiche, rendendo il modello più equo. Questo introduce un processo di apprendimento iterativo, che raffina costantemente il rilevamento e la correzione.

**Avanzamento**: Le GANs apportano un approccio proattivo alla mitigazione dei bias, correggendo in modo dinamico i problemi man mano che vengono rilevati, il che è una novità rispetto ai modelli statici di correzione dei bias.

1. **Rilevamento gerarchico e miglioramento progressivo**: L'uso della clusterizzazione gerarchica permette di scendere in profondità nell'analisi del bias, rivelando pregiudizi più complessi o nascosti che potrebbero emergere solo in contesti specifici. Ogni iterazione consente di perfezionare i cluster, migliorando la precisione del rilevamento e permettendo un'esplorazione approfondita delle sfumature culturali e sociali che influenzano il bias.

**Avanzamento**: Il miglioramento continuo attraverso iterazioni di clusterizzazione gerarchica consente di rilevare bias sempre più sofisticati, rappresentando un progresso rispetto ai sistemi esistenti che tendono a limitarsi alla rilevazione di bias espliciti.

1. **Combina rilevamento quantitativo e qualitativo**: La proposta utilizza sia analisi quantitative (es. clusterizzazione, analisi vettoriale) che qualitative (es. sentiment analysis, tecniche interpretative) per comprendere a fondo il bias. Questa integrazione rende il framework completo e versatile, consentendo un'analisi più dettagliata dei contesti in cui emergono i bias.

**Avanzamento**: Un approccio ibrido quantitativo-qualitativo fornisce una comprensione più ricca dei bias, superando i limiti dei framework che si concentrano esclusivamente su una delle due dimensioni.

1. **Stress testing con modelli generativi**: Il test sotto stress con frasi e scenari estremi generati dai LLM consente di identificare i bias latenti che potrebbero emergere solo in circostanze specifiche o non comuni. Questa strategia non è comunemente utilizzata nei metodi attuali e rappresenta un'innovazione nella scoperta dei bias nascosti.

**Avanzamento**: Questa capacità di anticipare e testare bias latenti porta il rilevamento del bias a un nuovo livello di robustezza e prevenzione.

### Limiti della Proposta

1. **Complessità computazionale**: Il processo di clusterizzazione iterativa e l'integrazione di GANs richiede un notevole potere computazionale e risorse di addestramento. La complessità aumenta man mano che i dati e le iterazioni crescono, rischiando di rendere il framework meno scalabile in contesti con dataset di grandi dimensioni.

**Limite**: L'alto costo computazionale potrebbe limitare l'applicabilità del framework in ambienti con risorse limitate.

1. **Rischio di overfitting**: Le reti avversarie, se non gestite correttamente, possono portare a un overfitting sui dati utilizzati per l'addestramento, riducendo la capacità del sistema di generalizzare correttamente su dati nuovi o inattesi.

**Limite**: È essenziale trovare un equilibrio tra la correzione dei bias e la capacità del modello di rimanere flessibile di fronte a nuovi dati. Un'implementazione errata potrebbe portare a una riduzione della generalizzazione.

1. **Difficoltà di interpretazione delle GANs**: Le reti avversarie, pur essendo potenti strumenti di mitigazione, sono spesso considerate una "black box". Capire come funzionano esattamente le correzioni apportate dalle GANs può essere difficile, limitando la trasparenza e l'affidabilità del sistema.

**Limite**: La scarsa interpretabilità potrebbe rendere difficile diagnosticare errori o comprendere le ragioni dietro le correzioni del bias.

1. **Validazione del processo di clusterizzazione**: La clusterizzazione automatica può essere sensibile alla qualità dei dati e alle metriche di similarità utilizzate. Un errore nella scelta di questi parametri può portare a cluster inappropriati o a una sovra-segmentazione, che rischia di non riflettere accuratamente i bias reali.

**Limite**: Il processo di clusterizzazione richiede un'attenta calibrazione per garantire che i risultati siano validi e riflettano correttamente i dati.

**ANALOGAMENTE PUO’ ESSERE APPLICATO UN PROCESSO SIMILE SUGLI VLM**

### Framework per il Rilevamento e la Mitigazione del Bias nei **Vision-Language Models (VLM)**

Il seguente framework è progettato per rilevare, monitorare e mitigare i bias presenti nei **Vision-Language Models (VLM)**. Questo sistema combina tecniche di **clusterizzazione automatica**, analisi di **rappresentazioni multimodali**, e l'integrazione di **reti avversarie** per l'auto-miglioramento continuo. L'obiettivo principale è quello di costruire un sistema in grado di **rilevare ogni tipo di bias** nei VLM e mitigare automaticamente gli effetti negativi, mantenendo alte prestazioni.

Il framework sfrutta immagini e testo, trattando sia bias visivi che testuali, e utilizza tecniche avanzate di **deep learning** e **machine learning** per garantire una rilevazione precisa e una mitigazione efficiente.

