Preguntas sobre QGANs

QGANS tecniques for anomaly detection

Miranda Carou Laiño

1 ¿Diferentes tipos de GANs (clásicas y cuánticas) son mejores para diferentes tipos de datos/distribuciones para el caso de anomaly detection? ¿Es GANomaly la mejor?

En el ámbito clásico en el documento nombro 3 tipos de GANs para detección de anomalías: **AnoGAN**, **EGBAD** y **GANomaly**. Y en el ámbito clásico tenemos laS **qGANs**.

1. **AnoGAN**: Es el primer enfoque que se propuso para la detección de anomalías usando GANs pero tiene un rendimiento muy lento porque siempre necesita realizar una optimización en cada vuelta.

2. **EGBAD**: Tiene un mejor tiempo de ejecución que **AnoGAN** porque introduce un codificador que mapea las muestras de entrada al espacio latente durante el entrenamiento.

3. **GANomaly**: Combina un generador y un autoencoder. Esto nos permite detectar de una manera más eficiente de anomalías y una interpretación mas clara de las puntuaciones de anomalía. **GANomaly** supera a **EGBAD** tanto en métricas de evaluación como en velocidad de ejecución.

4. **qGANs**:Son interesantes y pueden ser competitivas para la detección de anomalías en especial para cuando *los datos son escasos* o de *alta dimensionalidad*. De momento no se ha podido afirmar que sean superiores a las **GANomaly** en todos los casos pero si ofrecen ventajas e casos específicos como la detección de anomalías en HEP donde podríamos obtener resultados buenos con menos datos. Hasta donde yo entiendo todas las arquitecturas de GANs clásicas se pueden traducir a qGANs modificadas pero no tienen un nombre diferente.

En definitiva la estructura de GANomaly es como la opción mas sólida en el ámbito clásico. En principio los 3 tipos de GANs clásicas sirven de igual manera para cualquier tipo de distribución o datos. Es más bien la estructura interna lo que tu adaptas (generador, codificación, discriminador, etc) a tus necesidades.

2 Si se 'visualizan' los datos como imágenes, ¿se puede sacar más partido de las qGANs? ¿Hasta cierto punto se puede saber a priori cuál es el encoding más óptimo sabiendo la distribución de datos?

Efectivamente las **qGANs** pueden ser muy útiles si se visualizan los datos como imágenes. En especial en aplicaciones donde la estructura espacial de los datos es importante. Hablando del **encoding**, los tipos de encoding que hay son:

1. **Angle encoding**: Muy útiles para datos continuos y se basa en la codificación de ángulos en puertas cuánticas.

2. **Amplitude encoding**: Es más eficiente para datos de alta dimensionalidad pero requiere que los datos estén normalizados.

Dentro de mi conocimiento, no siempre podemos saber a priori cuál es el mejor encoding ya que depende mucho de la distribución de los datos y de el problema específico. Si que es verdad que **angle encoding** es más flexible y puede adaptarse mejor a diferentes tipo de datos mientras que **amplitude encoding** es más eficiente para datos de alta dimensionalidad (con un gran número de variables).

3 ¿Se pueden obtener resultados competitivos con datasets más pequeños y un número reducido de variables?

Sí, por lo que leí las qGANs pueden lograr resultados muy competitivos con menos datos en comparación con métodos clásicos. Por ejemplo, en el estudio de IBM, ETH Zurich y CERN (paper número 2), las qGANs lograron una precisión comparable a los métodos clásicos utilizando **10 veces menos muestras de entrenamiento**.

4 ¿Con qudits, hay un enfoque implícito y explícito en cuanto a metric learning? ¿Hay alguno que sea mejor? ¿El explícito?

Esta creo que es la parte en la que aún tengo que trabajar yo en entenderla mejor porque igual ando algo cortita porque no termino de visualizarla del todo bien. Dicho esto.

