Resumo do artigo: **Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling**

Contexto:
Neste artigo os autores fazem uma **comparação** entre três tipos de **redes recorrentes** (RNN - *recurrent neural networks*): a primeira com uma unidade mais antiga (proposta original de 1997) e bastante utilizada **LSTM** (long short-term memory); a segunda com uma unidade mais recente (2014) **GRU** (gated recurrent unit); e por fim a clássica tangente hiperbólica, ***tanh***. A comparação é feita tendo como base a modelagem de sequência, onde a probabilidade da sequência pode ser decomposta como: $\mathbf{p(x_1, \ldots, x_T) = p(x_1)\, p(x_2|x_1)\, p(x_3|x_1x_2) \ldots p(x_T|x_1, \ldots, x_T)}$, onde o último elemento é o *end-of-sequence*.

RNN background:
Dois **problemas** comumente encontrados em redes recorrentes são: **dificuldade de treinamento** (*vanishing and exploding gradient*) e dependências de curto prazo tendem a serem levadas mais em consideração do que **dependências de longo prazo**. Para a **solução** desses problemas, dois tipos de abordagens podem ser utilizadas: **melhores algoritmos de aprendizados**, ao invés de simplesmente usar SGD e/ou desenvolver **funções de ativação mais sofisticadas**, consistindo de transformação afim seguida de não-linearidades por meio do uso de *gating units*. Neste último contexto entram as unidades LSTM e GRU, sendo ambas competentes em capturar relações de longo prazo, como requerido por exemplo em tarefas de reconhecimento de fala e tradução.

- **LSTM:** ao contrário das redes recorrentes tradicionais que sobrescrevem seu conteúdo a cada passo de tempo, a unidade LSTM aprende quais memórias devem ser mantidas e quais podem ser esquecidas por meio das portas de entrada (***input gate***), saída (***output gate***) e esquecimento (***forget gate***). Assim, mesmo que uma *feature* importante tenha sido apresentada a rede vários passos atrás, a rede é capaz de detectá-la e mantê-la em sua memória, capturando portanto dependências de longas distâncias.
- **GRU:** A principal diferença entre a GRU e a LSTM, é que GRU apresenta apenas duas portas: ***update gate*** e ***reset gate***. A principal ideia por trás desta alteração é de que na unidade LSTM a atualização a cada passo se dá jogando parte da memória fora (através da porta de esquecimento) e agregando novas informações (por meio da porta de entrada). Como essas duas portas apresentam funcionalidades relativamente opostas, substituí-las por uma apenas seria equivalente. Essa é a intuição por trás das GRU.

Discussão:
Nas RNN tradicionais o conteúdo da unidade é sempre substituído por um novo valor que depende do último *hidden state* e da entrada. As unidades LSTM e GRU são capazes de guardar e adicionar novos conteúdos em cima. Outra vantagem dessas unidades é a de criar **atalhos** entre os passos temporais, evitando que o gradiente se anule rapidamente ao se retropropagar o erro. Por fim, a principal diferença entre a LSTM e GRU é que a primeira tem uma porta de saída que controla o tanto de memória que será exposto na saída, enquanto que a segunda expõe todo o conteúdo sem qualquer tipo de controle.

Experimentos:
As duas unidades mais a *tanh* foram testadas em uma duas tarefas: **modelagem de música polifônica** e **modelagem de sinal de fala**. Para a primeira tarefa, foram usados os *datasets Nottingham*, *JSB Chorales*, *MuseData* e *Piano-midi*, e usou-se a função logística na saída das unidades. Para a segunda tarefa, dois *datasets* foram utilizados: *Ubisoft A* e *Ubisoft B*, sendo que a rede foi projetada para olhar 20 amostras consecutivas e prever as próximas 10, considerando a técnica mistura de Gaussianas como camada de saída. Para a comparação ser justa, escolheu-se modelos com número de parâmetros próximos e modelos pequenos para evitar *overfitting*.

Resultados:
Para a **tarefa de música polifônica** os três modelos apresentaram **resultados semelhantes**, tendo a GRU ficado na frente em todos *datasets* com exceção do *Nottingham*. Para a **tarefa de fala**, tanto a GRU quanto LSTM apresentaram **resultados muito melhores** do que a rede *tanh* sem portas, tendo cada uma pontuado melhor em um *dataset*. Também são apresentados gráficos mostrando a convergência das redes por iterações (*epoch*) e por tempo de CPU (*wall clock time*).

Conclusão:
Além de as redes baseadas em unidades LSTM e GRU apresentarem resultados melhores do que a rede sem portas (*tanh*-RNN), elas também apresentam convergência geralmente mais rápida. Entretanto, os resultados não foram conclusivos a respeito da comparação entre LSTM e GRU, o que sugere que a melhor unidade vai depender do *dataset* e da tarefa em questão.