### 1. **Raccolta e Incrocio di Dataset Multimodali**

**Descrizione**: Il primo passo consiste nel raccogliere **dataset multimodali** che includono immagini e corrispondenti descrizioni testuali. Per garantire la massima copertura, i dataset dovrebbero essere diversificati per includere diverse culture, contesti geografici, generi e classi sociali. Incrociare più dataset permette di ottenere una rappresentazione più completa dei bias presenti.

**Esempi**:

* **COCO**: dataset di immagini annotate con descrizioni testuali dettagliate.
* **Visual Genome**: dataset di immagini con annotazioni semantiche complesse, che coprono scene e oggetti.
* **FairFace** o **UTKFace**: dataset di volti con annotazioni di genere ed etnia.
* **Flickr30K**: contiene immagini e descrizioni di utenti, spesso con opinioni e pregiudizi impliciti.

**Vantaggi**:

* **Ampiezza di contesto**: Incrociare più dataset garantisce che il modello riesca a identificare bias legati a contesti culturali e sociali diversi.
* **Rappresentazioni bilanciate**: Includendo dataset con annotazioni demografiche, si riesce a cogliere meglio le disuguaglianze legate alla rappresentazione visiva e testuale.

### 2. **Rappresentazione Multivettoriale delle Immagini e del Testo**

**Descrizione**: Utilizzare i **Vision-Language Models (VLM)**, come **CLIP** o **BLIP**, per generare rappresentazioni multivettoriali delle immagini e dei loro testi associati. Le immagini vengono mappate in uno spazio vettoriale, così come i testi, e possono essere confrontate per rilevare somiglianze semantiche e possibili bias.

**Esempi**:

* Crea rappresentazioni vettoriali di descrizioni testuali e immagini utilizzando modelli come **CLIP**, che mappa entrambe le modalità nello stesso spazio di embedding.
* Estrai i vettori per immagini contenenti volti di diversi generi ed etnie e rappresenta i testi descrittivi associati, come “uomo d’affari”, “madre amorevole”, per vedere come sono rappresentati nei diversi contesti.

**Vantaggi**:

* **Rappresentazione profonda e ricca**: I vettori generati catturano le relazioni semantiche complesse tra immagini e testi.
* **Flessibilità**: Le rappresentazioni multivettoriali consentono di analizzare più dimensioni di bias, sia visivo che testuale, senza perdita di dettaglio.

### 3. **Clusterizzazione Iniziale delle Rappresentazioni Multimodali**

**Descrizione**: Applicare algoritmi di **clusterizzazione** per raggruppare le immagini e i testi in base alle loro somiglianze semantiche o strutturali. Questo aiuta a individuare gruppi di immagini e testi che condividono tratti comuni e possono contenere bias simili.

**Esempi**:

* Usa **K-Means** o **DBSCAN** per raggruppare immagini e testi che hanno descrizioni simili (es. "uomo forte" per gli uomini, "donna gentile" per le donne).
* Clusterizza le rappresentazioni vettoriali di immagini di volti e testi associati per verificare se le descrizioni stereotipate si accumulano in certi gruppi.

**Vantaggi**:

* **Individuazione di pattern di bias**: La clusterizzazione iniziale permette di identificare immediatamente gruppi di immagini e descrizioni con tratti stereotipati.
* **Riduzione del rumore**: Raggruppando le rappresentazioni si può lavorare più facilmente su insiemi che condividono caratteristiche rilevanti per l'analisi.

### 4. **Aggiungere un Livello di Analisi Sentimentale**

**Descrizione**: Integrare un’**analisi del sentiment** sui testi descrittivi associati alle immagini per identificare eventuali connotazioni positive o negative associate a certi gruppi demografici. Questo livello aggiuntivo è particolarmente utile per identificare bias impliciti.

**Esempi**:

* Usa strumenti come **VADER** per misurare il sentiment delle descrizioni testuali associate a immagini di persone e verifica se le descrizioni positive ("successo", "potente") sono collegate a certi gruppi demografici (ad es. uomini bianchi).
* Analizza descrizioni testuali associate a immagini di diversi contesti (lavoro, scuola) per vedere se esistono connotazioni negative o stereotipi legati a certe professioni.

**Vantaggi**:

* **Rilevamento di bias impliciti**: L'analisi del sentiment permette di scoprire bias nascosti che non emergono esplicitamente dalle descrizioni.
* **Analisi combinata**: Questa fase consente di correlare il sentiment con i cluster di immagini, migliorando la rilevazione di bias nel dominio visivo.

### 5. **Applicare Tecniche di Interpretabilità dei Modelli**

**Descrizione**: Per comprendere meglio come i VLM generano i bias, utilizza tecniche di **interpretabilità** che ti aiutano a capire quali caratteristiche influenzano le decisioni del modello. Questo step permette di identificare le ragioni dietro certe rappresentazioni stereotipate.