1. **Enfoque implícito**: Este no utiliza estados de referencia fijos y maximiza la pureza de los estados dentro de la misma clase mientras que minimiza la superposición entre clases diferentes.
2. **Enfoque explícito**: Este utiliza estados de referencia fijos (centros) para cada clase y minimiza la diferencia entre los datos de entrenamiento y estos centros.

Lo que se sugiere en el paper que leí es que el enfoque implícito puede ser ligeramente superior en términos de rendimiento, ya que no depende de la elección de centros fijos y puede adaptarse mejor a la estructura de los datos. Sin embargo, el **enfoque explícito** es más práctico para implementaciones experimentales inmediatas, ya que simplifica la evaluación de la función de costo.

5 En la página 18 dice 'it may fail in scenarios where ignored features are crucial'. ¿Con ignored features se refiere a cosas que no se han usado en el training o que no se han medido?

Las **ignored features** se refieren a características de los datos que no se han utilizado en el entrenamiento o que no se han considerado en el proceso de encoding. Por ejemplo, en el **encoding g1** se decartan la mitad de los datos originales, tratándolos como fases relativas que no influyen en las proyecciones sobre la base ortonormal. Esto puede ser problemático en escenarios donde las características ignoradas pueden ser cruciales como por ejemplo para la clasificación.

6 ¿También se pueden usar VQNNs? ¿Son las VQNNs a las que se refiere la sección 5.2 parte de las qGANs? ¿Se usan como generador/discriminador?

Las **VQNNs** no on parte directa de las qGANS pero pueden ser utilizadas en conjunto con ellas. Por ejemplo las **VQNNs** podrían ser utilizadas como **generador** o **discriminador**, dependiendo de la arquitectura específica. Al final hasta donde yo entiendo en las qGANs los generadores y discriminadores no dejan de ser redes neuronales , que no dejan de ser en su forma más primitiva circuitos cuánticos. Por lo tanto, es una manera más de crear esas redes neuronales.

7 ¿Se puede buscar un caso específico de anomalía, tal que el dataset sea óptimo para QML con qudits? Aquí estoy pensando desde un punto de vista físico, el tipo de firma BSM que se espera.

Sí, es posible diseñar un dataset específico para la detección de anomalías en BSM que sea óptimo para QML con qudits. En alguno de los papers que leí se menciona que las qGANS son particularmente útiles en la detección de anomalías en HEP donde se esperan firmas BSM. Entiendo que la clave aquí es diseñar un dataset que capture las características específicas de las anomalías esperadas y utilizar un encoding adecuado (como puede ser **angle encoding**) para representar los datos en un estado cuántico.

8 ¿Hasta qué punto se puede sistematizar el uso del encoding?

Hasta donde yo se, el uso de encoding se puede sintetizar hasta cierto punto, pero depende en gran medida del tipo de datos y del problema específico. Por ejemplo tendriamos que tener en cuenta:

1. **Tipo de los datos**: Si los datos son continuos **angle encoding** puede ser más adecuado. Si los datos son de alta dimensionalidad, el **amplitude encoding**.
2. **Distribución de los datos**: Si los datos tienen una distribución específica (como pude ser log-normal), se puede elegir un encoding que se adapte mejor a ese tipop de distribución
3. **Recursos computacionales**: El encoding tiene que ser compatible con los recursos cuánticos disponibles.

También podríamos basándonos en los puntos anteriores crear un **encoding específico** a nuestras necesidades. ¿Que tipos de datos estamos manejando?¿Cual es nuestro propósito? Preparar los datos para el circuito, reducir la dimnsionalidad, etc. En el proceso de crear un encoding específico tendríamos que decidir que transformaciones necesitamos.

- Normalización de los datos
- Discretización
- Mapeo a un espacio de menor dimensionalidad (usando PCA o embeddings por ejemplo).
- Codificar los datos al formato que necesitamos para nuestro circuito cuántico. Como puede ser el mapeo de datos a ángulos de rotación en qutrits