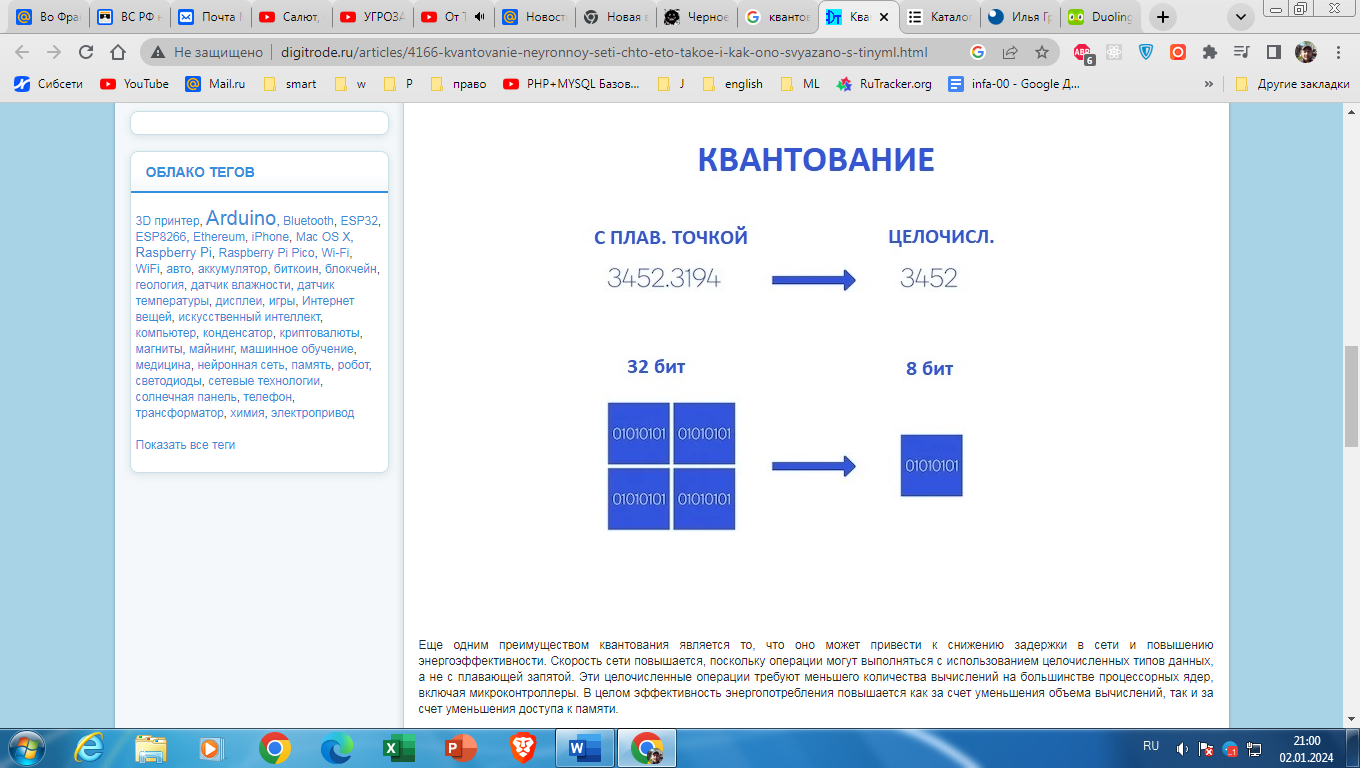
Квантование – это процесс снижения точности весов, смещений и активаций, чтобы они потребляли меньше памяти. Другими словами, процесс квантования – это процесс использования нейронной сети, которая обычно использует 32-битные числа с плавающей запятой для представления параметров, и вместо этого преобразует их для использования меньшего представления, такого как 8-битные целые числа. Например, переход от 32-битного к 8-битному формату уменьшит размер модели в 4 раза, поэтому одним из очевидных преимуществ квантования является значительное сокращение объема памяти. На следующем рисунке представлен пример квантования.



Еще одним преимуществом квантования является то, что оно может привести к снижению задержки в сети и повышению энергоэффективности. Скорость сети повышается, поскольку операции могут выполняться с использованием целочисленных типов данных, а не с плавающей запятой. Эти целочисленные операции требуют меньшего количества вычислений на большинстве процессорных ядер, включая микроконтроллеры. В целом эффективность энергопотребления повышается как за счет уменьшения объема вычислений, так и за счет уменьшения доступа к памяти.

Несмотря на преимущества, компромисс с квантованием заключается в том, что нейронные сети могут терять точность, поскольку они не точно представляют информацию. Однако, в зависимости от степени потери точности, архитектуры сети и схемы обучения/квантования сети, было показано, что квантование часто может привести к минимальной потере точности, особенно если сравнивать с улучшениями в задержке, использовании памяти и энергопотреблении.