**Esempi**:

* Usa tecniche come **SHAP** o **Grad-CAM** per capire quali parti delle immagini o quali parole nei testi associati contribuiscono maggiormente alla generazione di bias.
* Analizza come **CLIP** assegna importanza a determinate parole o oggetti nelle immagini, e verifica se certi oggetti (es. un vestito) portano a stereotipi di genere.

**Vantaggi**:

* **Trasparenza del modello**: Le tecniche di interpretabilità aiutano a capire come i modelli generano bias, offrendo spiegazioni più chiare.
* **Correzione mirata**: Sapendo quali caratteristiche sono responsabili dei bias, puoi intervenire in modo più preciso per ridurre i problemi.

### 6. **Clusterizzazione Gerarchica e Validazione**

**Descrizione**: Applicare una **seconda clusterizzazione** a livello gerarchico utilizzando tecniche di validazione come il **Silhouette Score** o il **Davies-Bouldin Index**. Questa clusterizzazione serve per affinare ulteriormente l'identificazione di bias sottili e correlazioni nascoste tra i cluster.

**Esempi**:

* Usa **clusterizzazione gerarchica** per suddividere ulteriormente i cluster di immagini e testi, cercando sottogruppi con bias latenti.
* Valida i cluster usando **Davies-Bouldin Index** per confermare che siano effettivamente rappresentativi dei bias rilevati.

**Vantaggi**:

* **Miglioramento della precisione**: La clusterizzazione gerarchica permette di esplorare i dati a un livello più fine, scoprendo bias che emergono solo in contesti molto specifici.
* **Validazione rigorosa**: L'uso di metriche di validazione garantisce che i cluster siano rappresentativi e precisi, riducendo il rischio di rilevare falsi positivi.

### 7. **Uso di Modelli Generativi per Stress Testing**

**Descrizione**: Usare modelli generativi come **DALL-E** o **Stable Diffusion** per creare scenari estremi con combinazioni di immagini e descrizioni che potrebbero indurre bias latenti nel modello. Questo permette di verificare la resistenza dei VLM a situazioni inusuali.

**Esempi**:

* Genera immagini di scenari ambigui (es. una persona con abiti non convenzionali) e osserva come i modelli trattano la descrizione testuale associata (es. "una persona rispettabile").
* Crea immagini con sfondi complessi e descrizioni ambigue per testare se i modelli assegnano bias legati a genere o etnia in contesti non chiari.

**Vantaggi**:

* **Rilevamento di bias nascosti**: Lo stress testing rivela bias latenti che si manifestano solo in situazioni estreme o ambigue.
* **Resistenza migliorata**: Testare il modello in scenari complessi aiuta a rendere i VLM più robusti e meno suscettibili ai bias.

### 8. **Integrazione con una Rete Avversaria**

**Descrizione**: Integrare il framework con una **rete avversaria** per la mitigazione automatica dei bias rilevati. Man mano che vengono rilevati bias attraverso la clusterizzazione e l'analisi, la rete avversaria impara a neutralizzare questi bias, migliorando continuamente la sua efficacia.

**Esempi**:

* Addestra una rete avversaria per correggere i bias presenti nelle descrizioni testuali associate a certe immagini, riducendo la frequenza di errori.
* Integra la rete avversaria con il VLM per applicare correzioni automatiche durante la generazione di testo o l'assegnazione di descrizioni a immagini.

**Vantaggi**:

* **Mitigazione automatica**: La rete avversaria elimina i bias appena vengono rilevati, riducendo la necessità di interventi manuali.
* **Auto-miglioramento continuo**: La rete avversaria diventa più efficace con il tempo, migliorando man mano che rileva più bias.

### 9. **Implementazione di un Loop di Clusterizzazione Automatica**

**Descrizione**: Per garantire il rilevamento continuo dei bias, implementa un **loop di clusterizzazione automatica**. Dopo ogni ciclo di rilevamento, il sistema riesegue la clusterizzazione, identificando nuovi bias man mano che emergono. Il processo si ripete fino a quando non viene raggiunta una copertura completa.

**Esempi**:

* Ogni volta che un nuovo cluster di immagini o testi viene identificato come potenzialmente biasato, il loop si attiva e ripete la clusterizzazione per confermare e approfondire il rilevamento.
* Applica la clusterizzazione in loop a dataset in tempo reale, come social media o news, per scoprire bias emergenti man mano che nuovi dati vengono generati.

**Vantaggi**:

* **Rilevamento completo**: Il loop garantisce che il processo continui fino a quando ogni bias potenziale viene identificato.

**Adattamento ai nuovi contesti**: Man mano che emergono nuovi dati o contesti, il loop si adatta automaticamente per includerli nell'analisi.